Metodología de Trabajo y Análisis de Necesidades en la Certificación de Sistemas Críticos Embarcados basados en IA

2024

Hoja de Identificación del documento

Título:	Resultados de actividades en proyectos de investigación en áreas de interés ATM
Código:	N/A
Fecha:	Septiembre 2024
Fichero:	N/A

Autor:	Vicent Ortolà Plaza
Revisor:	Raquel Delgado-Aguilera Jurado
Aprobado:	N/A

Versiones:			
Numero	Fecha	Autor	Comentarios
01	20/09/2024	Vicent Ortolà Plaza	

Resumen Ejecutivo

La Inteligencia Artificial ha evolucionado exponencialmente en los últimos años, ganando cada vez más importancia, sobre todo en determinados campos. Este crecimiento tan rápido hace fundamental la necesidad de un correcto proceso de certificación y de garantizar su buen funcionamiento.

El presente trabajo se sitúa en el contexto de la certificación de Inteligencia Artificial en aviónica embarcada. Para ello se realiza una investigación y se consulta a expertos en la materia.

Dada la gran extensión del trabajo y su larga duración en comparación con el desarrollo de esta investigación, el objetivo será tratar y desarrollar algunos aspectos que se consideran importantes y útiles en el proceso de certificación, más allá de proporcionar una solución final.

La finalidad del trabajo es múltiple. Por un lado, se pretende crear una propuesta de metodología de trabajo para que se pueda trabajar de la manera más eficiente, conduciendo al desarrollo de una primera versión de un documento de partida que combine aquello que sigue siendo válido de referencias existentes para este trabajo, y siendo finalmente un proceso incremental en el que se vayan generando nuevas versiones a partir de la extensión y modificación de la versión primitiva.

Por otro lado, también se hará un análisis de necesidades para poder definir un criterio “Safety Critical”, estudiando estándares de referencia de certificación de software tradicional y de certificación de Inteligencia Artificial.

Para llevar a cabo todo lo mencionado, inicialmente se realizan una serie de ejercicios para recopilar información. Esto sirve como punto de partida para realizar toda la investigación posterior.

Índice

Índice de ilustraciones

Índice de tablas

Abreviaturas y Terminología

Tabla 1: Abreviaturas y terminología

Acrónimo	Definición
AICI	Artificial Intelligence Criticallity Index
AL	Assurance Level
AMC	Acceptable Means of Compliance
BFS	Base Functional Scope
CAC	Collaborative Air Combat
CS	Certification Specifications
DAL	Design Assurance Level
DL	Deep Learning
EASA	European Union Aviation Safety Agency
EMAD	European Military Airworthiness Document
ENAC	Entidad Nacional de Acreditación
HW	Hardware
IA	Inteligencia Artificial
IEEE	Instituto de Ingenieros Eléctricos y Electrónicos
LOR	Level of Rigor
MIT	Massachussets Institute of Technology
ML	Machine Learning

OACI	Organización de la Aviación Civil Internacional
OODA	Observe, Orient, Decide, Act
RNAV	Navegación Aérea
RNP	Performance de Navegación Requerida
SCF	Safety Critical Functions
SW	Software
SWCI	Software Criticality Index
UE	Unión Europea

1 INTRODUCCIÓN

1.1 Evolución de la Inteligencia Artificial

La Inteligencia Artificial empezó a estudiarse después de la Segunda Guerra Mundial, y el nombre se acuñó en 1956. Las décadas de 1950 y 1960 se entienden como el nacimiento de la IA y en ellas los primeros investigadores se centraron en el desarrollo de mecanismos de búsqueda de propósito general, en los que se entrelazaban elementos de razonamiento básicos para encontrar así soluciones completas [1].

En este período surgieron conceptos tan influyentes como el Test de Turing, para saber si una máquina exhibe un comportamiento inteligente. En el test se requería al agente inteligente poseer capacidades de procesamiento de lenguaje natural, representación del conocimiento, razonamiento automático, "Machine Learning", visión computacional y robótica [2].

En 1964, Joseph Weizenbaum, científico informático de MIT, desarrolló ELIZA, el primer "chatbot" que podía conversar funcionalmente en inglés con una persona [3]. La idea del profesor era que ELIZA conversase de manera escrita con el humano de tal forma que pareciese que lo estaba escuchando y empatizaba con él. Con este invento, Weizenbaum pretendía demostrar la superficialidad de la comunicación entre una persona y una máquina [4].

El crecimiento y aplicación de la Inteligencia Artificial se ha evidenciado en los años 90, a partir de este momento los investigadores han concentrado su interés en el desarrollo de inteligencias más generales, teniendo como resultado sub campos de la IA tales como reconocimiento de habla e imagen, redes neuronales, robótica, ML, entre otras [5].

Además, el "Machine Learning" ganó importancia como enfoque clave de la IA. Los investigadores se centraron en desarrollar algoritmos capaces de aprender de los datos y mejorar su rendimiento con el tiempo. Se desarrollaron técnicas como las redes neuronales, las máquinas de vectores de soporte y los árboles de decisión, que permitieron avances en áreas como el reconocimiento de imágenes y del habla [6].

La IA ha ido evolucionando desde los primeros sistemas basados en reglas hasta enfoques basados en datos como el "Machine Learning" y el "Deep Learning", es decir, modelos más flexibles y adaptables, capaces de aprender grandes cantidades de datos. En las décadas de los 2000 y 2010 se produjo un aumento de la disponibilidad de grandes cantidades de datos, junto con mejoras en la potencia de cálculo. La transformación de la amplia gama de datos en conocimiento valioso es cada vez más importante en varios ámbitos [7].

Esto condujo al auge del DL, un sub campo del "Machine Learning" que utiliza redes neuronales con múltiples capas. El "Deep Learning" es responsable de importantes avances en diversos campos en los que la comunidad de la IA ha luchado durante muchos años [8].

Los algoritmos de DL demostraron ser muy eficaces en tareas como reconocimiento de imágenes y del habla [9] [10].

Dos cosas harían posible la revolución de aplicaciones de redes neuronales y algoritmos de "Deep Learning". En primer lugar, los avances de hardware especializado, que han acelerado drásticamente el entrenamiento y el rendimiento de las redes neuronales y reducido su consumo de energía. En segundo lugar, el aumento de datos abiertos disponibles "online", impulsan el desarrollo del DL [3].

La posibilidad de crear máquinas que son capaces de pensar plantea una serie de cuestiones éticas, relacionadas para asegurarse de que este tipo de máquinas no daña a los humanos y a otras cuestiones moralmente relevantes, y del estado moral de las mismas máquinas [11].

Recientemente, la preocupación por la ética y la transparencia ha ido en aumento a causa de la integración cada vez mayor de la IA en nuestra sociedad. De este modo, los investigadores y los responsables políticos están tratando de desarrollar sistemas de la IA para que sean responsables y justos, realizando esfuerzos para garantizar el despliegue responsable de la IA y abordar los riesgos potenciales asociados a las tecnologías de la IA.

El razonamiento de la IA debe ser capaz de tener en cuenta los valores sociales y las consideraciones morales y éticas; sopesar las prioridades respectivas de los valores de las distintas partes interesadas en diversos contextos multiculturales; explicar su razonamiento; y garantizar la transparencia [12].

La IA normalmente se compara con la inteligencia humana cuando se realiza una prueba específica. Por ello, es necesario conocer los mecanismos de funcionamiento de la mente humana, a través de la introspección y experimentos psicológicos. De acuerdo al libro "Inteligencia Artificial: Un enfoque Moderno" podemos visualizar que el concepto IA varía según su punto de vista y pueden definirse a través de los procesos mentales o de la conducta [13]. Véase Ilustración 1.

Sistemas que piensan como humanos	**Sistemas que piensan racionalmente**
«El nuevo y excitante esfuerzo de hacer que los computadores piensen… máquinas con mentes, en el más amplio sentido literal». (Haugeland, 1985) «[La automatización de] actividades que vinculamos con procesos de pensamiento humano, actividades como la toma de decisiones, resolución de problemas, aprendizaje…» (Bellman, 1978)	«El estudio de las facultades mentales mediante el uso de modelos computacionales». (Charniak y McDermott, 1985) «El estudio de los cálculos que hacen posible percibir, razonar y actuar». (Winston, 1992)
Sistemas que actúan como humanos	**Sistemas que actúan racionalmente**
«El arte de desarrollar máquinas con capacidad para realizar funciones que cuando son realizadas por personas requieren de inteligencia». (Kurzweil, 1990) «El estudio de cómo lograr que los computadores realicen tareas que, por el momento, los humanos hacen mejor». (Rich y Knight, 1991)	«La Inteligencia Computacional es el estudio del diseño de agentes inteligentes». (Poole *et al*., 1998) «IA… está relacionada con conductas inteligentes en artefactos». (Nilsson, 1998)

Ilustración 1: Definiciones de IA desde diferentes enfoques [13].

El futuro de la IA deparará nuevos avances en ámbitos como la IA explicable, el aprendizaje por refuerzo y el desarrollo de sistemas de IA que puedan colaborar eficazmente con los seres humanos.

1.2 Concepto de Inteligencia Artificial

Para definir el concepto de Inteligencia Artificial es conveniente centrarse en varias fuentes. La RAE define la IA como la disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico [14].

Por otro lado, en el libro "Inteligencia Artificial", Lasse Rouhiainen define el término IA como la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano. Sin embargo, se argumenta que, a diferencia de las personas, los dispositivos basados en IA pueden analizar grandes volúmenes de información a la vez. Asimismo, la proporción de errores es significativamente menor en las máquinas que realizan las mismas tareas que sus contrapartes humanas [15].

Por su parte, y situando el foco en documentos regulatorios de aviación, la European Union Aviation Safety Agency (EASA) en el documento "EASA Concept Paper: First usable guidance for Level 1 & 2 machine learning applications" define la IA como una tecnología que puede, para un conjunto determinado de objetivos definidos por el ser humano, generar resultados como contenidos, predicciones, recomendaciones o decisiones que influyen en los entornos con los que interactúan.

Como explica EASA, un sistema basado en la IA se compone de varios subsistemas tradicionales, y al menos uno de ellos es un subsistema basado en la IA. Los elementos de hardware y software tradicionales no incluyen un modelo de inferencia ML. Véase Ilustración 2.

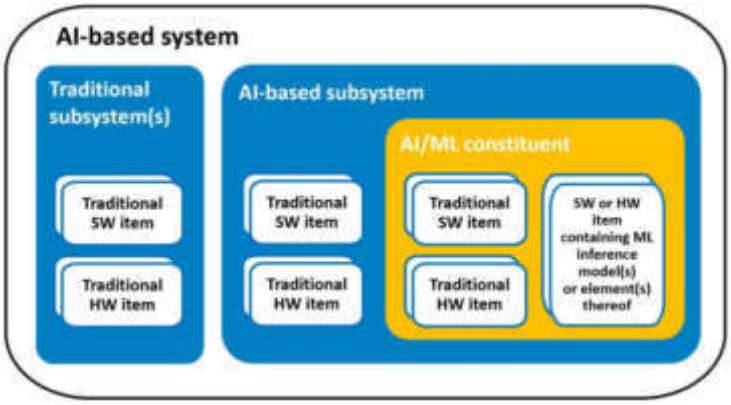


Ilustración 2: Subsistemas de un sistema basado en IA. EASA Concept Paper: First usable guidance for Level 1 & 2 machine learning applications.

En la línea de EASA, el documento "AIR 6988: Statement of concerns" sostiene que la IA es la teoría y el desarrollo de sistemas basados en software capaces de realizar tareas que hasta ahora eran competencia de la inteligencia humana. Algunos ejemplos son la percepción visual, el reconocimiento del habla, la toma de decisiones, la atención al cliente y la detección de anomalías. John McCarthy, que acuñó el término en 1956, lo define como "la ciencia e ingeniería de hacer máquinas inteligentes".

En el "White Paper On Artificial Intelligence - A European approach to exellence and trust", la Comisión Europea sostiene que la IA es una tecnología estratégica que ofrece muchos beneficios para los ciudadanos, las empresas y la sociedad en su conjunto, siempre que esté centrada en el ser humano, sea ética, sostenible y respete los derechos y valores fundamentales. La IA ofrece importantes ganancias de eficiencia y productividad que pueden reforzar la competitividad de la industria y mejorar el bienestar de los ciudadanos [16].

Atendiendo a las definiciones anteriores, se entiende que la Inteligencia Artificial es una tecnología que combina algoritmos con la finalidad de crear elementos o máquinas para automatizar y sustituir funciones de forma que imiten las tareas realizadas por el ser humano. Estas tareas pueden incluir la resolución de problemas, el aprendizaje, el reconocimiento de patrones y la toma de decisiones. Además de sustituir funciones humanas, la capacidad de la IA de trabajar con grandes volúmenes de información, con menos errores y sin necesidad de descanso, ha producido que ésta haya encontrado numerosas aplicaciones en diversos sectores. Esta compleja tecnología está creciendo de forma exponencial y cada vez está más presente en nuestro día a día.

En la actualidad, la IA ha tomado un papel importante debido al alto volumen de datos que se generan dentro de las diferentes industrias. Los algoritmos son más sofisticados, rápidos y pueden resolver bases de datos cada vez más extensas y heterogéneas. Además, debido a que la potencia computacional ha mejorado ha nacido la era del "Big Data", esta era es caracterizada por las 3 V: volumen, velocidad y variedad. Esto se debe a que se almacenan un gran volumen de datos, estos

datos tienen una amplia variedad de formatos (números, imágenes, textos y otras) y son analizados a una alta velocidad. Las herramientas usadas para el análisis del "Big Data" son "Machine Learning" y "Deep Learning" [13].

El "Machine Learning" consiste en un conjunto de métodos usados con el objetivo de encontrar patrones a partir de los datos de forma automática. Los patrones que son encontrados pueden ser usados para crear predicciones de datos nunca antes vistos y pronostica el comportamiento futuro, los pronósticos pueden ayudar a la identificación de acciones subsecuentes sin necesidad de entender totalmente el comportamiento de los datos. De esta manera el ML ha sido una herramienta efectiva en aplicaciones como toma de decisiones, detección de fraude, diagnóstico de cáncer, recomendación de sistemas, asistentes de voz, y otras [13].

El "Deep Learning" permite a los modelos computacionales compuestos por múltiples capas de procesamiento aprender representaciones de datos con múltiples niveles de abstracción. Estos métodos han mejorado drásticamente el estado del arte en el reconocimiento del habla, el reconocimiento visual y la detección de objetos y muchos otros dominios [17].

El DL descubre estructuras intrincadas en grandes conjuntos de datos utilizando el algoritmo de retropropagación para indicar cómo una máquina debe cambiar sus parámetros internos que se utilizan para calcular la representación en cada capa a partir de la representación en la capa anterior [17].

EASA explica la taxonomía definida anteriormente utilizando un esquema. Véase Ilustración 3.

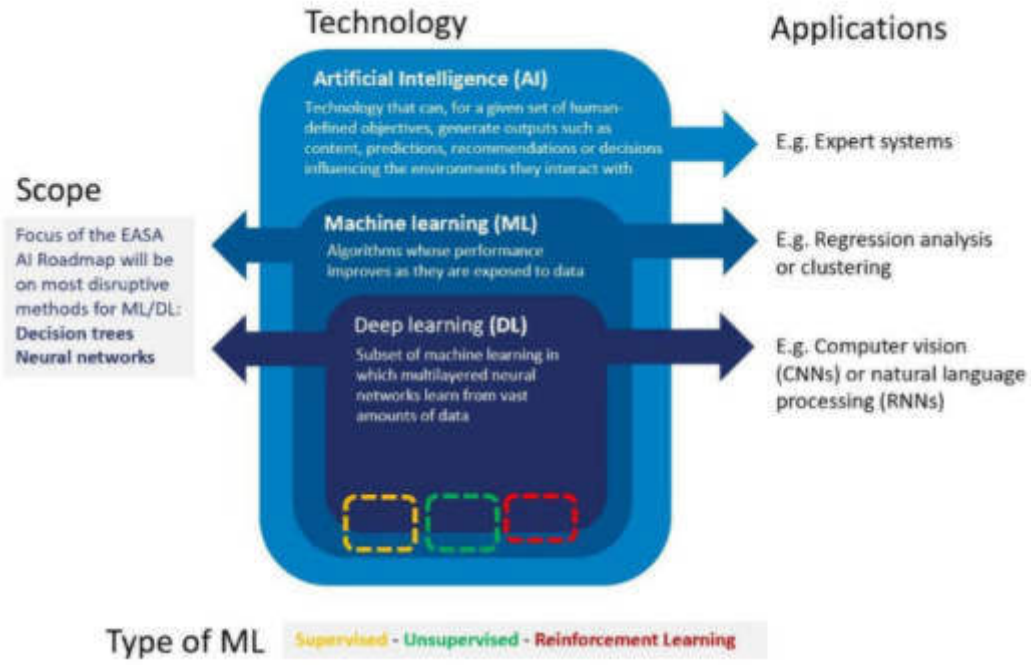


Ilustración 3: Taxonomía de la IA. EASA AI Roadmap.

1.3 Inteligencia Artificial en aviación.

El Transporte Aéreo es una de las industrias con mayor crecimiento actualmente, lo cual se traduce en mayor consumo de combustible, contaminación ambiental, tiempos de servicio, mantenimiento, consumibles y tripulación, que conlleva a mayores costos operacionales [18].

La IA se está integrando cada vez más en la aviónica para mejorar la seguridad, la eficiencia y la automatización en la industria de la aviación.

El actual aumento masivo de generación de datos permite utilizar algoritmos que transforman estos datos en información útil, algunas investigaciones han permitido que las aerolíneas utilicen sistemas de Inteligencia Artificial construyendo algoritmos de "Machine Learning" para recolectar y analizar datos con el propósito de predecir demoras, clima, performance, planes de vuelo o combustible. Algunas de estas compañías utilizan softwares como los desarrollados por "Airspace Intelligence", "FlightAware", entre otras [13].

