Aplicación de la IA al ferrocarril
Notas sobre seguridad en la operación

29 de octubre 2024

Jornada organizada por

ASOCIACIÓN NACIONAL
DE **INGENIEROS DEL ICAI**

AGENCIA ESTATAL DE
SEGURIDAD FERROVIARIA

adscrita al

MINISTERIO
DE TRANSPORTES
Y MOVILIDAD SOSTENIBLE

Introducción. El contexto (1)

Seguridad en la operación

¿Cuál es el valor aceptable?

- **De forma genérica se guía por la exigencia social de seguridad.**
- **La exigencia social de seguridad no es la misma para todos los medios de transporte.**

Principio básico

- **No regresión (o GAME): La introducción de un nuevo componente, subsistema, etc, o su sustitución no degrada el nivel global de seguridad del sistema.**

Introducción. El contexto (2)

Cambios en el sistema ferroviario

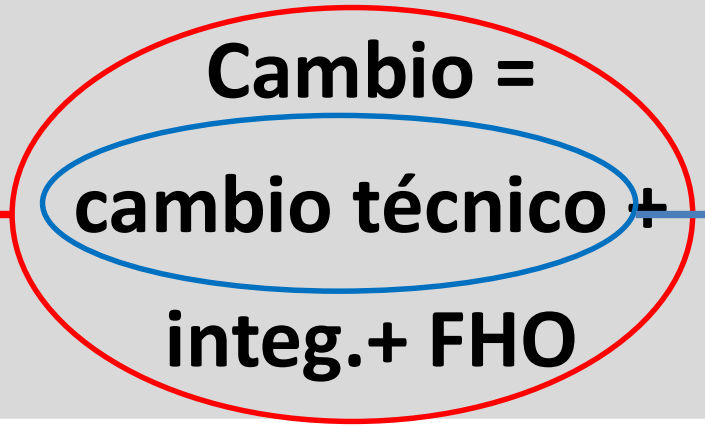
Cambio técnico

↓

Integración + cambio en factores humanos y organizativos (FHO)

(casi siempre)

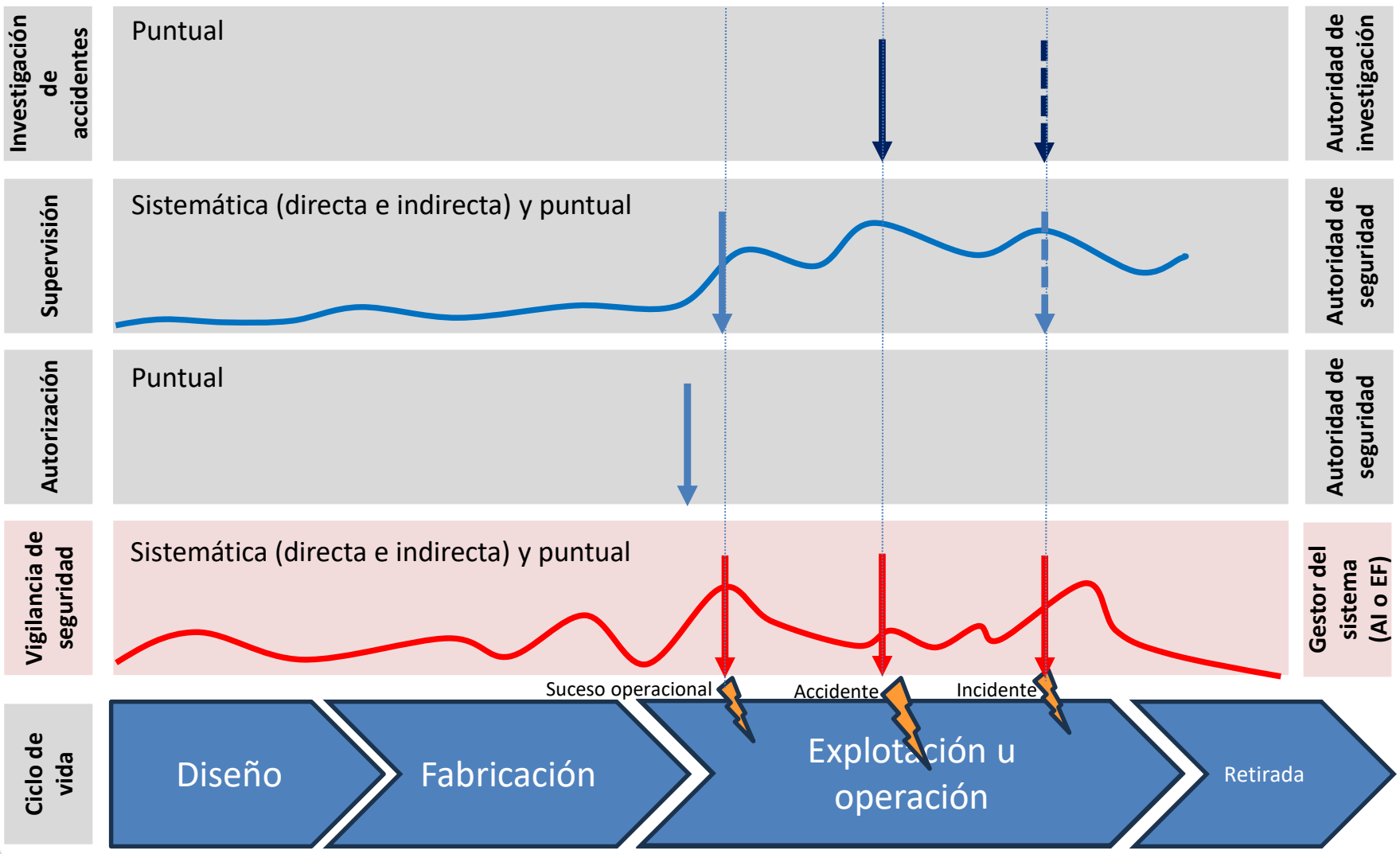
El gestor del sistema en que se integra el cambio técnico acredita globalmente su nivel de seguridad



