GT de Tecnologías Futuras de AMETIC - Webinar

# SEGURIDAD de la IA

24 de maayo de 2024

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

i2cat

# About us

**Albert Calvo**
**Research engineer (PhD Candidate) | Adjunct lecturer at UPC**
**TAI Research line**
Trust-aware systems for cybersecurity and utilities domain

Msc. in Artificial Intelligence

albert.calvo@i2cat.net **LinkedIn**/in/albertcalvo/

**Nil Ortiz**
**Senior R&D Cybersecurity engineer**

Incident response and threat intelligence analysis

Msc. in Cybersecurity

nil.ortiz@i2cat.net **LinkedIn**/in/nilortiz/


**Acerca de i2CAT ...**

- **Centro de investigación e innovación CERCA** ubicado en Barcelona.
- Un equipo interdisciplinario de 200 miembros.
- Diferentes áreas de investigación enfocadas tecnologías digitales (5G, IA, Ciberseguridad, IoT o Comunicaciones Espaciales).
- Hemos contribuido a 39 proyectos europeos y coordinado siete proyectos diferentes.

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

# Proyectos relevantes

Algunos de nuestros proyectos relevantes en la línea de investigación de Inteligencia Artificial Confiable:

- **SIEVA (SIEM visibility assessment)** visibilidad de logs i clasificacion en MITRE ATT&CK Framework.

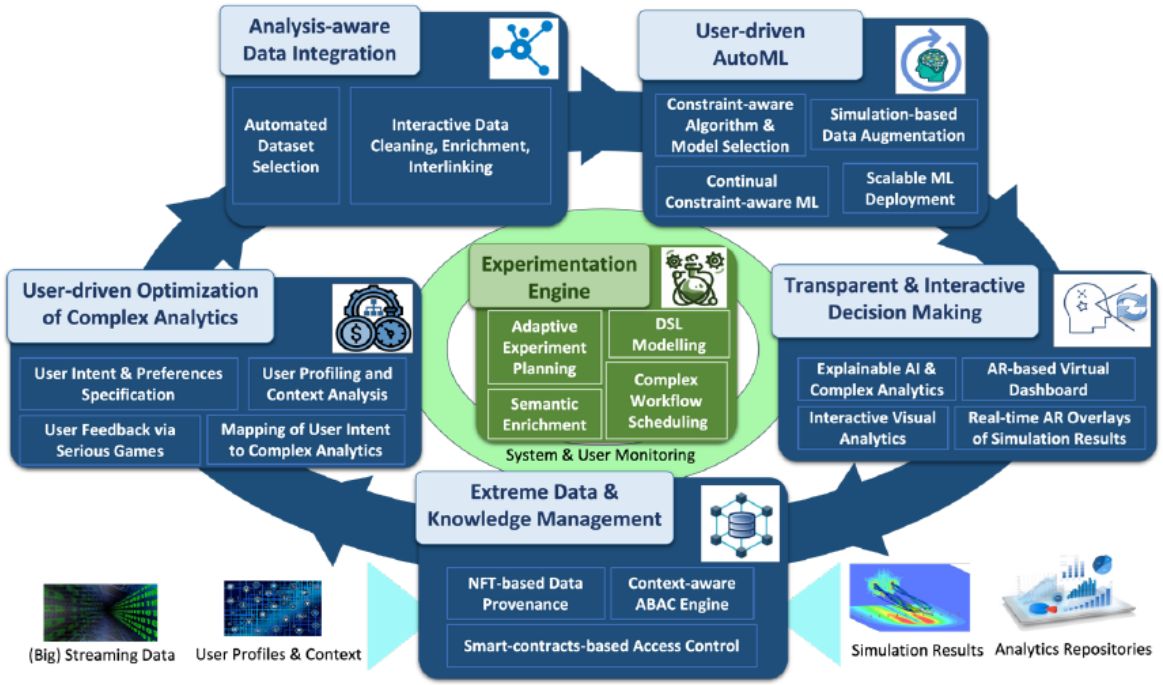
- **preventUEBA**: Calcular el grado de exposición de usuarios y entidades frente a amenazas específicas a través del poder de CTI y la IA.

- **detectUEBA:** Capacidades de detección de amenazas con el poder de los métodos de aprendizaje automático de última generación.

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

# ExtremeXP

Análisis impulsados por la experimentación para proporcionar información precisa, adecuada y confiable basada en datos, mediante la evaluación de diferentes variantes de análisis complejos, teniendo en cuenta las preferencias y comentarios del usuario final de manera automatizada.