El constante crecimiento ha obligado la necesidad de mejorar los sistemas de control de tráfico aéreo con el fin de prevenir demoras y mejorar la seguridad. Las soluciones que se han brindado es la aplicación de técnicas como Navegación de Área, método de navegación que permite la operación de aeronaves en cualquier ruta de vuelo deseada dentro de la cobertura de las ayudas a la navegación con referencia a la estación o dentro de los límites de la capacidad de las ayudas autónomas, o una combinación de estas. Este método fue mejorado con la introducción del

Rendimiento de Navegación Requerido, cuya diferencia radica en el monitoreo y alerta de rendimiento de la aeronave a bordo [19].

Estas metodologías han permitido operar una mayor cantidad de aeronaves en el mismo espacio aéreo y mejorar la seguridad, pero no están dedicadas a la optimización, concepto en el cual se interesan los operadores. Un plan de vuelo eficiente es uno de los factores más importantes en las operaciones aéreas, ya que permite operar con seguridad, mejorar la confianza de la tripulación y un ahorro significativo de combustible [20].

Por su parte, la IA desempeña un papel crucial en el desarrollo de sistemas autónomos para aeronaves. Esto incluye el control de vuelo autónomo, el papel en los sistemas de seguridad para evitar colisiones y los procedimientos automatizados de aterrizaje y despegue. Los algoritmos de IA analizan datos de sensores, como radares y cámaras, para tomar decisiones y controlar la aeronave [21] [22].

Asimismo, a IA se utiliza en los sistemas de gestión del tráfico aéreo para mejorar la eficiencia del uso del espacio aéreo, reducir la congestión y mejorar la seguridad. Los algoritmos de IA ayudan a predecir el flujo de tráfico, optimizar las rutas y proporcionar información en tiempo real a pilotos y controladores aéreos [23].

Los operadores aéreos utilizan softwares que permiten planear el vuelo teniendo en cuenta diferentes aspectos operacionales, el resultado de esta planificación no siempre resulta el más óptimo debido a las congestiones del espacio aéreo no pronosticadas o meteorología, ya que los despachadores se basan en las publicaciones que actualiza la autoridad aeronáutica [13].

La IA se utiliza para optimizar las rutas de vuelo, teniendo en cuenta factores como las condiciones meteorológicas, el tráfico aéreo y el consumo de combustible. Analizando grandes volúmenes de datos y utilizando algoritmos de aprendizaje automático, la IA puede sugerir las rutas más eficientes para minimizar el consumo de combustible y reducir el tiempo de vuelo [24].

En aviónica embarcada, cada vez son más los sistemas que emplean IA en una aeronave. Estos sistemas se suelen componer de una parte de sistemas técnicos (donde se puede emplear IA), una parte de procesos y una parte de decisión y supervisión humana, componiendo un sistema de sistemas [25].

En el mundo militar, como consecuencia del avance exponencial de estas tecnologías, cada vez son más los componentes que utilizan IA. El papel de la IA en la Nube de Combate gana importancia. Es necesario proporcionar una definición de la Nube de Combate, que, además, será de mucha utilidad para el desarrollo del trabajo.

La Nube de Combate es una reserva de recursos de combate flexible y dinámica formada por la reorganización orgánica de estos recursos desplegados de forma dispersa. El estilo de combate distribuido basado en la Nube de Combate es el comportamiento de concentrar rápida y flexiblemente todos los recursos de combate dispersos en uno o más objetivos de ataque o defensa [26].

Por otro lado, EMAD entiende Nube de Combate como la aplicación de una solución tecnológica avanzada a las capacidades militares que habilita su empleo, especialmente el mando y control, en el multidominio y que permite mediante la captura , procesamiento y distribución de datos, incluidos los que proporcionan sensores y sistemas e intercambio de información de datos, así como la prestación de servicios, que cada usuario, plataforma o nodo autorizado contribuya y reciba información esencial a tiempo para que sea capaz de utilizarla para la toma de decisiones y la ejecución de operaciones militares dentro de un espacio de batalla [27].

La Nube de Combate integraría sistemas tripulados y no tripulados y utilizaría los avances en sigilo, armas de precisión y herramientas avanzadas de mando y control, garantizando que ningún punto de ataque pueda paralizar las operaciones de combate. Este esfuerzo también brindaría la oportunidad de crear capacidades de combate modulares y escalables, en lugar de obligar a aeronaves individuales u otros activos a asumir cada vez más tareas [28].

1.4 Certificación de Inteligencia Artificial.

La utilización de la Inteligencia Artificial constituye una de las más significativas aportaciones tecnológicas que impregnará la vida de la sociedad los próximos años, en muchas de sus actividades cotidianas y en sus sectores más representativos, desde la industria al sistema financiero, pasando por la educación, la salud, el transporte y, por descontado, la defensa y la seguridad, aportando significativos beneficios, pero evidenciando también riesgos que es necesario valorar y minimizar.

Una realidad tan disruptiva como la IA exige que su tecnología y la de los productos y servicios sustentados en ella ofrezcan suficientes garantías de su adecuado funcionamiento [29].

La Inteligencia Artificial es una de las tecnologías de más rápido crecimiento del siglo XXI y nos acompaña en nuestra vida cotidiana al interactuar con aplicaciones técnicas. Sin embargo, la confianza en estos sistemas técnicos es crucial para su amplia aplicabilidad y aceptación. Las herramientas sociales para expresar la confianza suelen estar formalizadas por reglamentos legales, es decir, estándares, normas, acreditaciones y certificados.

Se debe intentar analizar las aplicaciones de aprendizaje automático desde múltiples perspectivas para evaluar y verificar los aspectos de desarrollo seguro de software, requisitos funcionales, calidad de datos, protección de datos y ética [30].

El debate sobre los principios éticos de una IA fiable ha atraído la atención de la Comunidad Europea, que ha publicado una propuesta legislativa para regular las aplicaciones de la IA. El objetivo es garantizar que las tecnologías más recientes se utilicen de forma segura y conforme a la ley, incluido el respeto de los derechos fundamentales.

La Inteligencia Artificial es un claro apoyo en muchos escenarios de toma de decisiones, pero cuando se trata de áreas sensibles como la sanidad, las políticas de contratación educación, banca

o justicia, con gran impacto en las personas y la sociedad, se hace crucial establecer directrices sobre cómo diseñar, desarrollar, desplegar y supervisar esta tecnología.

Todos los ámbitos sensibles mencionados aparecen en la propuesta de la Comunidad Europea, y los sistemas de IA relacionados se clasifican como de alto riesgo, con un llamamiento explícito a la calidad de los datos y al control de la solidez algorítmica contra los sesgos. En este escenario, la normalización desempeñaría un papel clave para definir soluciones técnicas que pueden utilizar los proveedores de IA para garantizar el cumplimiento de la normativa de la UE [31].

Desde pequeños proveedores de servicios hasta grandes empresas y organizaciones gubernamentales utilizan aplicaciones basadas en IA/ML para prestar servicios. Sin embargo, no todos disponen de los recursos o los conocimientos técnicos necesarios para comprobar la idoneidad de los sistemas de IA que están implantando [32].

Con el uso generalizado y omnipresente de la IA para sistemas automatizados de toma de decisiones, el sesgo de la IA es cada vez más evidente y problemático. Una de sus consecuencias negativas es la discriminación: el trato injusto o desigual de las personas en función de determinadas características [33].

Por lo tanto, se necesita un procedimiento de prueba estándar para comprobar si la aplicación de IA presenta sesgos. Un certificado de imparcialidad emitido por una autoridad de certificación neutral basado en dicho procedimiento de prueba estándar garantizaría a los usuarios la imparcialidad del sistema de IA/ML [32].

El procedimiento de prueba estándar debe proteger a los desarrolladores de sistemas de IA patentados, la privacidad y la seguridad del conjunto de datos de entrenamiento [34]. Lógicamente, también debe cumplir la normativa nacional e internacional vigente [31].

Al igual que existen procedimientos establecidos para la auditoría de seguridad y certificación de sitios web y portales por parte de organismos designados en función de criterios de referencia aceptados, asimismo debe evolucionar un ecosistema para la auditoría y certificación de la imparcialidad de los sistemas basados en IA [32].

Siguiendo la línea de lo comentado, cabe la posibilidad de que estos sistemas estén condicionados por sesgos no deseados o tengan algún tipo de mal funcionamiento. Si esta problemática produce fallos que causan un aumento en el riesgo de la seguridad de las personas y su entorno, pudiendo suponer lesiones o muertes humanas, daños materiales, pérdida económica como consecuencia de los daños o efectos sobre el medio ambiente podría considerarse al sistema como crítico para la seguridad.

Llegados a este punto, verificar que estos sistemas, subsistemas o programas son capaces de satisfacer sus requisitos especificados durante un período de tiempo definido en un entorno operativo es necesario.

2 CONTEXTO

El presente trabajo se sitúa en el ámbito de certificación de componentes o sistemas basados en Inteligencia Artificial críticos para la seguridad en aviónica embarcada. Se pretende elaborar una primera aproximación que pueda traducirse en propuestas de normas para cumplir con las necesidades requeridas por algunos de estos. Una parte de estos sistemas estarán orientados al ámbito militar, pudiendo ser componentes clave en el Combate Aéreo Colaborativo.

Para poder plantear un proceso o metodología de trabajo el primer paso era comprender los objetivos principales de la investigación, leyendo sobre los procesos de certificación de SW tradicionales y el funcionamiento de Inteligencia Artificial en el contexto de la aviación, para así realizar una primera aproximación de las funciones que puede englobar la Nube de Combate.

El objetivo de esta tarea consiste en el análisis de necesidades para el uso de IA centrada en cuestiones de aeronavegabilidad y seguridad.

La tarea mencionada tiene por objeto proporcionar una base para poder generar a futuro un documento en el que se incorporen contenidos que conformen el resultado de dicho análisis como términos de referencia y taxonomía; definición de las funciones críticas de seguridad dentro de los casos de uso e identificación de las cuestiones de aeronavegabilidad y seguridad, así como de los retos que plantea la aplicación de las tecnologías de IA en un entorno crítico para la seguridad.

La tarea conlleva una gran extensión y larga duración, por lo que en este documento tan solo se tratarán algunos aspectos que se consideran relevantes para el desarrollo de la misma, con la intención de que posteriormente se pueda seguir la línea de esta investigación.

3 OBJETIVOS

El objetivo principal es generar una posible metodología de trabajo para poder abordar la tarea explicada con mayor eficiencia. Para ello, se han llevado a cabo una serie de acciones que se explicarán más adelante y que servirían como guía para realizar el análisis de necesidades.

Asimismo, se tratarán otros aspectos relevantes, como analizar cuáles son las necesidades para definir un criterio de criticidad para la seguridad, estudiar una primera versión de un documento de partida (V0) que combine los aspectos que siguen siendo válidos y usables de referencias existentes (EASA, ARP, MIL-882) y sea de utilidad para el análisis de algunos estándares.

Como ya se ha explicado, el objetivo no consiste en resolver la tarea de certificar IA embarcada, sino desarrollar y tratar algunos factores que puedan ser útiles para iniciar esta tarea.

Para cumplir los objetivos, se realiza una investigación inicial en la que se toma nota de expertos en la materia que sirve como introducción a los procesos de certificación tradicionales, en la que se reúne información de gran utilidad para el desarrollo general de la tarea y en la que se estudia y debate que implicación puede tener la IA y su certificación en un contexto CAC.

4 CONCEPTOS PREVIOS. CONTEXTO DEL CAC.

4.1 Conceptos previos. Contexto del CAC.

Previamente a trabajar en los objetivos propuestos, se realiza una breve introducción de algunos conceptos clave que se debían tratar y considerar. Esta parte consiste en una explicación de lo que significan algunos conceptos en el contexto del CAC. Además, se pidió opinión a expertos en la materia para poder recopilar información que permitiera realizar una investigación más completa.

4.1.1 Concepto de aeronavegabilidad

Un objetivo clave del proceso de aeronavegabilidad es garantizar que el diseño es seguro para operar dentro de su ámbito de operación previsto. Por tanto, es un concepto principal que es necesario esclarecer desde el principio.

Por ello, es necesario plantearse la pregunta de si es necesario adaptar, ampliar o extender el concepto de Aeronavegabilidad al considerar un Sistema de Sistemas como la Nube de Combate.

Atendiendo al manual de aeronavegabilidad de la OACI, se define condición de aeronavegabilidad como el estado de una aeronave, motor, hélice o pieza que se ajusta al diseño aprobado correspondiente y está en condiciones de operar de modo seguro [35].

Otra definición de aeronavegabilidad es la siguiente: La aeronave se ajusta a su diseño de tipo y está en condiciones de operar con seguridad [36].

Para una aeronave, o parte de una aeronave, aeronavegabilidad es la posesión de los requisitos necesarios para volar en condiciones seguras, dentro de los límites permitidos [37].

En el contexto militar se pueden proporcionar las siguientes definiciones de aeronavegabilidad:

Aeronavegabilidad: “Cualidad que hace a una aeronave segura para el vuelo. Es la propiedad de un sistema aéreo, en una determinada configuración, de conseguir, mantener y acabar un vuelo de forma segura de acuerdo a las limitaciones de uso aprobadas" [38].

Aeronavegabilidad: “Capacidad de una aeronave, u otro equipo o sistema aerotransportado, para operar en vuelo y en tierra sin peligro significativo para la tripulación, el personal de tierra, los pasajeros (en su caso) u otros terceros" [39].

En cuanto al concepto de aeronavegabilidad, una vez se han planteado las diferentes definiciones, pueden surgir una serie de discrepancias. En primera instancia, puede no entenderse la necesidad de plantear una definición de aeronavegabilidad, si es algo que habitualmente se “da por hecho” y se coge la del regulador. Sin embargo, en este contexto es necesario por el cambio de enfoque a

la Nube de Combate. Pero no se trata de inventar una nueva definición, sino valorar si alguna de las existentes es aplicable, o si se debe adaptar, según se crea necesario.

En segunda instancia, se cuestiona la necesidad de considerar la definición de OACI si hay que enfocarse en el ámbito militar, atendiendo a que las características en cuanto a safety de una misión militar son muy distintas al ámbito civil, por lo que no se debería aplicar OACI. No obstante, las definiciones militares dependen de cada Estado. Por ejemplo, el concepto de safety militar incluye aspectos de la misión, si no se cumple un aspecto de la misión, eso no es aceptable. Precisamente se argumenta que se han incluido definiciones militares y que se han tenido presentes.

4.1.2 Concepto de aeronavegabilidad

En la línea de lo que se busca en necesario plantearse qué significan las "Safety Critical Functions" en un sistema de sistemas como la nube de combate.

Este aspecto se desarrolla con más profundidad en otro apartado del trabajo. Sin embargo, por ahora una SCF se puede definir como función cuyo fallo de funcionamiento o funcionamiento incorrecto provocará directamente un percance de gravedad catastrófica o crítica [40].

Desde la perspectiva de la aeronavegabilidad, la identificación de las SCF es esencial para el proceso de verificación de todas las funciones que contribuyen al riesgo de aeronavegabilidad.

En cuanto al concepto de "Safety Critical Functions" pueden llegar a surgir discrepancias similares a las de la definición de Aeronavegabilidad. Se puede interpretar que, en cuanto a la misión militar, puede haber algunos factores que no afecten directamente a vidas humanas, pero que sean críticos para la campaña.

En este punto se considera la importancia de las cuestiones relativas a "fitness for flight", que se trata de uno de los aspectos que pueden ser críticos.

"Fitness for flight" se refiere a los requisitos de salud física y mental que pilotos, tripulación y otro personal determinado deben cumplir para garantizar un desempeño seguro y eficiente de sus funciones. Los requisitos específicos de aptitud para el vuelo pueden variar en función del país, la autoridad aeronáutica y el tipo de personal de aviación de que se trate [41].

4.1.3 Clasificación del riesgo y la criticidad en función de varias dimensiones

Otro aspecto importante es definir qué dimensiones deben considerarse a la hora de definir los niveles de garantía en la Nube de Combate para aplicaciones de IA.

Se debe determinar cuáles son los criterios para identificar las funciones y cuáles de ellos se pueden aplicar. Se tiene distintos tipos de clasificaciones en función de las referencias existentes

de las asociaciones de seguridad. Hay que decidir si estas referencias se deben adaptar, si se pueden aplicar directamente o si se tienen que crear otras.

De manera muy resumida, se exponen los diferentes tipos de clasificaciones de algunas referencias para tener una noción inicial de los criterios que se utilizan para determinar el riesgo y la criticidad.

- **RTCA DO-178C**
  - Efecto en el avión
  - Efecto sobre la tripulación
  - Efecto sobre otros ocupantes, excepto la tripulación de vuelo

- **EASA AI Framework**
  - Enfoque sin riesgo
  - Nivel de asistencia humana, colaboración y autonomía

- **EU Regulatory Framework**
  - Sin riesgo
  - Riesgo mínimo
  - Riesgo limitado
  - Riesgo elevado
  - Riesgo inaceptable

- **German Data Ethics Commission**
  - Pirámide de criticidad de cinco niveles

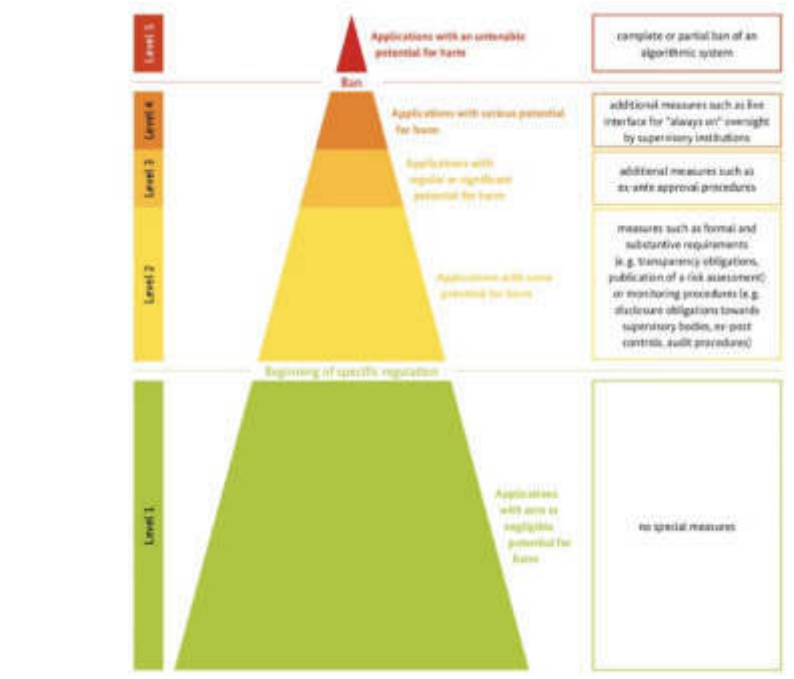


Ilustración 4: Pirámide de criticidad. German Data Ethics Commission.