El productor o el implantador acredita su nivel de seguridad

Introducción. Ciclo de vida de un subsistema

La seguridad en la operación durante el ciclo de vida de un SS



Modelos de inferencia ←→ seguridad en la operación

- **Una posible clasificación de los modelos de inferencia basados en aprendizaje automático en relación con la seguridad en la operación**
  - a) Modelo de inferencia influye directamente en una acción de seguridad
    - a.1) El subsistema que incluye un modelo de inferencia proporciona un dato de seguridad imprescindible para la decisión de un autómata
    - a.2) El subsistema que incluye un modelo de inferencia analiza los datos y toma solo una decisión
  - b) Modelo de inferencia apoya a un operador humano en la toma de decisiones que afectan a la seguridad
  - c) Otros

Modelos de inferencia ←→ seguridad en la operación

- **Generalidades para todos los modelos de inferencia basados en aprendizaje automático**

Proceso de autorización del SS que lo incluye:

- **Proceso de evaluación y valoración del riesgo (RUE 402/13) que debe incluir:**
  - **Adecuación de los datos de salida:**
    - El dato de salida se corresponde con lo esperado.
    - Tiempo de obtención compatible con las demás funciones del SS en que se integra.
  - **Dominio preciso de utilización** (velocidad, territorio, climatología, luminosidad, etc).
  - **Descripción de la arquitectura del modelo de inferencia** utilizado y razones que motivan su elección.

Modelos de inferencia ←→ seguridad en la operación

- **Particularidades**

  a) Modelo de inferencia influye en una acción de seguridad (sistemas o equipos de seguridad)

  Proceso de autorización del SS que lo incluye:

  - Adecuación del modelo utilizado para **garantizar** un valor máximo de la probabilidad de aparición de un dato de salida erróneo susceptible de crear una situación de peligro
  - Si es un modelo de aprendizaje supervisado, conocer los conjuntos de datos para entrenamiento y validación, así como su representatividad y pertinencia.
  - Causas de "caducidad" de los conjuntos de datos utilizados para el entrenamiento.
  - Valorar la conveniencia de "congelar" el aprendizaje.
  - FHO: Evaluación de la aptitud para realizar algunas funciones únicamente en condiciones degradadas (Evitar el olvido cuando una actividad se realiza ocasionalmente).

Modelos de inferencia ←→ seguridad en la operación

- **Particularidades**

  b) Modelo de inferencia apoya a un operador humano en la toma de decisiones que afectan a la seguridad (sistemas o equipos de seguridad).

  Proceso de autorización del SS que lo incluye:

  - Adecuación esperada (no garantizada) del modelo utilizado y su dominio de utilización (tasa de fiabilidad, intervalo de confianza, etc,)
  - Si es un modelo de aprendizaje supervisado analizar la pertinencia del conjunto de datos para entrenamiento y validación.
  - FHO: Interacción entre el modelo de inferencia y el operador humano, que debe entender los datos suministrados en condiciones nominales y degradadas.

Modelos de inferencia. Transparencia y sus atributos

Transparencia: conocer cómo y porqué un sistema toma una decisión determinada.

Poner a disposición de los agentes, tanto dentro como fuera de la organización, información sobre las características de las operaciones de un desarrollador de IA o de sus sistemas de IA [marco de ética, transparencia y rendición de cuentas para la toma de decisiones automatizadas" (Gobierno de Reino Unido)]

¿En que atributos se puede concretar la transparencia?

- Explicable: Como tomó el sistema una decisión determinada. Necesario para la supervisión "ordinaria".
- Auditable: Además deben poderse conocer las causas de la decisión adoptada por el sistema. Necesario para el análisis de eventos de seg.
- Reproducible: Datos de entrada, salida y proceso registrados en una "caja negra" independiente del elemento que incorpora la IA.

Estos atributos **deben verificarse en el proceso de evaluación y valoración del riesgo que demuestre un nivel aceptable de seguridad en la operación,** ***en la autorización del subsistema*** **que integra el modelo, para** ***hacer posible la supervisión posterior, especialmente tras un evento de seguridad (precursor de accidente, incidente, o accidente).***

Algunas lecturas

- IEEE Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems. IEEE 7009-2024. 20.05.24.
- Conditions d'autorisation des systèmes comprenant des algorithmes d'apprentissage automatique. EPSF SYS-DOCT-001-V1. 23.10.23.
- Key requirements for the effective, safe and ethical use of Artificial Intelligence (AI). Rod Muttram. IRSE News, issue 305, december 2023.

Muchas gracias

pmlekuona@seguridadferroviaria.es

AGENCIA ESTATAL DE
SEGURIDAD FERROVIARIA