- **TÜV Austria Holding AG**
  - Efectos en las personas, el medio ambiente y las organizaciones
  - Efectos sobre los seres vivos, pérdida de confianza

- **Multi-domain Combat Cloud sensitivity of the situation**
  - Riesgo para las tripulaciones propias
  - Riesgo de daños colaterales
  - Riesgo para el resto de la campaña aérea

- **AI Methods, Capabilities and Criticality Grid — AI_MC2**

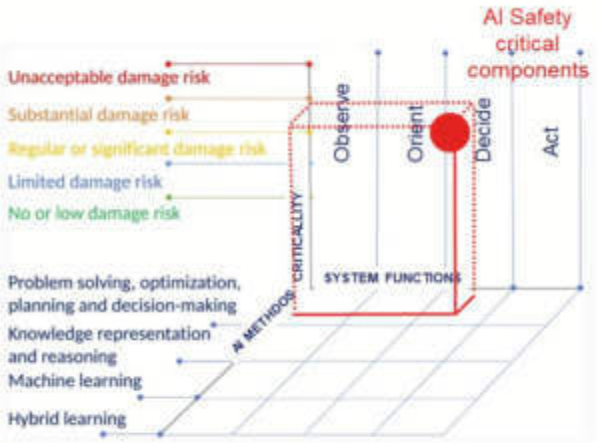


Ilustración 5: Matriz 3D criticidad/métodos IA/funciones.

Sobre las dimensiones a considerar, se debe plantear si es necesario establecer una dimensión para las técnicas de IA, ya que puede haber solape entre ellas. Además, siguiendo el razonamiento de algunos expertos, se argumenta que lo importante no es la técnica, sino la criticidad. No obstante, la finalidad no es tanto para ver si se considera la técnica de IA como dimensión, sino para resaltar de forma general que se debe decidir qué dimensiones considerar en el análisis de safety, puede que al final la técnica de IA no vaya en el análisis que se realizará. Por ello, aunque efectivamente haya solape entre las técnicas de IA y se esté considerando, es necesario pensar en las dimensiones que se van a adoptar desde una perspectiva más genérica.

También hay que cuestionarse si las técnicas de ML serían de aprendizaje online u offline. Por el momento los sistemas adaptativos que aprendan en tiempo real parece que están fuera del alcance de esta tarea, y probablemente lo estén al menos por unos años más. Precisamente la parte de decidir qué tecnologías se considerarán y cuáles no forma parte del análisis de necesidades y del análisis del estado del arte, este es uno de los propósitos de ambos.

4.1.4 Clasificación tradicional de software en la aviación

Un enfoque integral de la seguridad del software implica aplicar una serie de medidas para identificar, gestionar y mitigar los riesgos de seguridad a lo largo del ciclo de vida de desarrollo del software. Elementos de un enfoque integral de la seguridad del software:

- Establecimiento de niveles de software, normalmente de acuerdo con las normas prescritas del sector
- Identificación de las funciones críticas para la seguridad y su software crítico para la seguridad asociado

- Enfoque basado en el riesgo: Niveles de garantía de SW definidos en función de la gravedad del efecto de un fallo.
- Análisis y solución de fallos puntuales causados por software
- Elaborar los planes de seguridad y de software necesarios, así como otra documentación.

Se debe integrar un programa completo de seguridad del software (que incluya todas las cuestiones clave de seguridad del software) en el programa general de seguridad del sistema.

En cuanto a la certificación, un proceso de trabajar, a modo de breve resumen, puede ser el siguiente: tomar de EASA los códigos de aeronavegabilidad, analizar los "High Level Safety Objectives", hacer un "tailoring"[1] de lo que dicen los códigos de EASA para adaptarlos a la certificación de su sistema, se van concretando las "boundaries"[2] del proceso, y al llegar a un nivel más concreto se emplean los estándares ARP o EUROCAE según se requiera, y el safety con RTCA para los componentes.

Hacer un "tailoring" del código forma parte de incorporar cuestiones del proceso de certificación, en concreto la selección de la base de certificación y de posibles condiciones especiales para certificación aprobadas por la autoridad. Por lo que el proceso mencionado puede ser de aplicación para la tarea a realizar.

[1] Práctica de adaptar normas, metodologías o marcos predefinidos a una situación concreta.

[2] Límites, directrices o reglas que definen el comportamiento, las acciones o las condiciones aceptables y apropiadas dentro de un contexto o relación concretos. Establecen los parámetros y definen lo que se considera permisible o aceptable en una situación determinada.

5 METODOLOGÍA DE TRABAJO DE EL *BASE FUNCTIONAL SCOPE* Y EN LA REVISIÓN DE ESTÁNDARES APLICABLES.

Este capítulo tratará dos secciones. Por un lado, una sección recogerá el conjunto de funciones IA que se consideran críticas en el contexto CAC. Por otro lado, la segunda sección puede consistir en el análisis de los estándares existentes que sean de aplicación en la tarea.

5.1 *Base Functional Scope*. Necesidad y metodología de trabajo.

Uno de los objetivos principales de la tarea consiste en definir las funciones críticas de seguridad dentro de los casos de uso e identificar las cuestiones de aeronavegabilidad y seguridad que surgen a partir de estas funciones.

Como se ha explicado anteriormente, no se dispone de los casos de uso para determinar las funciones críticas, al depender de otro paquete de trabajo. Además, hay que recordar que el objetivo de este trabajo no es realizar la tarea, sino proponer un proceso que sirva como guía posteriormente y que pueda suponer la base de algunas secciones del documento final.

5.1.1 Necesidad y finalidad de un BFS CAC.

Esta sección explica la necesidad de considerar un "Base Functional Scope" como punto de partida del trabajo para una mejor comprensión y dimensionamiento de los procesos CAC, así como los problemas, restricciones, limitaciones y necesidades de certificación de los componentes IA como parte de las funciones críticas de seguridad CAC. Este BFS CAC es necesario para el posterior desarrollo del trabajo.

La necesidad de una BFS surge por tener que reducir todo el marco de funciones de Inteligencia Artificial a las necesidades específicas de la tarea.

Para definir la BFS se deben identificar y describir las funciones críticas para la seguridad desempeñadas por IA que deben abordarse, definiendo la forma de uso y las tecnologías IA candidatas. Por tanto, para definir y desarrollar la BFS:

- Hay que identificar que funciones IA de la Nube de Combate son importantes para la tarea, diferenciando cuales son críticas para la seguridad. Al no disponer de los casos de uso en primera instancia, es muy probable que sea necesario realizar iteraciones sobre el listado inicial de funciones identificado.

- Cuando se disponga de ellos, se debe elaborar y profundizar internamente en los casos de uso según sea necesario: descomposición funcional, subsistemas y componentes de IA, evaluación de la seguridad funcional, asignación de niveles y objetivos de seguridad.

- Se deben considerar las particularidades de los algoritmos de IA que se emplearán y las necesidades de las aplicaciones críticas para la seguridad. De esta manera, además de las funciones, se han de considerar las tecnologías IA.

- Para definir la BFS pueden ser necesarios distintos niveles de conformidad, ya que la criticidad de las aplicaciones puede variar en función de la misión y los algoritmos.

- Se ha de estudiar la herramienta que se empleará. Una misma función la pueden desempeñar distintas tecnologías de Inteligencia Artificial con algoritmos diferentes. Siguiendo esta explicación, puede ser necesaria la construcción de una matriz funciones/tecnologías IA. Véase Ilustración 6.

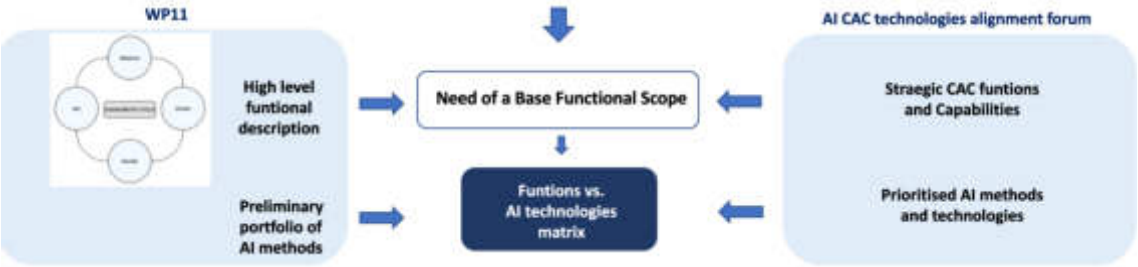


Ilustración 6: Necesidad de una BFS y una matriz funciones vs. Tecnologías IA.

5.1.2 Metodología del BFS.

La metodología propuesta para realizar esta parte del trabajo es la siguiente:

- Identificar una función inicial que sirva de ejemplo. Se propone el ciclo OODA como marco genérico para la identificación de funciones, este bucle se explicará más adelante.

- Lista preliminar de funciones. Se proporcionará una plantilla para la descripción y un ejemplo. La siguiente tabla sintetiza las funciones identificadas considerando el ciclo OODA. Esta lista es demasiado larga y se debe hacer un filtro preliminar. Esta sección también incluirá los criterios utilizados para filtrar y priorizar la lista de funciones. Véase Ilustración 7.

Observe	Orient	Decide	Act
• Detecting enemy • High level tactical assistance • Imagery adquisition (COP/CROP) • Global situational awareness • Server comms • ISTAR (Intelligence, Surveillance, Target Acquisition and Reconnaissance) • Classify sensor data • Battle damage assessment • Detect health issues • Detect cybersecurity issues (intrusion) • Automatic Detection and Tracking of objects • Normalized navigation maps • Detect critical infrastructure • Secure communication • EOA • Threat detection • Population-wide Monitoring • Cross share information from sensor and dynamically combine sensing • Perform reconnaissance and discover targets • Management of dynamic data stream and priorities • Use mil. TTP and doctrines	• Data fusion of sensor signals • High level tactical assistance • Security • Time requirements • Tasking • Data presentation for decision taking • Training • Dynamic target reallocation • Gaps on ARP • Detect off-nominal behaviour • Create and monitor an actionable picture of the operational area • Create and monitor the real time Common Operational Picture (COP) • Detect changes in the environment in relation to the initial situation and initiate a revision of the initial manoeuvre • Evaluation of collateral damage • Strike coordination and armed reconnaissance • Predictive asset maintenance	• Planning aircraft actions • Planning formation/dispositive actions • Dynamic Mission Replanning • Support to Operator Decision • Prioritize threats • High level tactical assistance • Identification Friend or Foe (IFF) • Cloud/fog/edge data distribution • Automatic landing or Take Off • Securely managing MILS • Create and propose to the mission commander a mission plan • Dynamic targeting • In case of unforeseen ("unplanned") situation/enemy reaction, propose mission plan update, based on the current COP • Production of Master Air Operation Plan (MAOP) Production or handling of Air Tasking Orders (ATO) • Dynamically combine effecting and C2 capabilities • Simultaneous effect pats using the action nodes to create the desired effects on a designated target • Planning and tasking • Production or handling of Airspace Control Orders (ACO) • Air Crew Planning	• Coalition Shared Data Base (CSD) • Flying in formation • Counteract threat • End-to-End Comms • Ensuring rules of engagement • Tactical communication • High level tactical assistance • Engage • Launch weapon • Launch Defensive Aid • EMCON/stealthiness • Secure communications • Swarm strategies • Consider ethical principles • Is human approval required? • Delegation to the lowest possible level of subsidiary • Propose to the authority the appropriate level of delegation in line with sensitivity of the situation • Execution of Air Operations and mission monitor • Reporting, Logistics, Electronic logbooks including management of events and info flow

Ilustración 7: Lista preliminar de funciones considerando el ciclo OODA.

- Definir una taxonomía de las características del problema y crear una plantilla para la descripción y caracterización de las funciones CAC en el contexto del paquete de trabajo. El propósito es estandarizar la caracterización de los procesos de CAC en términos de los retos que conllevan para los métodos de IA y su certificación. La taxonomía ayudará a categorizar las dimensiones de la naturaleza distintiva de los procesos de CAC, los problemas, las restricciones, las limitaciones y las necesidades de certificación. Esta taxonomía se propone para seguir determinando cuáles son las necesidades y limitaciones de las funciones CAC SC relevantes para el proceso de certificación en la tarea, y como primer paso para determinar qué métodos de IA son los más adecuados para abordarlas.

- Identificar un conjunto reducido de funciones candidatas IA en el contexto CAC para cada fase del concepto genérico OODA (propuesto, puede ser otro similar).
  La taxonomía definida en la sección anterior se traducirá en una plantilla para la evaluación de las funciones identificadas. Se proporcionará un ejemplo cumplimentado. La lista reducida de funciones seleccionadas se describirá de acuerdo con esta plantilla.

- Proporcionar un ejemplo de evaluación para una de las funciones identificadas.

5.1.3 Ciclo OODA.

El bucle OODA se propone como marco genérico para identificar ejemplos de funciones relevantes. El bucle OODA es un proceso de cuatro pasos para tomar decisiones eficaces en situaciones de alto riesgo. Consiste en recopilar información relevante, reconocer posibles sesgos, decidir y actuar, para luego repetir el proceso con nueva información. En esta sección se describe el concepto y cómo se utiliza.

Esta sección también incluirá los criterios utilizados para filtrar y priorizar la lista de funciones.

“Thales”, explica el bucle OODA de la siguiente manera:

El bucle OODA, es una forma de restaurar la capacidad de respuesta ágil y dinámica de la toma de decisiones militares. En combate, o en cualquier forma de actividad competitiva en circunstancias inciertas, siempre nos movemos en cuatro fases:

Primero observamos las condiciones. No solo "ver", sino absorber activamente toda la situación: "recopilación de datos" en el sentido más amplio.

A continuación, nos orientamos de acuerdo con lo que hemos aprendido, estamos sintetizando información y preguntándonos 'dado lo que sabemos, ¿qué opciones tenemos disponibles?' Esta es la fase más importante porque todo está en juego: nuestro entrenamiento, nuestra experiencia, incluso nuestras expectativas culturales. Y debido a que la buena orientación "engloba", afecta a todas las demás fases: cómo observamos, decidimos y actuamos.

Después, partiendo de las opciones que tenemos a nuestra disposición, decidimos qué debemos hacer y luego actuamos. Al actuar, hacemos que las cosas cambien, dando lugar a nuevas alternativas que observar ... y entramos de nuevo en el circuito [42].

En su forma más simple, el bucle OODA se presenta así:

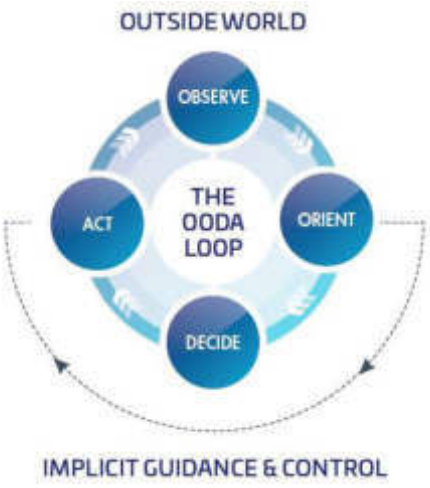


Ilustración 8: Bucle OODA [42].

Como se explica en el libro "Air power's quest for strategic paralysis", las ideas de Boyd sobre la parálisis estratégica están orientadas al proceso y apuntan a la incapacitación psicológica. Habla de replegar al adversario sobre sí mismo operando dentro de su bucle de observación-orientación-decisión-acción (OODA). Boyd ofrece esta forma de gimnasia mental diseñada para permitir una construcción más rápida de estrategias más precisas en la batalla [43].

La siguiente imagen muestra el ciclo OODA de una forma más detallada. Véase Ilustración 9.

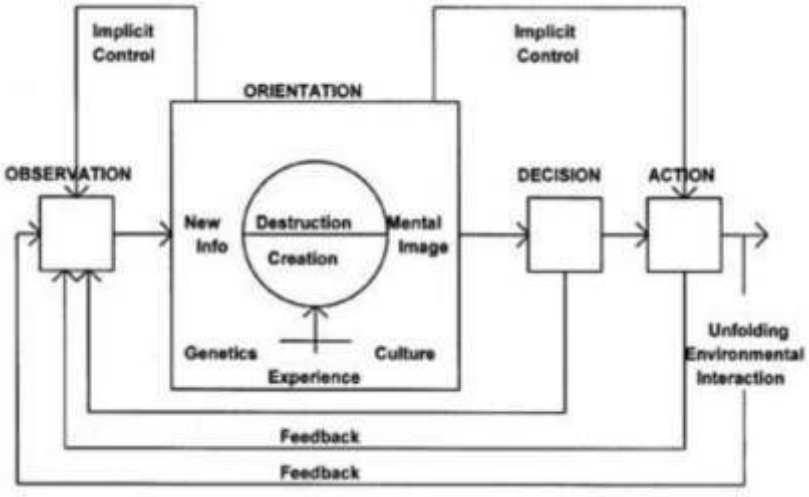


Ilustración 9: Bucle OODA [43].

Es, en palabras de Boyd, el proceso de "examinar el mundo desde varias perspectivas para poder generar imágenes mentales o impresiones que correspondan a ese mundo". Si se hace bien, es la clave para ganar en lugar de perder. Si se hace muy bien, es la marca del genio [43].

5.2 Revisión y evaluación de las normas aplicables y pertinentes.

Para llevar a cabo el trabajo, otro aspecto clave es la consideración del proceso y las normas actuales de certificación de la aviación, así como de las normas de IA ya elaboradas o en fase de desarrollo. Es necesario alinearse con el proceso de normalización actual. Para ello, hay que identificar la documentación pertinente.

Por lo explicado anteriormente, una parte del trabajo, consistirá en revisar las normas aplicables y relevantes, disponibles o en curso de producción, para determinar qué partes de dichas normas son aplicables a la certificación militar de componentes críticos de seguridad basados en IA. Esta evaluación tiene por objeto identificar los componentes existentes o en producción del futuro marco de certificación.

Por ello, las siguientes 4 áreas de regulación y estandarización se consideran muy relevantes para construir el marco de estandarización para el desarrollo, validación y certificación de componentes críticos de seguridad basados en IA del CAC, y para entender las necesidades e identificar los gaps actuales en la construcción de dicho marco:

- La normativa de aeronavegabilidad de la aviación militar aplicable al sistema CAC, incluido el marco de política, reglas, directivas, normas, procesos y la dirección, asesoramiento y orientación asociados, que rigen la actividad de la aviación militar y con arreglo a los cuales se evalúa la seguridad aérea.

- Reglamentación de la aviación civil aplicable a los ámbitos de la aviación implicados en el sistema CAC y, especialmente, la normalización de los sistemas críticos para la seguridad, el SW crítico para la seguridad y las aplicaciones de la IA en los ámbitos de la aviación civil.

- Procesos generales de normalización de la IA, ya que gran parte de los objetivos, requisitos y medios de cumplimiento relativos a la IA son horizontales y aplicables a múltiples ámbitos.

- Normas sobre sistemas autónomos, en particular para los que implican aplicaciones de IA.

Considerando los estándares recogidos en una investigación inicial (Véase *Anexo I*) pertenecientes a estas cuatro áreas y otra información aplicable, los objetivos de esta tarea consisten en examinar en profundidad toda esta documentación para:

- Identificar cuál es la base de certificación aplicable a un sistema CAC, diferenciar qué disposiciones específicas seguirán siendo aplicables a los componentes críticos de seguridad basados en IA, y cuáles podrían requerir una adaptación o ampliación adicional para dar cabida a la certificación de componentes críticos de seguridad IA.

- Identificar las necesidades específicas y distintivas de la CAC que no están cubiertas ni está previsto cubrir con los trabajos de normalización en curso o a corto plazo.

- Identificar los gaps derivadas para lograr el marco de normalización de seguridad crítica de la IA de la CAC deseado.

5.2.1 Metodología de la revisión de normas aplicables.

Se propone organizar este análisis como un proceso incremental. Se quiere combinar algunos documentos para formar la base del marco de trabajo, para posteriormente completar esta base con otros estándares adicionales de forma que se aborde todo el marco de certificación actual.

De alguna manera se busca hacer como una especie de “puzle”, en el que el marco de trabajo base (V0) constituye las piezas fundamentales y principales, y poder completar este puzle con la incorporación de otros estándares.

La metodología que se propone es ir produciendo versiones incrementales del marco teniendo en cuenta la revisión de otras normas generales de IA. Así, al revisar o analizar otros estándares con respecto a V0 se irán generando N versiones del mismo, hasta VN versiones. Véase Ilustración 10.

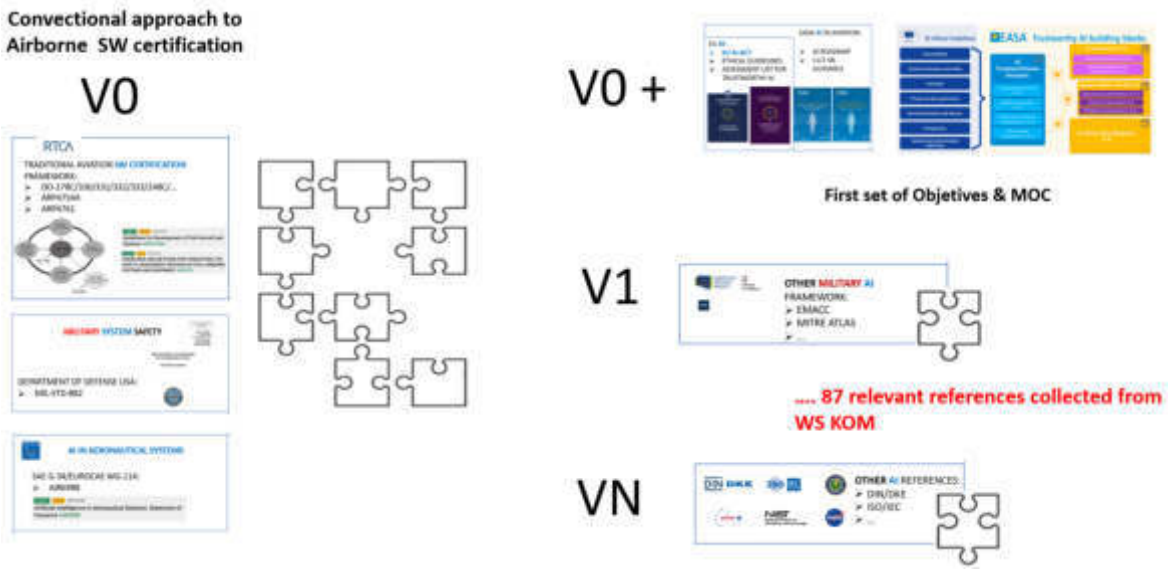


Ilustración 10: Construcción del "puzle" que representa el proceso incremental de la revisión de estándares.

Esta versión marco V0 considera la revisión de las normas actuales aplicables a la verificación, validación y certificación de SW de seguridad crítica en el contexto militar. Estas normas incluyen el Reglamento Europeo de Aeronavegabilidad Militar (EMARs) y las bases de certificación aplicables a un sistema CAC.

Además, la V0 también incluye la evaluación de las normas actuales de desarrollo de SW aeroespacial utilizadas en el proceso de certificación/aprobación de sistemas críticos para la seguridad, tanto aéreos como terrestres, en la aviación civil.

Así pues, algunas de las referencias que considerará V0 son las siguientes:

- **DO-178C**, Software Considerations in Airborne Systems and Equipment Certification.
- **MIL-HDBK-516C**, MANUAL DEL DEPARTAMENTO DE DEFENSA: CRITERIOS DE CERTIFICACIÓN DE AERONAVEGABILIDAD (12-DIC-2014)
- DIRECTRICES Y MÉTODOS PARA LLEVAR A CABO EL PROCESO DE EVALUACIÓN DE LA SEGURIDAD EN SISTEMAS Y EQUIPOS AEROTRANSPORTADOS CIVILES **ARP4761**.
- Directrices para el desarrollo de aeronaves y sistemas civiles **ARP4754A**.

- **CS 25**. Especificaciones de certificación y medios aceptables de cumplimiento para grandes aeronaves.

- **AMC-20**. Medios de cumplimiento aceptables generales para la aeronavegabilidad de productos, componentes y equipos.

- Otros

La versión V0+ del marco de certificación incluye una revisión del documento "**EASA AI Roadmap"** y del documento "**EASA AI Concept Paper"**. En las siguientes versiones (V1, V2, V3...VN), se incluirán el listado de referencias recogidas en la investigación inicial y también se considerarán otras que vayan proponiendo los expertos en el desarrollo del trabajo.

A la vista de V0+, quedarían aquellas áreas que abarca la Inteligencia Artificial y que todavía no han sido estandarizadas. Esas áreas, que son los "huecos del puzle" que faltan por completar en una versión VN, son las necesidades que hay que identificar para trabajar en tareas posteriores.

Uno de los documentos que compondrá V0, y en consecuencia VN, es el **AIR 6988**. El documento realiza un análisis de otros documentos de normativa, identificando los "gaps". Sin embargo, sólo cubre un conjunto de posibles técnicas o métodos de IA, el ML. Por lo que se conoce, aún no se ha realizado ningún análisis similar teniendo en cuenta otras ramas de la IA.

Por tanto, se identifica una necesidad a nivel general para el análisis de los estándares de referencia.

Necesidad: Análisis teniendo en cuenta otras ramas de la IA, como los métodos de IA estadística, la IA simbólica, ya sea lógica o basada en el conocimiento, los enfoques algorítmicos, la IA híbrida y las aplicaciones de aprendizaje en línea. Véase Ilustración 11.

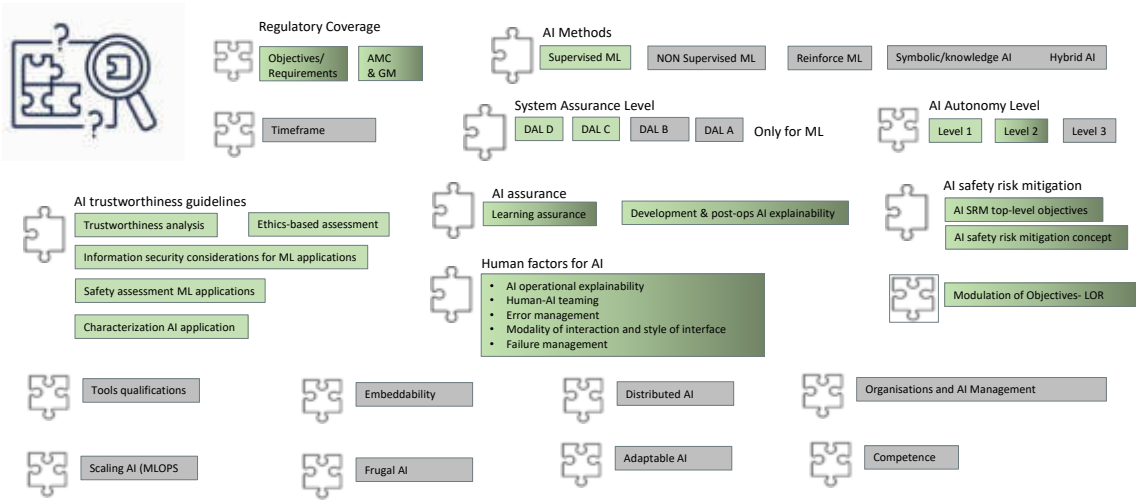


Ilustración 11: Áreas de la IA estudiadas y por estudiar al considerar VO+.

Se proporciona un esquema de áreas que componen la IA. Algunas de ellas ya se han estudiado y otras están en proceso de desarrollo o todavía no se han estudiado. En verde están las áreas que ya han sido tratadas y están reguladas por EASA u otros documentos normativos. En gris están aquellas que siguen en proceso de desarrollo o que todavía no han sido estudiadas. Esta imagen es muy útil para identificar que campos pueden componer la necesidad explicada.

5.2.2 Plantilla para el análisis de normas.

Con el fin de llevar a cabo un análisis estructurado de las referencias y normas identificadas, se elabora una plantilla. Esta plantilla pretende albergar la revisión de las referencias de forma útil, permitiendo extraer la información más relevante. En concreto, se pretende extraer de cada referencia:

- Breve resumen de los principales conceptos contenidos en la norma.
- Análisis estructurado por dimensiones de los principales conceptos recogidos en la norma en comparación con la V0.
- Lista de necesidades y gaps detectadas en el análisis de la referencia.
- Conclusiones potencialmente aplicables a otras áreas.

Por tanto, la plantilla desarrollada es fundamental para las identificaciones del bloque de construcción del marco y los elementos ya incluidos en la V0, los elementos parcialmente cubiertos en la V0 o los elementos que presenten contradicciones o incompatibilidades con V0.

La plantilla, que se utilizará para la revisión de estándares, divide el análisis en diferentes partes:

- **Información de identificación:** Para identificar de forma sencilla de qué revisión se trata, en la parte superior hay un espacio destinado a señalar:
  - Contribuidor: Organización encargada de revisar el estándar.
  - Fecha: Cuando se realizó el análisis.
  - Fuente designada: Nombre del estándar que se revisa.
  - Otras fuentes: Si se ha consultado algún otro estándar porque se cree que puede ser relevante.

- **Resumen los conceptos principales**: El objetivo de esta sección es ofrecer un breve resumen de los principales contenidos desarrollados en la referencia analizada. No se pretende desarrollar un resumen detallado de la referencia, sino únicamente situar las principales conclusiones/conceptos tratados.

- **Análisis por dimensiones de los principales conceptos en comparación con la V0 del marco de normalización:** El objetivo es realizar un análisis de la referencia bajo las dimensiones establecidas en la V0+ del marco de certificación basado en IA de la CAC. De esta forma, podremos identificar cuáles de estas dimensiones se tratan en la referencia y desde qué enfoque. Así, se puede ver las partes que se solapan o contradicen entre estándares y con V0+.

- **Necesidades y gaps identificados:** Esta sección tiene por objeto identificar qué partes de los gaps no cubiertos por V0+ están cubiertos por la referencia analizada. Permitirá identificar nuevos gaps no detectados hasta el momento. Del mismo modo, tratará de recoger las posibles necesidades identificadas respecto a V0+ con la intención de unificar conceptos que vayan en la misma línea o sobre los que sea necesario alcanzar un criterio común.

- **Conclusiones o conceptos principales potencialmente aplicables a otras áreas de trabajo:** Este último apartado pretende facilitar la identificación de partes concretas de la referencia analizada que puedan ser aplicables o útiles para otros usos o trabajos. Se trata de un apartado donde el revisor puede apuntar conceptos que encuentre relevantes para otras secciones de este extenso campo. Se pretende que el esfuerzo invertido en la revisión de las normas sea lo más práctico posible y fácilmente extrapolable a otras áreas. En la plantilla se incluyen las siguientes columnas:
  - Sección / Tarea / Paquete de Trabajo: Identificación de la sección, tarea o paquete de trabajo a la que se cree aplicable el contenido identificado.
  - Descripción: Contenido identificado.

- Motivo: Razón por la que se cree que el contenido es potencialmente aplicable a otras secciones.

- **Proposals Submission:** Por último, la parte final incorpora una tabla donde se pueden sugerir conceptos que se deberían añadir al análisis, modificar, debatir, etc. Esta parte recoge las cuestiones que no pertenecen a las otras secciones de la plantilla pero que el contribuidor asignado para el análisis cree conveniente reflejar. Se incluye la columna "SOURCE" en esta tabla por si se quieren añadir propuestas referentes a otras fuentes que son relevantes en el análisis.

A continuación, se muestra el modelo de plantilla para el análisis de estándares que se ha diseñado.

PLANTILLA COMPLETA

IDENTIFICATION INFORMATION			
CONTRIBUTOR		**DATE**	
DESIGNATED SOURCE		**OTHER SOURCES**	

Tabla 2: Información de identificación.

SUMMARY OF THE MAIN CONCEPTS

Tabla 3: Resumen de los conceptos principales.

ANALYSIS BY DIMENSIONS OF THE MAIN CONCEPTS COVERED BY THE STANDARD IN COMPARISON WITH V0 (EASA)	
Identified dimensions	The content identified in the reference with respect to that established by EASA is considered to be: applicable, sufficient, duplicative, overlapping or contradictory
Dimension 1 - **Subdimension 1**	

Add a row for each of the identified dimensions. Consider both those included in the list in section ***¡Error! No se encuentra el origen de la referencia.*** *and new dimensions identified.*	

Tabla 4: Análisis por dimensiones de los principales conceptos en comparación con VO.

LIST OF IDENTIFIED NEEDS AND GAPS
Identified Needs: **1.** **2.** **Identified Gaps:** **1.** **2.**

Tabla 5: Necesidades y "gaps" identificados.

FINDINGS OR MAIN CONCEPTS POTENTIALLY APPLICABLE TO OTHER AREAS		
Section / Task	**Description**	**Reason**

Tabla 6: Conclusiones o conceptos aplicables a otras áreas.

PROPOSALS SUBMISSION				
Type	**Section**	**Description**	**Reason**	**Source**
ADD				
MODIFY				
ERASE				
SUGGESTION				
DISCUSSION TOPIC				
…				

Tabla 7: Propuestas del análisis.

Estos análisis hechos en la plantilla quedarán en un repositorio, y se podrán consultar cuando sea necesario, evitando así repetir el trabajo ya hecho por alguien en caso de necesitar consultar el estándar. De esta manera, si se tiene que consultar un estándar, se puede tomar el análisis de la plantilla ya realizado por otro socio en lugar de tener que recurrir al documento completo.

Se recuerda que el análisis no solo incluye resúmenes de las referencias, sino que se encarga de identificar solapes, contradicciones, “gaps”, necesidades, etc.

Es cierto que si la información que se quiere buscar en la norma es muy específica probablemente no esté recogida en el análisis. Sin embargo, si no se está familiarizado con la norma será de ayuda haberse leído el análisis previamente.

De este modo, la cobertura de las distintas normas se irá añadiendo a la V0 en un proceso de revisión progresiva hasta perfilar el marco más completo.

6 CLASIFICACIÓN *SAFETY CRITICAL* PARA COMPONENTES/ FUNCIONES IA DEL CAC.

Como se explica en el apartado de objetivos, se debe dar respuesta a los retos y problemas desde la perspectiva de la aeronavegabilidad y la seguridad con respecto a los componentes críticos de seguridad basados en la IA del CAC.

La certificación de sistemas críticos para la seguridad ha seguido tradicionalmente un enfoque basado en el riesgo con la definición de niveles de garantía en función de la gravedad del efecto de un fallo. Esta sección tratará de revisar qué dimensiones adicionales deben tenerse en cuenta en la definición de los niveles de garantía en el caso de los sistemas CAC y los sistemas basados en IA. Se considerarán como punto de partida las normas vigentes en materia de aeronavegabilidad militar/civil y certificación de la seguridad.

Dado que el trabajo se realiza para componentes críticos de la seguridad, es necesario definir un criterio para separar entre los componentes que son críticos y los que no. Sin embargo, según el documento de referencia estudiado, la criticidad puede venir a raíz de diferentes factores o criterios.

En Inteligencia Artificial no solo se ha de tener en cuenta la severidad de las consecuencias potenciales, como se ha hecho tradicionalmente. Sino que también es necesario considerar criterios como el nivel de autonomía, el nivel de interacción entre el ser humano y la IA, la robustez o el nivel de seguridad de la información a la hora de definir los niveles de garantía de los componentes críticos de seguridad basados en IA para el CAC.

En el presente trabajo se intentará explicar estos criterios a partir de estándares de referencia. Además, se definirá qué adoptamos como “Safety Critical” y “non-Safety Critical” y se identificarán las posibles necesidades, las cuales son el objetivo principal de la tarea e indispensables para proceder en el futuro trabajo de certificación.

Para ello, algunos de los estándares que se tomarán como referencia para el análisis de los criterios serán **RTCA (DO-178C for certification of SW in airborne systems)**, **MIL-STD-882E System Safety Standard Practice** y **EASA Roadmap and Concept Paper**. Se han elegido principalmente estos estándares por varios motivos, aunque también se considerarán otros.

En primer lugar, es necesario estudiar la certificación de software tradicional, aspecto que cubre **RTCA**. En segundo lugar, siguiendo en el mundo civil, **EASA** cubre algunos criterios en materia de certificación de Inteligencia Artificial. Por último, el estándar militar de debe tener en cuenta en el ámbito CAC. Este análisis inicial del esquema de clasificación se amplía con las dimensiones consideradas por otras fuentes pertinentes.

Como resultado, la necesidad de adaptar los actuales esquemas de seguridad y criticidad al contexto y las necesidades específicas de los sistemas basados en IA para CAC se ha recogido en un conjunto de necesidades que se explicarán más adelante.

6.1 Revisión y evaluación de las normas aplicables y pertinentes.

El primer paso, antes de adentrarse en el análisis de los esquemas de clasificaciones de los documentos de los que se partirá, se quiere estudiar la definición o concepto de “Safety” y “Safety Critical”. No obstante, el término "Safety Critical" referido a un componente o función no se define en algunos de los distintos documentos normativos y reglamentos de aviación civil y militar**.**

Una definición general de seguridad es la "ausencia de aquellas condiciones que pueden causar la muerte, lesiones, enfermedades, daños o pérdidas de equipos o bienes o daños medioambientales" [44].

El Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) define software crítico para la seguridad como: "software cuyo uso en un sistema puede dar lugar a un riesgo inaceptable. El software de seguridad crítica incluye software cuyo funcionamiento o fallo de funcionamiento puede conducir a un estado peligroso, software destinado a recuperarse de estados peligrosos y software destinado a mitigar la gravedad de un accidente".

El Estándar de Seguridad del Software publicado por la Administración Nacional de Aeronáutica y del Espacio de EE.UU. (NASA) identifica el software como crítico para la seguridad si se cumple al menos uno de los siguientes criterios: 1. Reside en un sistema crítico para la seguridad (determinado por un análisis de riesgos) y al menos causa o contribuye a un peligro, proporciona control o mitigación de peligros, controla funciones críticas para la seguridad, procesa comandos o datos críticos para la seguridad, mitiga los daños en caso de peligro o reside en el mismo sistema (procesador) que el software crítico para la seguridad. 2. Proporciona verificación o validación total o parcial de sistemas críticos para la seguridad, incluidos sistemas de hardware o software.

A partir de estas definiciones, se podría deducir que el software por sí mismo no es ni seguro ni inseguro; sin embargo, cuando forma parte de un sistema crítico para la seguridad, puede causar o contribuir a condiciones inseguras. Este tipo de software se considera crítico para la seguridad [44].

Si nos centramos en el **MIL-STD-882E,** sí se proporciona una definición “Safety Critical” exacta. Se definen los siguientes conceptos:

- “**Safety**. Freedom from conditions that can cause death, injury, occupational illness, damage to or loss of equipment or property, or damage to the environment (MIL-STD-882E 3.2.30).

- **Safety-critical.** A term applied to a condition, event, operation, process, or item whose mishap severity consequence is either **Catastrophic** or **Critical** (e.g., safety-critical function, safety-critical path, and safety-critical component) (MIL-STD- 882E 3.2.31).

- **Safety-critical function (SCF).** A function whose failure to operate or incorrect operation will directly result in a mishap of either **Catastrophic or Critical** severity (MIL-STD- 882E 3.2.32).

- **Safety-critical item (SCI).** A hardware or software item that has been determined through analysis to potentially contribute to a hazard with Catastrophic or Critical mishap potential, or that may be implemented to mitigate a hazard with **Catastrophic or Critical** mishap potential (MIL-STD- 882E 3.2.33).

- **Safety-related**. A term applied to a condition, event, operation, process, or item whose mishap severity consequence is either **Marginal or Negligible**. (MIL-STD- 882E 3.2.24).

- **Safety-significant.** A term applied to a condition, event, operation, process, or item that is identified as either safety-critical or safety-related (MIL-STD- 882E 3.2.35).

- **Severity**. The magnitude of potential consequences of a mishap to include: death, injury, occupational illness, damage to or loss of equipment or property, damage to the environment, or monetary loss (MIL-STD- 882E 3.2.36)".

EASA utiliza el término "componente crítico" en varias especificaciones de certificación. Sin embargo, el término no siempre está definido, y no existe una definición general porque depende del contexto en el que se utilice.

RTCA (en los documentos **DO-178C** o **DO-278A,** para certificación de SW en aviación), tampoco proporciona una definición de "Safety Critical". La situación es similar en otras normas de SW de aviación, como EUROCAE ED-109A y sus suplementos ED-153, ED-76A.

6.2 Nivel de severidad. Clasificación.

Las normas de seguridad establecen esquemas de severidad para clasificar los efectos de las condiciones de fallo a nivel de un sistema.

Se analizarán los criterios de severidad que contemplan los distintos estándares por ser la práctica tradicional en certificación. Aunque los estándares, además de la severidad, contemplen otros criterios, éstos se explicarán más adelante.

6.2.1 MIL-STD-882. Práctica habitual en materia de seguridad de los sistemas

Se empezará estudiando el documento militar, MIL-STD-882E. El documento proporciona definiciones para describir cada categoría de severidad. Los niveles o categorías de severidad están clasificados como en la Ilustración 13.

TABLE I. Severity categories

SEVERITY CATEGORIES		
Description	**Severity Category**	**Mishap Result Criteria**
Catastrophic	1	Could result in one or more of the following: death, permanent total disability, irreversible significant environmental impact, or monetary loss equal to or exceeding $10M.
Critical	2	Could result in one or more of the following: permanent partial disability,injuries or occupational illness that may result in hospitalization of at least three personnel, reversible significant environmental impact, or monetary loss equal to or exceeding $1M but less than $10M.
Marginal	3	Could result in one or more of the following: injury or occupational illness resulting in one or more lost work day(s), reversible moderate environmental impact, or monetary loss equal to or exceeding $100K but less than $1M.
Negligible	4	Could result in one or more of the following: injury or occupational illness not resulting in a lost work day, minimal environmental impact, or monetary loss less than $100K.

Ilustración 12: Descripción de las categorías de severidad. MIL-STD-882E.

Centrándose en las definiciones de los conceptos generales del apartado anterior y en las de las categorías de severidad de la Ilustración 13, se podría hacer un esquema que resuma, relacione y permita visualizar como se clasifica cada nivel de severidad dentro de los conceptos explicados al principio. Véase Ilustración 14.

Este esquema podría ser una primera aproximación de qué se considera como “Safety Critical”.

Ilustración 13: Esquema resumen relacionando las categorías de severidad con los conceptos definidos.

Los riesgos evaluados se expresan como un Código de Evaluación de Riesgos (CER), que es una combinación de una categoría de gravedad y un nivel de probabilidad [40]. Lo que proporciona la matriz de riesgo siguiente:

TABLE III. Risk assessment matrix

RISK ASSESSMENT MATRIX				
SEVERITY / PROBABILITY	Catastrophic (1)	Critical (2)	Marginal (3)	Negligible (4)
Frequent (A)	High	High	Serious	Medium
Probable (B)	High	High	Serious	Medium
Occasional (C)	High	Serious	Medium	Low
Remote (D)	Serious	Medium	Medium	Low
Improbable (E)	Medium	Medium	Medium	Low
Eliminated (F)	Eliminated			

Ilustración 14: Matriz de riesgo. MIL-STD-882E.

Lo explicado en este subapartado se refiere a los criterios del documento militar en cuanto a nivel de severidad. Como se ha explicado anteriormente, más adelante se tratarán otros criterios de este estándar y se compararán con los otros estándares.

6.2.2 ARP 7554/4761 – RTCA DO 178C

De la misma manera que se ha hecho con el documento militar, se expone la clasificación de severidad empleada junto con las definiciones para cada nivel. Para cada nivel se establece una probabilidad determinada y un “Assurance Level”.

Failure Condition Severity, Probabilities, and Levels

Severity Classification	Potential Failure Condition Effect	Likelihood of Occurrence	Exposure Per Flight Hour (Part 25)	Assurance Level
Catastrophic	Failure conditions, which would result in multiple fatalities, usually with the loss of the airplane	Extremely improbable	1E–9	A
Hazardous/ Severe major	Failure conditions, which would reduce the capability of the airplane or the ability of the flight crew to cope with adverse operating conditions to the extent that there would be • A large reduction in safety margins or functional capabilities • Physical distress or excessive workload such that the flight crew cannot be relied upon to perform their tasks accurately or completely • Serious or fatal injury to a relatively small number of the occupants other than the flight crew	Extremely remote	1E–7	B
Major	Failure conditions, which would reduce the capability of the airplane or the ability of the crew to cope with adverse operating conditions to the extent that there would be a significant reduction in safety margins or functional capabilities, a significant increase in crew workload or in conditions impairing crew efficiency, discomfort to the flight crew, or physical distress to passengers or cabin crew, possibly including injuries	Remote	1E–5	C

Severity Classification	Potential Failure Condition Effect	Likelihood of Occurrence	Exposure Per Flight Hour (Part 25)	Assurance Level
Minor	Failure conditions, which would not significantly reduce airplane safety and which involve crew actions that are well within their capabilities. Minor failure conditions may include a slight reduction in safety margins or functional capabilities; a slight increase in crew workload, such as routine flight plan changes; or some physical discomfort to passengers or cabin crew	Reasonably probable	1E–3	D
No safety effect	Failure conditions that would have no effect on safety; e.g., failure conditions that would not affect the operational capability of the airplane or increase crew workload	Probable	1.0	E

Ilustración 15: Descripción de las categorías de severidad. RTCA DO-178C.

RTCA clasifica el software en cinco niveles de criticidad. DO-178C define los niveles de software “A”, “B”, “C”, “D” y “E” que corresponden a las categorías de una condición de fallo de software de "Catastrófico", "Grave", "Mayor", "Menor" y "Sin efecto sobre la seguridad", respectivamente.

A cada nivel de software se le asignan niveles de garantía de desarrollo (DAL), y los objetivos y requisitos de independencia aplicables varían en función del nivel.

Es sencillo cerciorarse de que la clasificación de severidad de las consecuencias entre el MIL-STD-882E y el RTCA se asemeja. Se podría incluso hacer una comparativa para desarrollar una combinación entre ambas clasificaciones. Véase Ilustración 17.

6.2.3 Certificación de SW tradicional. Comparación MIL-STD_882E/RTCA DO 178C

		MIL-STD-882 Severity condition: potential consequences or damages to human life, damage to or loss of equipment or property, damage to the environment, or monetary loss.		ARP 7554/4761 - RTCA DO 178C Severity condition: effect on the aircraft, crew and occupants to determine the associated severity classification considering crew awareness, flight phase, environmental and operational conditions.	
SAFETY RELATED	SAFETY CRITYCAL AI COMPONENTS	**Catastrophic**	Could result in one or more of the following: death, permanent total disability, irreversible significant environmental impact, or monetary loss equal to or exceeding $10M.	**Catastrophic**	Failure conditions, which would result in multiple fatalities, usually with the loss of the airplane.
		Critical	Could result in one or more of the following: permanent partial disability, injuries or occupational illness that may result in hospitalization of at least three personnel, reversible significant environmental impact, or monetary loss equal to or exceeding $1M but less than $10M.	**Hazardous Severe**	Failure conditions, ...would be: • A large reduction in safety margins or functional capabilities. • Physical distress or excessive workload such that the flight crew cannot be relied upon to perform their tasks accurately or completely. • Serious or fatal injury to a relatively small number of the occupants other than the flight crew.
	NON-SAFETY CRITYCAL AI COMPONENTS	**Marginal**	Could result in one or more of the following: injury or occupational illness resulting in one or more lost work day(s), reversible moderate environmental impact, or monetary loss equal to or exceeding $100K but less than $1M.	**Major**	Failure conditions, ... would be a significant reduction in safety margins or functional capabilities, a significant increase in crew workload or in conditions impairing crew efficiency, discomfort to the flight crew, or physical distress to passengers or cabin crew, possibly including injuries.
		Negligible	Could result in one or more of the following: injury or occupational illness not resulting in a lost work day, minimal environmental impact, or monetary loss less than $100K.	**Minor**	Failure conditions, ... would not significantly reduce airplane safety and which involve crew actions that are well within their capabilities. ..... may include a slight reduction in safety margins or functional capabilities; a slight increase in crew workload, such as routine flight plan changes; or some physical discomfort to passengers or cabin crew.
		Non Safety related: No negative potential consequences on safety			

Ilustración 16: Esquemas de clasificación de severidad en MIL-STD-882E y RTCA.

Atendiendo a las definiciones de las consecuencias potenciales para cada nivel, se pueden correlacionar ambos estándares correspondiéndose los niveles "Catastrophic, Severe Major, Major y Minor" del RTCA con los niveles "Catastrophic, Critical, Marginal y Negligible" del MIL-STD-882E, respectivamente. Adaptando la Ilustración 18 se puede apreciar esta comparativa.

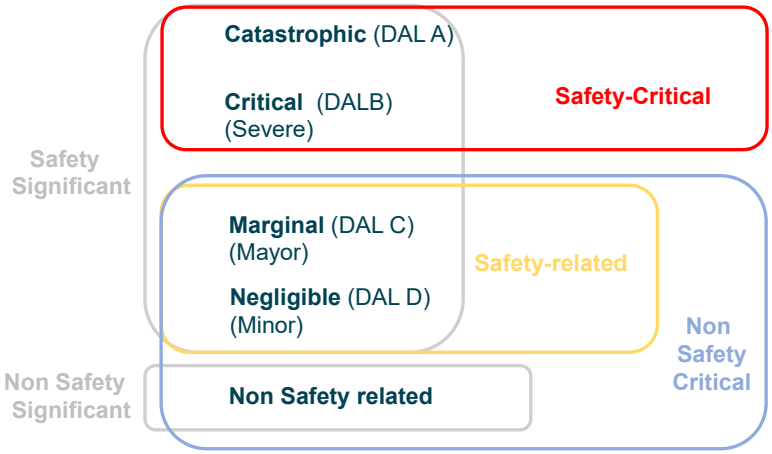


Ilustración 17: Definición del concepto "Safety Critical" en MIL-STD-882E y RTCA.

No obstante, ya se ha mencionado que, así como el documento militar define qué es un elemento o función crítica para la seguridad, explicando que solo aborda los casos de peligro potencial catastrófico o crítico; el RTCA no habla de criticidad para la seguridad según las consecuencias, sino que asume que todo es crítico para la seguridad excepto el último nivel, "No safety effect" en el que no hay efectos de seguridad en las consecuencias.

Por tanto, aquí existe una discrepancia entre ambos estándares. La Ilustración 18 es una propuesta para resolver esta cuestión, asumiendo que los dos niveles superiores son los considerados como "Safety Critical". Sin embargo, esto no es tan sencillo debido a los objetivos que hay que cumplir en el proceso de certificación. Hay objetivos en los niveles de severidad inferiores que se deben entender para cumplir los niveles de severidad superiores. Esto se verá en el siguiente apartado.

Aunque la definición "Safety Critical" pertenece al apartado anterior, es necesario comprender los esquemas de severidad para poder entender la propuesta de la Ilustración 18. Es por esta razón, que este apartado acaba combinándose con el anterior.

6.3 Nivel de rigor para demostrar el cumplimiento en el proceso de garantía.

En un proceso de certificación existen una serie de requisitos que se tienen que cumplir para poder garantizar un elevado porcentaje de fiabilidad. Con respecto al desarrollo de software tradicional, en el mundo civil, RTCA detalla unos objetivos en función de la severidad; en el mundo militar, MIL-STD-882E, habla de tareas. Estos objetivos y tareas son los requisitos que garantizan que un componente tiene la capacidad de ejercer la función para la cual se ha diseñado.

Por otro lado, EASA establece unos objetivos para los niveles "non-Safety Critical" de IA, sin embargo, los objetivos para niveles de IA "Safety Critical" todavía no se han desarrollado.

6.3.1 ARP 7554/4761-RTCA DO 178C

En cuanto al nivel de rigor, RTCA fija un número de objetivos que se requiere cumplir para cada *"Design Assurance Level*". El DAL viene dado por la categorización de las consecuencias potenciales. Véase Ilustración 19.

Category	Failure Condition Description	Design Assurance Level (DAL)	Number of Required DO-178C Process Objectives
Catastrophic	Failure condition results in multiple fatalities with probable loss of aircraft	A	71/30
Severe	Failure condition would significantly reduce the ability of the crew and/or aircraft capabilities required to compensate for adverse operating conditions	B	69/38
Major	Failure condition would reduce the ability of the crew and/or aircraft capabilities needed to compensate for adverse operating conditions	C	62/5
Minor	Failure condition has no significant impact on safety margins or crew workload	D	26/2
No Safety Effect	Failure condition has no impact on safety	E	0/0

Ilustración 18: Niveles de garantía basados en la gravedad/criticidad. RTCA DO 178C.

Es decir, una consecuencia severa tendrá un DAL alto, y por tanto un mayor número de objetivos que cumplir. El número de objetivos en el proceso de certificación disminuye con la criticidad. Véase Ilustración 20.

Ilustración 19: DAL con el número de objetivos a cumplir [44].

Un ejemplo de un caso real puede ser el siguiente: Suponiendo que un accidente causara "múltiples víctimas mortales" estaríamos en el nivel Catastrófico, y, por tanto, el "Design *Assurance Level*" sería A, el más alto posible. DAL A es el más crítico y el que mayor número de objetivos ha de cumplir en el proceso de certificación. En cambio, si se produjese una "ligera reducción de los márgenes de seguridad" nos situaríamos en el nivel Menor, teniendo un DAL D, y con un menor número de objetivos. Véase Ilustración 20.

6.3.2 MIL-STD-882E

El documento militar, por su parte, utiliza un criterio distinto. Para establecer el nivel de rigor combina el nivel de control de software con la criticidad.

El MIL-STD-882E desarrolla una categorización en función del nivel de control que tiene el software, siendo el nivel 1 el de máxima autonomía en las funciones de seguridad y el 5 el nivel en qué las funciones que desempeñan los componentes de IA no tienen ningún impacto sobre la seguridad. Véase Ilustración 21. Además, para cada nivel de control del software proporciona una descripción.

TABLE IV. Software control categories

SOFTWARE CONTROL CATEGORIES		
Level	**Name**	**Description**
1	**Autonomous (AT)**	• Software functionality that exercises autonomous control authority over potentially safety-significant hardware systems, subsystems, or components without the possibility of predetermined safe detection and intervention by a control entity to preclude the occurrence of a mishap or hazard. *(This definition includes complex system/software functionality with multiple subsystems, interacting parallel processors, multiple interfaces, and safety-critical functions that are time critical.)*
2	**Semi-Autonomous (SAT)**	• Software functionality that exercises control authority over potentially safety-significant hardware systems, subsystems, or components, allowing time for predetermined safe detection and intervention by independent safety mechanisms to mitigate or control the mishap or hazard. *(This definition includes the control of moderately complex system/software functionality, no parallel processing, or few interfaces, but other safety systems/mechanisms can partially mitigate. System and software fault detection and annunciation notifies the control entity of the need for required safety actions.)* • Software item that displays safety-significant information requiring immediate operator entity to execute a predetermined action for mitigation or control over a mishap or hazard. Software exception, failure, fault, or delay will allow, or fail to prevent, mishap occurrence. *(This definition assumes that the safety-critical display information may be time-critical, but the time available does not exceed the time required for adequate control entity response and hazard control.)*
3	**Redundant Fault Tolerant (RFT)**	• Software functionality that issues commands over safety-significant hardware systems, subsystems, or components requiring a control entity to complete the command function. The system detection and functional reaction includes redundant, independent fault tolerant mechanisms for each defined hazardous condition. *(This definition assumes that there is adequate fault detection, annunciation, tolerance, and system recovery to prevent the hazard occurrence if software fails, malfunctions, or degrades. There are redundant sources of safety-significant information, and mitigating functionality can respond within any time-critical period.)* • Software that generates information of a safety-critical nature used to make critical decisions. The system includes several redundant, independent fault tolerant mechanisms for each hazardous condition, detection and display.
4	**Influential**	• Software generates information of a safety-related nature used to make decisions by the operator, but does not require operator action to avoid a mishap.
5	**No Safety Impact (NSI)**	• Software functionality that does not possess command or control authority over safety-significant hardware systems, subsystems, or components and does not provide safety-significant information. Software does not provide safety-significant or time sensitive data or information that requires control entity interaction. Software does not transport or resolve communication of safety-significant or time sensitive data.

Ilustración 20: Categorías del nivel de control del software. MIL-STD-882E.

La tabla de Categorías de Severidad cruzada con la de Categorías de Control de SW proporciona la matriz "Software Safety Criticality Matrix", que determina el Nivel de Rigor en la verificación y desarrollo de SW.

TABLE V. Software safety criticality matrix

SOFTWARE SAFETY CRITICALITY MATRIX				
	SEVERITY CATEGORY			
SOFTWARE CONTROL CATEGORY	Catastrophic (1)	Critical (2)	Marginal (3)	Negligible (4)
1	SwCI 1	SwCI 1	SwCI 3	SwCI 4
2	SwCI 1	SwCI 2	SwCI 3	SwCI 4
3	SwCI 2	SwCI 3	SwCI 4	SwCI 4
4	SwCI 3	SwCI 4	SwCI 4	SwCI 4
5	SwCI 5	SwCI 5	SwCI 5	SwCI 5

SwCI	Level of Rigor Tasks
SwCI 1	Program shall perform analysis of requirements, architecture, design, and code; and conduct in-depth safety-specific testing.
SwCI 2	Program shall perform analysis of requirements, architecture, and design; and conduct in-depth safety-specific testing.
SwCI 3	Program shall perform analysis of requirements and architecture; and conduct in-depth safety-specific testing.
SwCI 4	Program shall conduct safety-specific testing.
SwCI 5	Once assessed by safety engineering as Not Safety, then no safety specific analysis or verification is required.

Ilustración 21: Índices de criticidad del software (SWCI) basados en una combinación de gravedad/criticidad y Autonomía o grado de control que el SW ejerce sobre el HW. MIL-STD-882E.

Como ya se ha mencionado, el MIL-STD-882E combina el nivel de autonomía en las funciones críticas para la seguridad con el nivel de severidad o criticidad de las consecuencias potenciales. De esta manera, se establece un nivel de rigor de las tareas, siendo SWCI 1 el nivel más riguroso y SWCI 5 el nivel menos riguroso. Con el SWCI se determina el LOR de las tareas detalladas en el documento, es decir, la exigencia requerida. Véase Ilustración 22.

Se puede apreciar que, en comparación con el RTCA, el MIL-STD-882E considera la autonomía además de la criticidad para establecer los niveles de rigor. Por tanto, el objetivo que se busca es encontrar la manera de homogeneizar estos criterios entre sí, y al mismo tiempo intentar cohesionarlos con otros de diferentes estándares, de manera que sean coherentes.

6.4 Introducción de IA. MIL-STD-882F y EASA

6.4.1 MIL-STD-882F

El MIL-STD-882E habla de autonomía, pero las versiones preliminares del MIL-STD-882F son las que introducen Inteligencia Artificial. Se trata de un documento elaborado a partir del MIL-STD-882E, pero incorpora comentarios de expertos en la materia sobre qué se debería modificar al incluir IA. Este documento, junto con EASA que se explicará más adelante, deben considerarse, además de los ya mencionados.

La Inteligencia Artificial está cada vez más presente en el diseño de sistemas. Parte de la IA es la capacidad del sistema para aprender, o aprendizaje automático. Sería necesario estudiar las tecnologías de IA que se emplearán para poder clasificarlas en diferentes niveles de estratificación que dan cuenta de los diferentes grados de aprendizaje de los sistemas [45].

Esta clasificación la propone el MIL-STD-882F, que contempla la introducción de una nueva categoría de control de IA/aprendizaje automático o SAII y un nuevo índice de control de inteligencia artificial o AICI, que serán un reflejo del constructo de categoría de control de software ya existente, pero con contenidos/objetivos diferentes.

Esta nueva categoría incluye una plantilla para la categorización específica de Inteligencia Artificial. Véase Ilustración 23. Esta nueva clasificación no pretende duplicar el apartado que define las categorías de control de software (autonomía), sino centrarse en los aspectos de la IA y el aprendizaje automático que van más allá del "software tradicional" [45].

TABLE V: Artificial Intelligence Categories

Artificial Intelligence Categories		
Level	**Name**	**Description**
1	Add categories	• Add definitions
2	TBD	•
3	TBD	•
4	TBD	•
5	TBD	•
6	TBD	•
7	TBD	•
8	TBD	•
9	TBD	•
10	No AI/Machine Learning Incorporated	The system design does not possess AI or Machine Leaning in its design.

Ilustración 22: Categorías de Inteligencia Artificial. MIL-STD-882F.

Por tanto, el Nivel 10 en esta categorización de IA correspondería al nivel en que no se incorpora IA ni aprendizaje automático. En el Nivel 1, en cambio, la incorporación de IA sería máxima. Se trata de una clasificación que determina en qué medida se incluye la Inteligencia Artificial.

Llegados a este punto, puede ser necesario hacer alguna distinción entre el software que captura el aprendizaje automático frente al software que ejecuta el comportamiento aprendido. Un enfoque más convencional/determinista para el software de aprendizaje y un enfoque más probabilístico para el software que ejecuta el aprendizaje [45].

El AICI se utilizará para determinar el LOR de las actividades de garantía de seguridad del software que deben imponerse al software de IA [45]. Correlacionando los resultados de las tablas de categorías de severidad y de categorías de IA planteadas, se obtiene una designación AICI. Véase Ilustración 24.

TABLE VII: Artificial Intelligence Criticality Matrix

ARTIFICIAL INTELLIGENCE CRITICALITY MATRIX				
	SEVERITY CATEGORY			
AI / MACHINE LEARNING CONTROL CATEGORY	Catastrophic (1)	Critical (2)	Marginal (3)	Negligible (4)
1	AICI 1	AICI 1	AICI 3	AICI 4
2	AICI 1	AICI 2	AICI 3	AICI 4
...	...	...	...	...
#	AICI #	AICI #	AICI #	AICI #

Ilustración 23: Matriz IA de criticidad. MIL-STD-882F.

Cada nivel SWCI o AICI corresponde a una designación SW LOR o AI LOR para la unidad de software designada. Véase Ilustración 25. Las actividades de garantía de las LOR de AI/Aprendizaje Automático deben realizarse además de las especificadas en las LOR de SW.

TABLE VIII. Level of Rigor Activities

LEVEL OF RIGOR ACTIVITIES				
SwCI 1	SW LOR 1		AICI 1	AI LOR 1
SwCI 2	SW LOR 2		AICI 2	AI LOR 2
SwCI 3	SW LOR 3		…	….
SwCI 4	SW LOR 4		AICI #	AI LOR #

Ilustración 24: Level of Rigor Activities. MIL-STD-882F.

Se está asumiendo que no existen interdependencias entre niveles o IA. Si esta suposición no fuera cierta, entonces cada LOR de IA será una lista distinta de actividades frente a la lista en cascada utilizada en las LOR de SW. El software que no implique IA o Aprendizaje Automático no impondrá actividades LOR adicionales.

Finalmente, MIL-STD-882F, relaciona los niveles de SWCI y AICI con un "Assurance Level". Véase Ilustración 26.

Table IX: Software Safety Assurance Risk

SWCI	AICI	Software Safety Assurance Risk Level	Para 4.4.8.1/4.4.8.2 Non-Compliance Risk Acceptance Authority
I	I	High	SAE/CAE
II	II	Serious	PEO or Designated Equivalent
III	III	Medium/Low	PM
IV	IV	Not Safety	PM

Ilustración 25: Software Safety Assurance Risk. MIL-STD-882F.

Todas los conceptos y tablas explicadas de los documentos MIL-STD-882E y MIL-STD-882F quedan resumidos en el siguiente esquema. Véase Ilustración 27. En el esquema, la parte izquierda corresponde al desarrollo del MIL-STD-882E mientras que la parte derecha pertenece a lo introducido por el MIL-STD-882F.

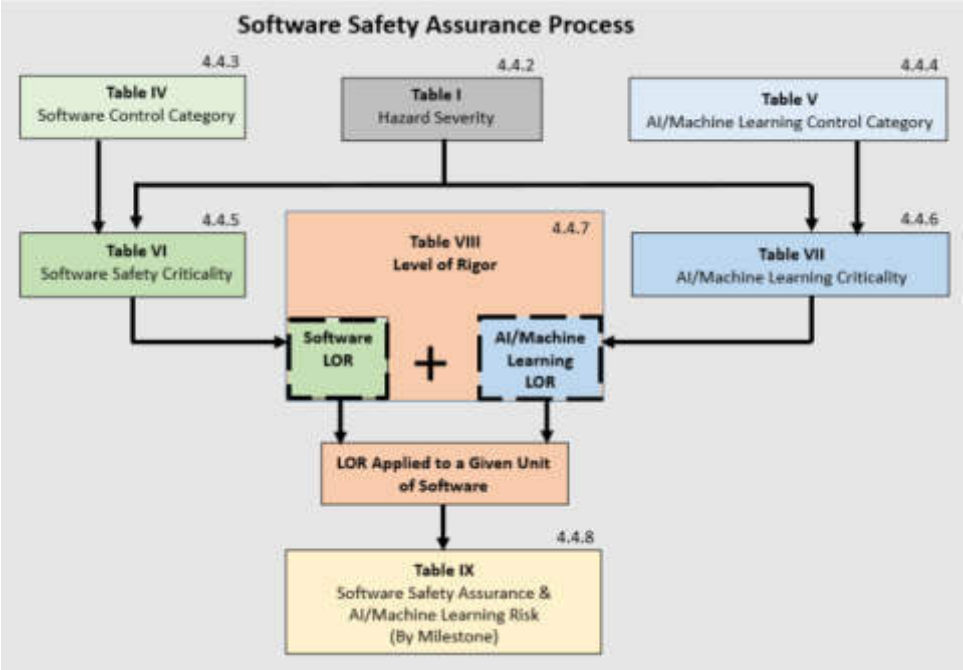


Ilustración 26: Level of Rigor process overview. MIL-STD-882F.

La autonomía se considera una dimensión relevante para establecer el nivel de garantía y los LOR para el desarrollo y la validación de SW en las normas militares, y, por lo tanto, también es una dimensión relevante a tener en cuenta en los niveles de garantía.

La autonomía es una dimensión relevante para la certificación de la IA, no solo en las normas militares sino también a nivel general, ya que ésta se percibe como un habilitador necesario para mayores niveles de automatización.

La importancia del concepto de autonomía en relación con los componentes de la IA ha sido reconocida por los trabajos iniciales relativos a la certificación de la IA en la aviación, en particular el trabajo ya realizado por la EASA en su hoja de ruta y su documento conceptual.

6.4.2 EASA

EASA introduce una clasificación de las aplicaciones de IA en niveles de IA en función del nivel de autonomía y de las interacciones entre el ser humano y la IA, y establece actividades de garantía de diferencias para cada uno de esos niveles de IA. “EASA AI Roadmap” identifica tres tipos principales de aplicaciones de IA. Éstas son la asistencia humana (Nivel 1A y 1B), el trabajo en equipo entre humanos e IA (Nivel 2A y 2B) y la automatización avanzada (Nivel 3 de IA). Véase Ilustración 28.

AI level	Function allocated to the system to contribute to the high-level task	Authority of the end user
Level 1A Human augmentation	Automation support to information acquisition	Full
	Automation support to information analysis	Full
Level 1B Human assistance	Automation support to decision-making	Full
Level 2A Human-AI cooperation	Overseen and overridable automatic decision	Full
	Overseen and overridable automatic action implementation	Full
Level 2B Human-AI collaboration	Overseen and overridable automatic decision	Partial
	Overseen and overridable automatic action implementation	Partial
Level 3A Supervised advanced automation	Supervised automatic decision	Upon alerting
	Supervised automatic action implementation	Upon alerting
Level 3B Autonomous AI	Non-supervised automatic decision	Not applicable
	Non-supervised automatic action implementation	Not applicable

Ilustración 27: Niveles IA. EASA Concept Paper.

El documento "EASA Concept Paper" presenta la aplicabilidad de los objetivos a cada nivel de IA (es decir, los Niveles 1A, 1B, 2A y 2B), y se completará en una fase posterior con consideraciones relativas al Nivel 3 de IA. Lógicamente, la mayor exigencia en la aplicabilidad de los objetivos se dará en el Nivel 3.

"EASA Concept Paper" también considera la necesidad de una nivelación adicional basada en el riesgo de los objetivos relacionados con la seguridad de la información en determinados ámbitos de la aviación y prevé que a las medidas de seguridad de la información se les pueda asignar también un nivel de garantía de la seguridad (SAL).

EASA reconoce el concepto de proporcionalidad y modulación en el nivel de rigor aplicado a los componentes de IA, de modo que los objetivos y requisitos de las aplicaciones de IA se regirán por diferentes dimensiones o criterios. EASA identifica inicialmente 3 dimensiones o criterios principales para anticipar la proporcionalidad en los objetivos de certificación y, por tanto, influir en el Nivel de Rigor que será aplicable a los componentes de IA:

- **Autonomía:** el nivel de IA como resultado de la caracterización de la aplicación de la IA.

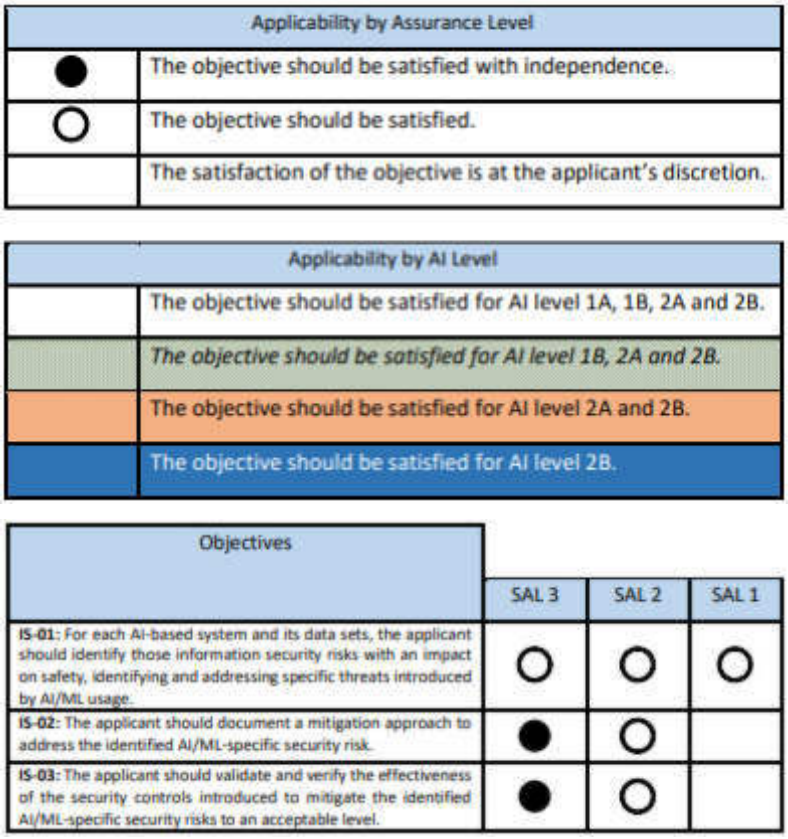
- **Safety Assurance:** el nivel de garantía o criticidad de la aplicación como resultado de la evaluación de la seguridad. Se refiere al nivel de garantía de desarrollo (DAL) para la aeronavegabilidad inicial y continuada o las operaciones aéreas, o al nivel de garantía del

software (SWAL) para la gestión del tránsito aéreo/servicios de navegación aérea (ATM/ANS).

- **Security Assurance:** el nivel de garantía de seguridad (SAL) específico de la aplicación como resultado de las evaluaciones de seguridad. Véase AMC 20-42 (ámbito de la certificación de productos). EASA solo recoge 3 objetivos referentes al SAL.

Por tanto, EASA no define lo que es crítico para la seguridad y lo que no. EASA presenta unos objetivos necesarios en función del "Assurance Level" (nivel de garantía de desarrollo o software), del "AI Level" (autonomía) y "Security Assurance Level" (nivel de seguridad de la información). Véase Ilustración 29.

No obstante, contrariamente al MIL-STD-882E, EASA trata la aplicabilidad de los objetivos con independencia, es decir, no existe una combinación de los tres criterios para determinar el Nivel de Rigor en la aplicabilidad de los objetivos.



Applicability by Assurance Level	
●	The objective should be satisfied with independence.
○	The objective should be satisfied.
	The satisfaction of the objective is at the applicant's discretion.

Applicability by AI Level	
	The objective should be satisfied for AI level 1A, 1B, 2A and 2B.
	The objective should be satisfied for AI level 1B, 2A and 2B.
	The objective should be satisfied for AI level 2A and 2B.
	The objective should be satisfied for AI level 2B.

Objectives	SAL 3	SAL 2	SAL 1
IS-01: For each AI-based system and its data sets, the applicant should identify those information security risks with an impact on safety, identifying and addressing specific threats introduced by AI/ML usage.	○	○	○
IS-02: The applicant should document a mitigation approach to address the identified AI/ML-specific security risk.	●	○	
IS-03: The applicant should validate and verify the effectiveness of the security controls introduced to mitigate the identified AI/ML-specific security risks to an acceptable level.	●	○	

Ilustración 28: Dimensiones o criterios para determinar el Nivel de Rigor aplicable. EASA.

Con la intención de entender adecuadamente lo explicado sobre el documento de EASA, a continuación, se muestran algunos de los objetivos que proporciona el documento con los criterios que determinan el Nivel de Rigor que se debe cumplir.

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
Trustworthiness analysis	**CO-04:** The applicant should define and document the ConOps for the AI-based system, including the task allocation pattern between the end user(s) and the AI-based system. A focus should be put on the definition of the OD and on the capture of specific operational limitations and assumptions.	○	○	○	○	○
	CO-05: The applicant should document how end users' inputs are collected and accounted for in the development of the AI-based system.	○	○	○	○	○
	CO-06: The applicant should perform a functional analysis of the system.	○	○	○	○	○
	CL-01: The applicant should classify the AI-based system, based on the levels presented in Table 2, with adequate justifications.	○	○	○	○	○
	SA-01: The applicant should perform a safety (support) assessment for all AI-based (sub)systems, identifying and addressing specificities introduced by AI/ML usage.	●	●	○	○	○
	ICSA-01: The applicant should identify which data needs to be recorded for the purpose of supporting the continuous safety assessment .	●	●	○	○	○
	ICSA-02: The applicant should use the collected data to perform a continuous safety assessment. This includes: — the definition of target values, thresholds and evaluation periods to guarantee that design assumptions hold; — the monitoring of in-service events to detect potential issues or suboptimal performance trends that might contribute to safety margin erosion, or, for non-ATS providers, to service performance degradations; and — the resolution of identified shortcomings or issues.	●	●	○	○	○

Ilustración 29: Objetivos del bloque "Trustworthiness analysis". EASA Concept Paper.

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
AI assurance	**DA-03:** The applicant should describe the system and subsystem architecture, to serve as reference for related safety (support) assessment and learning assurance objectives.	○	○	○	○	
	DA-04: Each of the captured requirements should be validated.	●	●	○	○	○
	DA-05: The applicant should document evidence that all derived requirements have been provided to the (sub)system processes, including the safety (support) assessment.	○	○	○	○	○
	DA-06: The applicant should document evidence of the validation of the derived requirements, and of the determination of any impact on the safety (support) assessment and (sub)system requirements.	○	○	○	○	○
	DA-07: Each of the captured (sub)system requirements allocated to the AI/ML constituent should be verified.	●	●	○	○	○
	DM-01: The applicant should define the set of parameters pertaining to the AI/ML constituent ODD.	○	○	○	○	○

Ilustración 30: Objetivos del bloque "AI assurance". EASA Concept Paper.

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
Human factors for AI	***EXP-09:*** *Where a customisation capability is available, the end user should be able to customise the level of details provided by the system as part of the explainability.*	O	O	O	O	O
	EXP-10: *The applicant should define the timing when the explainability will be available to the end user taking into account the time criticality of the situation, the needs of the end user, and the operational impact.*	O	O	O	O	O
	EXP-11: *The applicant should design the AI-based system so as to enable the end user to get upon request explanation or additional details on the explanation when needed.*	O	O	O	O	O
	EXP-12: *For each output relevant to the task(s), the applicant should ensure the validity of the specified explanation, based on actual measurements (e.g. monitoring) or on a quantification of the level of uncertainty.*	O	O	O	O	O
	EXP-13: *The AI-based system should be able to deliver an indication of the degree of reliability of its output as part of the explanation based on actual measurements (e.g. monitoring) or on a quantification of the level of uncertainty.*	O	O	O	O	O
	EXP-14: The AI-based system inputs should be monitored to be within the operational boundaries (both in terms of input parameter range and distribution) in which the AI/ML constituent performance is guaranteed, and deviations should be indicated to the relevant users and end users.	O	O	O	O	O
	EXP-15: The AI-based system outputs should be monitored to be within the specified operational performance boundaries, and deviations should be indicated to the relevant users and end users.	O	O	O	O	O

Ilustración 31: Objetivos del bloque "Human factors for AI". EASA Concept Paper.

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
Human Factors for AI	**HF-04:** If a decision is taken by the AI-based system, the applicant should design the AI-based system with the ability to request from the end-user a cross-check validation. Corollary objective: The applicant should design the AI-based system with the ability to cross-check and validate a decision made by the end user automatically or on request	O	O	O	O	O
	HF-05: For complex situations under normal operations, the applicant should design the AI-based system with the ability to identify suboptimal strategy and propose through argumentation an optimised solution. Corollary objective: The applicant should design the AI-based system with the ability to accept rejection required by the end user on the proposal.	O	O	O	O	O
	HF-06: For complex situations under abnormal operations, the applicant should design the AI-based system with the ability to identify the problem, share the diagnosis including the root cause, the resolution strategy and the anticipated operational consequences. Corollary objective: The applicant should design the AI-based system with the ability to consider the arguments shared by the end user.	O	O	O	O	O
	HF-07: The applicant should design the AI-based system with the ability to detect poor decision-making by the end user in a time-critical situation.	O	O	O	O	O
	HF-08: The applicant should design the AI-based system with the ability to take the appropriate action outside of a collaboration scheme, in case of detection of poor decision-making by the end user in a time-critical situation.	O	O	O	O	O
	HF-09: The applicant should design the AI-based system with the ability to negotiate, argue, and support its positions.	O	O	O	O	O
	HF-10: The applicant should design the AI-based system with the ability to accept the modification of task allocation / task adjustments (instantaneous/short-term).	O	O	O	O	O

Ilustración 32: Objetivos del bloque "Human factors for AI". EASA Concept Paper.

7 CONCLUSIONES Y ANÁLISIS DE NECESIDADES PARA LA DEFINICIÓN DE UN CRITERIO *SAFETY CRITICAL*.

Para resumir lo explicado en esta sección y a modo de conclusiones del trabajo, y en particular del análisis de necesidades, el siguiente hilo argumental recoge las necesidades identificadas e identifica algunos “gaps”.

7.1 Definición de seguridad crítica.

Para definir la seguridad crítica, en primer lugar, es necesario tener en cuenta las definiciones ya redactadas por las normas.

En el ámbito militar, la norma MIL-STD-882E proporciona una definición exacta que permite diferenciar entre un componente crítico para la seguridad y uno que no lo es, como se ha visto anteriormente.

En el mundo civil, EASA explica que el término "componente crítico" se utiliza en varios requisitos de EASA, especificaciones de certificación y en el acuerdo bilateral UE-EE.UU., pero no siempre se define. No existe una definición general porque depende del contexto en el que se utilice el término.

Por otra parte, RTCA no proporciona una definición exacta de seguridad crítica. Se da a entender que, con mayor o menor criticidad, todo lo que se aborda es crítico para la seguridad, aparte de la categoría "No safety effect".

Necesidad: Debe considerarse si las definiciones existentes de componentes o funciones críticas para la seguridad son aplicables al contexto del CAC. En caso negativo, habría que estudiar cómo adaptarlas al CAC.

7.2 Esquemas de criticidad y nivel de rigor.

Tradicionalmente, la criticidad se ha regido por la gravedad de las consecuencias potenciales. En el mundo civil, tenemos RTCA, que establece “Design Assurance Levels” basados en la severidad. Para cada DAL, se establecen objetivos de cumplimiento en el proceso de certificación. Por otro lado, en el mundo militar, la norma MIL-STD-882E también tiene en cuenta la severidad.

Podríamos considerar los objetivos a cumplir en la certificación como una pirámide invertida, ya que el número de objetivos de un DAL es superior a los objetivos de los DAL inferiores. Es decir, el DAL D tendrá unos objetivos y el nivel superior, el DAL C, cumplirá los objetivos del DAL D más otros para ese nivel. Así, la DAL B cumplirá todos los de la DAL C más los suyos propios. Lo mismo ocurre con el DAL A.

Así, hay objetivos en niveles de gravedad inferiores que deben ser comprendidos para cumplir los niveles de gravedad superiores. Por tanto, hay que determinar cuál es el salto sustancial de los objetivos DAL C y D a los objetivos DAL A y B.

EASA especifica los objetivos de la Inteligencia Artificial que están relacionados con los "Design Assurance Level" C y D (relacionados con la seguridad) y E (no relacionados con la seguridad). Sin embargo, no se abordan los DAL A y B (críticos para la seguridad).

Gap: Los documentos RTCA y MIL-STD-882E tratan de la certificación tradicional de software y EASA está orientada a la certificación de Inteligencia Artificial. Aún no se han desarrollado los objetivos correspondientes a las DAL A y B de EASA.

Necesidad: Considerar los esquemas de severidad en el contexto del CAC, y determinar si alguno de los posibles esquemas aplicables y ya disponibles será adoptado para la certificación de componentes críticos de seguridad basados en IA, o debe servir de base para la definición de un esquema específico de clasificación de severidad.

7.3 Autonomía.

Además de la certificación tradicional basada en la severidad, MIL-STD-882E y EASA abordan otras dimensiones, como la autonomía.

MIL-STD-882E desarrolla una categorización basada en el nivel de control que tiene el software, siendo el Nivel 1 el de máxima autonomía en funciones de seguridad y el nivel 5 el nivel en el que las funciones realizadas por los componentes de IA no tienen impacto en la seguridad. También combina el nivel de control con la gravedad de las consecuencias.

MIL-STD-882F incluye la Inteligencia Artificial. Combina las categorías de nivel de autonomía de IA con la gravedad de las consecuencias. Estas nuevas categorías de nivel de autonomía de IA no pretenden duplicar la anterior categorización de nivel de control de software, sino centrarse en aspectos de IA y aprendizaje automático que van más allá del "software tradicional".

Gap: MIL-STD-882F no ha desarrollado las categorías de nivel de autonomía de IA.

EASA, en cambio, sí realiza una aproximación a las categorías de nivel de autonomía de IA. Asimismo, se incluyen los Niveles 1 y 2, pero aún no se contempla el Nivel 3. EASA trata el nivel de autonomía independientemente del "Assurance Level" en el cumplimiento de objetivos para el proceso de certificación.

Gap: EASA aún no aborda el Nivel 3 de autonomía.

Hemos visto que se están desarrollando clasificaciones con conceptos similares, pero al mismo tiempo con diferencias que hacen necesario encontrar una forma de combinarlas.

Necesidad: Integración de los enfoques iniciales civil (EASA) y militar (MIL-STD-882E y MIL-STD-882F) de forma coherente y adaptada al contexto del CAC. Puede que la clasificación EASA se ajuste a las categorías de nivel de autonomía de la IA MIL-STD-882F.

7.4 Mayores niveles de garantía y autonomía.

Para alcanzar los nuevos niveles, es necesario ampliar los datos para aumentar la gama de conocimientos. Para el DAL A, que contempla una probabilidad de fallo de 10E-9, se estima que es necesario un dominio de datos[3] del orden de 10E12 o 10E13 [46]. De esta forma, se consigue que las predicciones e inferencias que haga el ML tengan una robustez[4] y estabilidad más adecuadas. Las predicciones no son fiables, mientras que las inferencias verifican que el comportamiento del modelo real actúa igual que el modelo de simulación o entrenamiento. Además, otro aspecto no trivial de las inferencias a considerar es que con ligeros cambios en la entrada, se mantienen las mismas propiedades que el modelo anterior produciendo que su comportamiento no cambie drásticamente.

Por otro lado, es lógico que a mayor DAL, factores como la robustez o los aspectos éticos sean más relevantes. La robustez podría interpretarse en este contexto como la propiedad del sistema, no sólo del algoritmo de ML, de ser tolerante a fallos.

Por tanto, es necesario reflexionar sobre cómo completar e integrar estos nuevos datos y otros aspectos a considerar en los CS y AMC cuando se estudien niveles superiores.

Necesidad: Identificar las dimensiones adicionales, distintas del nivel de garantía de seguridad y el nivel de IA, que son necesarias para alcanzar los niveles superiores de autonomía y niveles de garantía. Por ejemplo: dominio de los datos[5], robustez, ética y supervisión humana[6]. Cómo tratar estos conceptos en términos de nuevos objetivos por desarrollar.

[3] Sin datos no hay IA. El funcionamiento de muchos sistemas de IA, y las acciones y las decisiones a las que pueden conducir, dependen en gran medida del conjunto de datos con los que se han entrenado los sistemas. Por lo tanto, deben tomarse las medidas necesarias para garantizar que, en lo que respecta a los datos utilizados para entrenar los sistemas de IA, se respeten los valores y las normas UE [55].

[4] Los sistemas de IA deben ser técnicamente sólidos y precisos para ser fiables. El desarrollo y funcionamiento de estos sistemas deben ser tales que garanticen que los sistemas de IA se comportan de forma fiable según lo previsto. Deben tomarse todas las medidas razonables para minimizar el riesgo de daños [55].

[5] Teniendo en cuenta elementos como la complejidad y la opacidad de muchos sistemas IA, se exigen requisitos relativos a la conservación de registros en relación con la programación del algoritmo, los datos utilizados para entrenar los sistemas y, en algunos casos, la conservación de los propios datos [55].

[6] La supervisión humana ayuda a garantizar que un sistema de IA no socave la autonomía humana ni cause otros efectos adversos. El objetivo de una IA fiable, ética y centrada en el ser humano sólo puede alcanzarse garantizando una participación adecuada de los seres humanos en relación con las aplicaciones de IA de alto riesgo [55].

8 CONCLUSIONES Y ACCIONES FUTURAS

Durante el desarrollo del presente documento se ha mostrado una posible metodología para comenzar a trabajar en la estandarización de procesos de certificación de componentes críticos para la seguridad basados en Inteligencia Artificial.

Para llevarlo a cabo, se ha hecho una propuesta de plantilla para el análisis de documentos y su forma de trabajarla. También se ha explicado el posible desarrollo de alguna sección de gran importancia como el BFS. Para ello, se ha hecho una investigación inicial con la finalidad de ubicarse en el contexto de trabajo y plantar las bases de la tarea.

Además, un primer paso imprescindible para la tarea era realizar un análisis de necesidades de la implementación de las tecnologías IA en el contexto "Safety Critical". En este trabajo se identifican algunas de éstas, por considerarse de gran importancia.

Sin embargo, todo esto es solo un pequeño paso en el largo camino que esta labor supone. Así pues, se quiere introducir unas pequeñas pinceladas de lo que podrían ser los siguientes pasos a realizar y las acciones que se deben abordar en el futuro.

En primer lugar, es necesario establecer los límites del marco de trabajo. En consecuencia, se deben identificar los casos de uso del CAC para posteriormente poder completar y definir el BFS. En este trabajo, se ha introducido el BFS, basándose en el ciclo OODA e indicando cuál sería su funcionamiento. Esta primera versión de la lista se genera solo de forma temporal hasta determinar casos de uso. En este momento de podrá producir una BFS claramente definida y completa que abarque todas las funciones que puedan darse en el contexto del CAC. El hecho de no disponer de los casos de uso es una gran limitación, y por esta razón debería ser uno de los primeros pasos a realizar.

En segundo lugar, se debe completar la lista de estándares de referencia a analizar. Como se explica anteriormente, la primera versión V0 de análisis se debe ir iterando de forma que abarque todos los estándares de referencia que se deban incluir. De esta forma, el documento final recogerá todos los aspectos que ya han sido tratados por los estándares y son aplicables en esta tarea, y, además, aquellos aspectos que no se han tratado o que no se pueden aplicar en este contexto CAC. Esto último serán las necesidades y los "gaps". Como se ha mencionado anteriormente, en este documento ya se ha empezado a analizar algún estándar y se han podido extraer las primeras necesidades. No obstante, habrá muchas otras que se deben obtener del análisis del resto de referencias.

A continuación, se deberán estudiar los requisitos específicos de cada tecnología IA para las funciones deseadas. Este punto está fuera del alcance del presente documento, pero debería ser un paso necesario en el futuro desarrollo del trabajo. Esto significa estudiar en profundidad el proceso de programación y los criterios de cumplimiento de las tecnologías IA.

Finalmente, sería necesario como trabajo a futuro ejecutar un demostrador.

Por último, resaltar que habrá muchas acciones futuras no mencionadas y que también serán necesarias. Sin embargo, en este apartado se mencionan algunas de ellas, haciendo hincapié en aquellos aspectos que se consideran especialmente relevantes.

9 BIBLIOGRAFÍA

[1] S. J. R. y. P. Norving, «Inteligencia Artificial. Un enfoque moderno,» PEARSON EDUCACIÓN, Madrid, 2004.

[2] A. C. Vicci, «¿Qué Pasó con el Test de Turing?,» *TEACS,* nº 23, pp. 60-69, 2018.

[3] A. A. &. C. Gutiérrez, «Historia y evolución de la Inteligencia Artificial,» *Bits de Ciencia,* nº 21, pp. 14-21, 2021.

[4] R. Holgado, «Así es Eliza, el primer chatbot de la historia,» *20 minutos,* 6 Abril 2021.

[5] P. &. C. C. McCorduck, Machines who think: A personal inquiry into the history and prospects of artificial intelligence, CRC Press, 2004.

[6] Z.-H. Zhou, Machine Learning, Springer Nature, 2021.

[7] B. L. S. Y. Seonwoo Min, «Deep Learning in Bioinformatics,» *Briefings in Bioinformatics,* vol. 18, nº 5, pp. 851-869, 2017.

[8] Y. B. &. G. H. Yann LeCun, «Deep Learning,» *Nature,* pp. 436-444, 2015.

[9] G. Hinton, «Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,» *IEEE Signal Processing Magazine,* vol. 29, nº 6, pp. 82-97, 2012.

[10] J. J. J. A. L. Y. &. B. C. Tompson, Joint training of a convolutional network and a graphical model for human pose estimation. Advances in neural information processing systems, 2014.

[11] N. &. Y. E. Bostrom, «The ethics of artificial intelligence,» de *The Cambridge handbook of artificial intelligence*, 2014, pp. 316-334.

[12] V. C. Müller, «Ethics of Artificial Intelligence and Robotics,» de *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, 2021.

[13] C. L. B. R. G. C. D. P. M. R. J. O. R. I. B. d. E. E. &. L. L. F. Tafur, «Inteligencia Artificial en las Operaciones Áereas,» de *Memorias 3er Seminario Internacional en Mantenimiento e Investigación Aeronáutica*, Bogotá, 2022, p. 35.

[14] «Real Academia Española,» [En línea]. Available: https://dle.rae.es/inteligencia?m=form2#2DxmhCT.

[15] L. Rouhiainen, Inteligencia Artificial, Madrid: Alienta Editorial, 2018.

[16] V. Tiple, Recommendations on the European Commission's WHITE PAPER on Artificial Intelligence-A European approach to exellence and trust, 2020.

[17] Y. B. Y. &. H. G. LeCun, «Deep Learning,» *Nature,* nº 521, pp. 436-444, 2015.

[18] S. &. H. A. Gössling, «The global scale, distribution and growth of aviation: Implications for climate change,» *Global Environmental Change,* vol. 65, 2020.

[19] D. M. C. S. J. M. R. &. B. K. Medeiros, «RNAV and RNP AR approach systems: the case for Pico Island airport.,» *International Journal of Aviation Management,* pp. 181-200, 2012.

[20] K. Simoes Spencer, Fuel Consumption Optimization using Neural Networks and Genetic Algorithms, Lisboa, 2011.

[21] A. M. G. &. R. C. Gaudín, Planificación de vuelo utilizando algoritmos evolutivos, 2020.

[22] M. H. R. K. D. S. D. O. A. Aloqali, «On the Role of Futuristic Technologies in Securing UAV-Supported Autonomous Vehicles,» *IEEE Consumer Electronics Magazine,* vol. 11, nº 6, pp. 93-105, 2022.

[23] J. G. F. G. M. B. G. J. J. Arboleda, Aplicación de la inteligencia artificial en el transporte internacional de mercancías, Institución Universitaria Esumer, 2021.

[24] M. K. A. A. M. M. T. Borhani, A Multicriteria Optimization for Flight Route Networks in Large-Scale Airlines Using Intelligent Spatial Information, 2020.

[25] EASA Concept Paper: First usable guidance for Level 1 & 2 machine learning applications, 2023.

[26] A. H. J. B. E. &. W. S. Kiser, The combat cloud: enabling multi-domain command and control across the range of military operations, 0039: Air Command and Staff College, 2017.

[27] L. Rojo Pinilla, Cobertura 5G para la integración de radios tácticas SDR, Vigo: Calderón, repositorio institucional del Centro Universitario de la Defensa, ENM, 2023.

[28] M. Schanz, «The Combat Cloud,» *Air Force Magazine,* pp. 38-41, 2014.

[29] C. Galán, «La certificación como mecanismo de control de la inteligencia artificial en Europa,» de *Boletín IEEE*, 2019, pp. 622-637.

[30] P. M. E. S. W. J. S. C. D. T. V. T. .. &. N. B. Winter, Trusted artificial intelligence: Towards certification of machine learning applications, 2021.

[31] T. B. S. C. B. &. A. P. Tommasi, Towards fairness certification in artificial intelligence, 2021.

[32] A. A. H. &. A. N. Agarwal, «Fairness Score and process standardization: framework for fairness certification in artificial intelligence systems,» de *AI and Ethics*, 2023, pp. 267-279.

[33] X. v. N. T. S. J. M. C. M. &. C. N. Ferrer, «Bias and discrimination in AI: a cross-disciplinary perspective,» *IEEE Technology and Society Magazine,* vol. 40, nº 2, pp. 72-80, 2021.

[34] S. A. Y. P. B. B. C. G. C. &. K. J. Sedal, «Fairness in the eyes of the data: Certifying machine-learning models,» de *In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

[35] Manual de aeronavegabilidad de la OACI, tercera edición, 2014.

[36] «Declaraciones sobre productos, componentes, equipos y materiales,» de *USA CFR Título 14, Capítulo I, Subcapítulo A, Parte 3, Subparte A*.

[37] F. De Florio, Airworthiness: An Introduction to Aircraft Certification and Operations, 2016.

[38] «Reglamento de aeronavegabilidad de la Defensa,» 2015. [En línea]. Available: https://www.boe.es/diario_boe/txt.php?id=BOE-A-2015-11426.

[39] A. E. d. Defensa, EMAD 1, Edición 1.4, 2021.

[40] MIL-STD-882E. Department of Defence Standard Practice., 2012.

[41] W. Kellum, «An early attempt to evaluate psycological fitness for flight training,» *Contact,* vol. 6, nº 4, pp. 232-235, 1948.

[42] «Thales,» 21 Octubre 2020. [En línea]. Available: https://www.thalesgroup.com/es/group/journalist/magazine/estamos-el-centro-toma-decisiones-observacion-decision-actuacion.

[43] D. S. B. J. &. W. J. Fadok, Air power's quest for strategic paralysis, Maxwell AFB, Alabama, 1995.

[44] L. Rierson, Developing Safety-Critical Software: A Practical Guide for Aviation Software and DO-178C Compliance, Boca Raton: Taylor & Francis Group, 2013.

[45] MIL-STD-882F. Department of Defence Standard Practice. System Safety.

[46] M. B. a. D. F. &. C. D. L. v. D. ConsuNova's Head of Advanced Avionics Systems, «Applying Design Assurance in Aerospace to AI,» 19 10 2022. [En línea]. Available: https://www.youtube.com/watch?v=LYHQFbJDwAM .

[47] E. C. Fernández, «Planificación de trayectorias 4D de aeronaves en entornos multiobjetivos,» 2017.

[48] Department of Transport Studies, «European airline delay cost reference values,» University of Wetsminster, 2013.

[49] «Global Air Traffic Management Operational Concept. Doc 9854-AN/458,» ICAO, Montreal, Canadá, 2005.

[50] European Organisation for the Safety of Air Navigation, «Manual for Airspace Planning,» EUROCONTROL, Brussels, Belgium, 2003.

[51] F. &. P. P. Y. Saéz Nieto, «Descubrir la Navegación Aérea,» Centro de Documentación y Publicaciones de AENA, 2003.

[52] ICAO, «Procedures for Air Navigation Services - Air Traffic Management. Doc 9854-ATM/501,» Montreal, Canadá, 2007.

[53] F. J. Saéz Nieto, «Navegación Aérea - Posicionamiento, Guiado y Gestión del Tráfico Aéreo,» Ibergaceta Publicaciones, 2013.

[54] F. P. S. L. &. G. C. F. Saéz Nieto, «La Navegación Aérea y el Aeropuerto,» Fundación AENA, 2003.

[55] G. R. O. S. S. Kilian, WHITE PAPER On Artificial Intelligence-A European approach to excellence and trust, 2020.

ANEXO I

ANÁLISIS DE ESTÁNDARES PARA LA VERSIÓN V0

Los estándares recopilados en la investigación inicial y otros añadidos posteriormente se recopilaron y organizaron en la siguiente lista para poder organizar su análisis de la mejor manera posible.

Se muestra la totalidad de la lista recopilada. Véase las siguientes Ilustraciones.

DOC. NUMBER IDENTIFICATION	STANDARD, REGULATIONS, OTHER REFERENCES	SOURCE	DOC. TYPE
1	AIR6987 (WIP) Artificial Intelligence in Aeronautical Systems: Taxonomy"	SAE.ORG	WIP
2	ARP4754 - Guidelines for Development of Civil Aircraft and Systems	SAE.ORG	PAYMENT
3	ARP4761 - GUIDELINES AND METHODS FOR CONDUCTING THE SAFETY ASSESSMENT PROCESS ON CIVIL AIRBORNE SYSTEMS AND EQUIPMENT	SAE.ORG	PAYMENT
4	AS6983 (WIP) - Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI	SAE.ORG	WIP
5	CEN-CENELEC JTC 21 Artificial Intelligence	CENCENELEC.EU	WORKING GROUP
6	CRISP-DM METHODOLOGY	TBD	SUPPORT METHOD
7	DO-178B, Software Considerations in Airborne Systems and Equipment Certification	RTCA	PAYMENT
	DO-178C, Software Considerations in Airborne Systems and Equipment Certification	RTCA	PAYMENT
8	DOT/FAA/TC-16/4 - Verification of Adaptive Systems	FAA	FREE
9	EASA Concept Paper 'First usable guidance for Level 1 machine learning applications'	EASA	FREE
10	EASA Artificial Intelligence Roadmap 1.0	EASA	FREE
11	EASA AMC-20 - General Acceptable Means of Compliance for Airworthiness of Products, Parts and Appliances	EASA	FREE
12	EASA AMCS UAS VTOL LEVEL OF AUTO		
13	EASA Concepts of Design Assurance for Neural Networks (CoDANN)	EASA	FREE
14	EUROPEAN MILITARY AIRWORTHINESS CERTIFICATION CRITERIA (EMACC)	EDA	FREE
15	ENISA SECURING AI, AND AI FOR SECURITY	ENISA	UNKNOWN
16	ER-022 - EUROCAE. 2021. Artificial Intelligence in aeronautical systems: Statement of concern. s.l. : EUROCAE, 2021. ER-022.	EUROCAE	PAYMENT
17	EU COMISSION . Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM/2021/206 final. 2021.	EU COMISSION	FREE
18	EU High Level Expert Group on AI. 2020. Assessment List for Trustworthy AI (ALTAI). s.l. : European Commission, 2020. Ethics Guidelines for Trustworthy AI. s.l. : European Commission, 2019	EU COMISSION	WORKING GROUP

Ilustración 33: Listado de estándares/referencias.

DOC. NUMBER IDENTIFICATION	STANDARD, REGULATIONS, OTHER REFERENCES	SOURCE	DOC. TYPE
19	EUROCAE WG-114 Artificial Intelligence	EUROCAE	WORKING GROUP
20	ISO/IEC JTC 1/SC 42 Artificial Intelligence	ISO	PAYMENT
21	ISO/IEC 22989, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology	ISO	PAYMENT
22	ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology	ISO	PAYMENT
23	ISO/IEC 23053:2022 - Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)	ISO	PAYMENT
24	ISO/IEC AWI 5259-2, Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures	ISO	PAYMENT
25	ISO/IEC AWI 5259-3, Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines	ISO	PAYMENT
26	ISO/IEC AWI TR 5469, Artificial intelligence — Functional safety and AI systems	ISO	PAYMENT
27	ISO/IEC AWI TS 29119-11, Information technology — Artificial intelligence — Testing for AI systems — Part 11	ISO	PAYMENT
28	ISO/IEC AWI TS 5471, Artificial intelligence — Quality evaluation guidelines for AI systems	ISO	PAYMENT
29	ISO/IEC AWI TS 6254, Information technology — Artificial intelligence — Objectives and approaches for explainability of ML models and AI systems	ISO	PAYMENT
30	ISO/IEC AWI TS 8200, Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems	ISO	PAYMENT
31	ISO/IEC CD 42001.2, Information Technology — Artificial intelligence — Management system	ISO	PAYMENT
32	ISO/IEC CD 5339, Information Technology — Artificial Intelligence — Guidelines for AI applications	ISO	PAYMENT
33	ISO/IEC DIS 24029-2, Artificial intelligence (AI) — Assessment of the Robustness of neural networks — Part 2: Methodology for the use of formal methods	ISO	PAYMENT
34	ISO/IEC DIS 25059, Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems	ISO	PAYMENT
35	ISO/IEC DIS 42001 - Information technology — Artificial intelligence — Management system	ISO	PAYMENT
36	ISO/IEC DTS 4213.2, Information technology — Artificial Intelligence — Assessment of machine learning classification performance	ISO	PAYMENT

Ilustración 34: Listado de estándares/referencias.

DOC. NUMBER IDENTIFICATION	STANDARD, REGULATIONS, OTHER REFERENCES	SOURCE	DOC. TYPE
37	ISO/IEC FDIS 23894 - Information technology — Artificial intelligence — Guidance on risk management	ISO	PAYMENT
38	ISO/IEC JTC 1/SC 42 Artificial intelligence	ISO	PAYMENT
39	ISO/IEC TR 24027:2021 - Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making	ISO	PAYMENT
40	ISO/IEC TR 24028:2020, Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence	ISO	PAYMENT
41	ISO/IEC TR 24029-1:2021, Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview	ISO	PAYMENT
42	ISO/IEC TR 24030:2021 - Information technology — Artificial intelligence (AI) — Use cases	ISO	PAYMENT
43	ISO/IEC TR 24372:2021 - Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems	ISO	PAYMENT
44	ISO/IEC TR 5469 - Artificial intelligence — Functional safety and AI systems	ISO	PAYMENT
45	MIL-HDBK-516C - AIRWORTHINESS CERTIFICATION CRITERIA	MIL	FREE
46	MIL-STD-882 - SYSTEM SAFETY	MIL	FREE
47	MITRE ATLAS FRAMEWORK	https://attack.mitre.org/	UNKNOWN
48	NASA/CR-2015-218702 - Certification Considerations for Adaptive Systems	NASA	FREE
49	NCSC PRINCIPLES FOR THE SECURITY OF ML	TBC	
50	PRCEN/CLC/TR 17894 - Artificial Intelligence Conformity Assessment	CENCENELEC.EU	WORKING GROUP
51	S079L03T00-005 - DEEL CG - White Paper Machine Learning in Certified Systems	HAL OPEN SCIENCE	FREE
52	EUROCAE WG114 / SAE G34 Artificial Intelligence in Aviation- AIR6988 - Artificial Intelligence in Aeronautical Systems: Statement of Concerns	EUROCAE	PAYMENT
53	ISO/IEC TR 24028 Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence	ISO	PAYMENT
54	ISO/IEC 23894 Information technology - Artificial intelligence - Guidance on risk management	DIN ISO IEC	PAYMENT
55	ISO/IEC 38507 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations	ISO	PAYMENT
56	ISO/IEC TR 20547-1 Information technology — Big data reference architecture Part 1: Framework and application process	ISO	PAYMENT
57	ISO/IEC TR 20547-2 Information technology — Big data reference architecture Part 2: Use cases and derived requirements	ISO	PAYMENT

Ilustración 35: Listado de estándares/referencias.

DOC. NUMBER IDENTIFICATION	STANDARD, REGULATIONS, OTHER REFERENCES	SOURCE	DOC. TYPE
58	ISO/IEC 20547-3 Information technology — Big data reference architecture Part 3: Reference architecture	ISO	PAYMENT
59	ISO/IEC TR 20547-4 Information technology — Big data reference architecture Part 4: Security and privacy	ISO	PAYMENT
60	ISO/IEC TR 20547-5 Information technology — Big data reference architecture Part 5: Standards roadmap	ISO	PAYMENT
61	ISO 9001:2015 Quality Management Systems - Requirements	ISO	PAYMENT
62	ISO/IEC 27000 Series - Information security management	ISO	PAYMENT
63	NIST AI Risk Management Framework NIST AI 100-1	NIST	FREE
64	Defence Artificial Intelligence Strategy	MoD UK	FREE
65	DO-254 Design Assurance Guidance for Airborne Electronic Hardware	RTCA	PAYMENT
66	INCOSE AI Systems WG	INCOSE	WORKING GROUP
67	ISO 8000-61:2016(en), Data quality — Part 61: Data quality management: Process reference model	ISO	PAYMENT
68	NATO RTG SAS-181 Exploiting Reinforcement Learning to Achieve Decision Advantage	NATO	WORKING GROUP
69	ECSS-Q-ST-40C ECSS Space product assurance: Safety	ECSS	FREE
70	ECSS-E-ST-40C ECSS Space engineering: Software	ECSS	FREE
71	ECSS-Q-ST-80C ECSS Space product assurance: Software	ECSS	FREE
72	Artificial Intelligence: A European perspective. AI flagship report	EU COMISSION	FREE
73	The Fly AI Report Demstifying and Acelerating AI in aviation/ATM	EUROCONTROL/ AI HLEG	FREE
74	RTCA DO-326A Airworthiness Security Process Specification	RTCA	PAYMENT
75	RTCA DO-355 Information Security Guidance for Continuing Airworthiness	RTCA	PAYMENT
76	RTCA DO-356A Airworthiness Security Methods and Considerations	RTCA	PAYMENT
77	STANAG 5524 - NATO Interoperability	STANAG	PAYMENT
78	STANAG 4671 - USAR	STANAG	PAYMENT
79	STANAG 4703 - LightUSAR	STANAG	PAYMENT
80	CS-ACNS - ATM	EASA	FREE
81	EASA AI Concept Paper (proposed Issue2) open for consultation	EASA	FREE

Ilustración 36: Listado de estándares/referencias.

DOC. NUMBER IDENTIFICATION	STANDARD, REGULATIONS, OTHER REFERENCES	SOURCE	DOC. TYPE
79	STANAG 4703 - LightUSAR	STANAG	PAYMENT
80	CS-ACNS - ATM	EASA	FREE
81	EASA AI Concept Paper (proposed Issue2) open for consultation	EASA	FREE
82	DOT/FAA/TC-21/48 - Neural Network Based Runway Landing Guidance for General Aviation Autoland	FAA	FREE
83	CEN-CENELEC Road Map Report on AI, version 2020-09	CENCENELEC.EU	FREE
84	Machine Learning in Certified Systems - White Paper	DEEL	FREE
85	Concepts of Design Assurance for Neural Networks (CoDANN) II	EASA	FREE
86	RTCA DO-178C and suplements - Software Considerations in Airborne Systems and Equipment Certification	RTCA	PAYMENT
87	AI WATCH. Defining Artificial Intelligence	EU COMISSION	FREE
88			
89			
90			
91			
92			

Ilustración 37: Listado de estándares/referencias.