**I - Improvement of flash flood forecasting thanks to the use of AI – CS**

**II - Increased Cybersecurity situation awareness for efficient threat mitigation i2CAT/UPC/ITU**

**III - Situational intelligence and decision making for PPDR (Public Protection and Disaster Relief) – ADS**

**IV - Flexible transportation analysis and visualization – MOBY**

**V - Failure prevention for manufacturing industry – IDK**

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

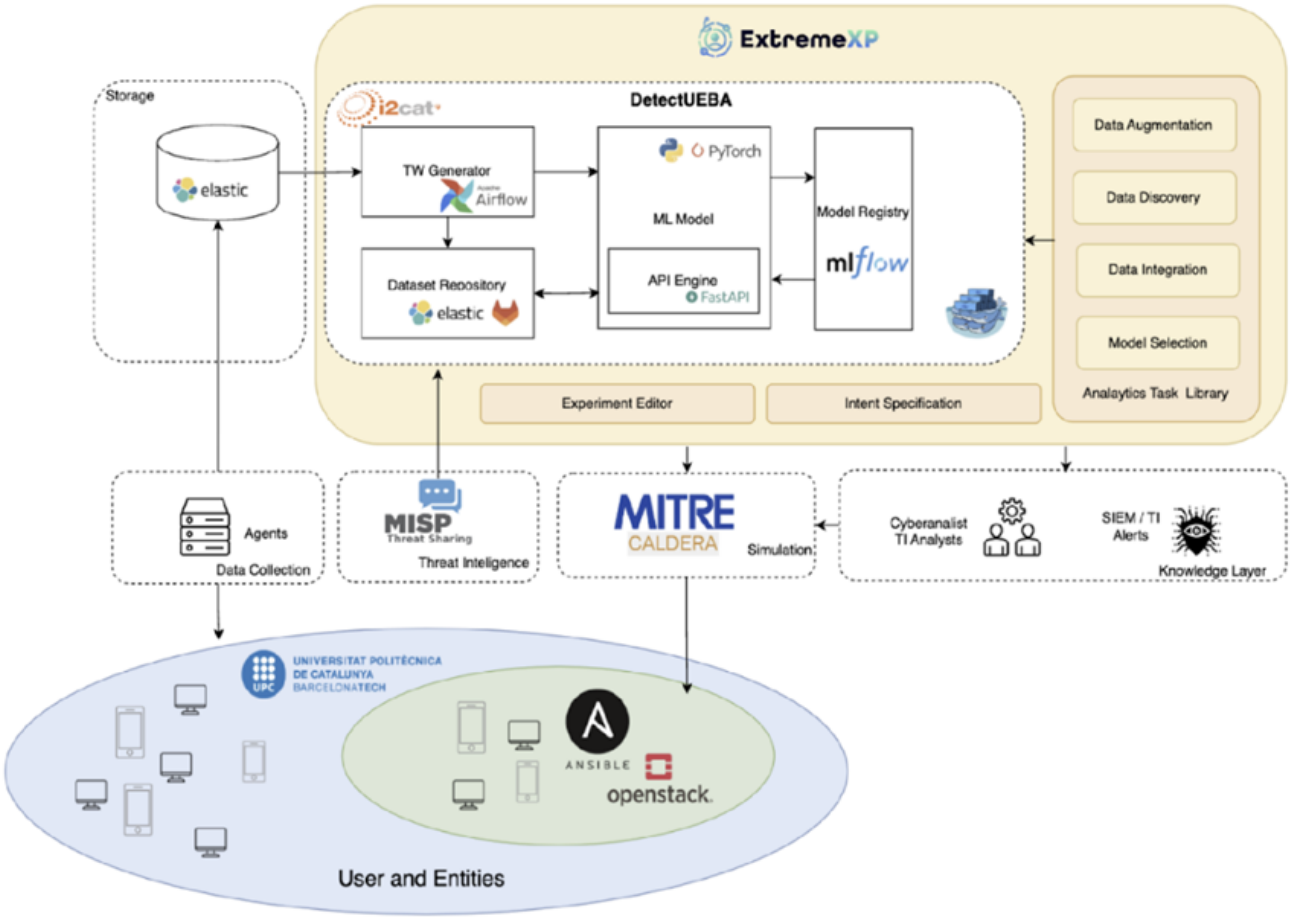
# Extreme XP

## Enhancing DetectUEBA

- SIEM technologies offer visibility but reliance on **expert knowledge** presents challenges.
- Qualifying indicators (TTPs) often involves **decision-making** problems for security analysts.
- SOAR integration is **expensive** and role of SOC has become **time intensive**.

## Use case Capabilities

- This use case aims to design a SIEM demonstrator for **multi-modal** threat detection and classification.
- It features **training with cybersecurity expert knowledge** and utilises ExtremeXP for efficient **AI pipelines** and **feedback** from SOC operators using human-in-the-loop approach.
- Qualify the behaviour of a threat through **attack simulations** and **real-time validations**.



# Fundamentos

Que tienen en común estos proyectos?

- **Critical Decision Making**: Los errores en estos casos de uso pueden tener consecuencias graves en la vida humana, la seguridad o la estabilidad financiera.
- **Limited training and bias**: Estos escenarios críticos carecen de conjuntos de datos ricos, ya que son difíciles de obtener (por ejemplo, conjuntos de datos de diagnóstico de drogas) y este entrenamiento limitado podría generar modelos sesgados.
- **Regulatory Scrutiny**: Las aplicaciones críticas están sujetas a consideraciones legales y éticas actuales.

Trustworthy AI

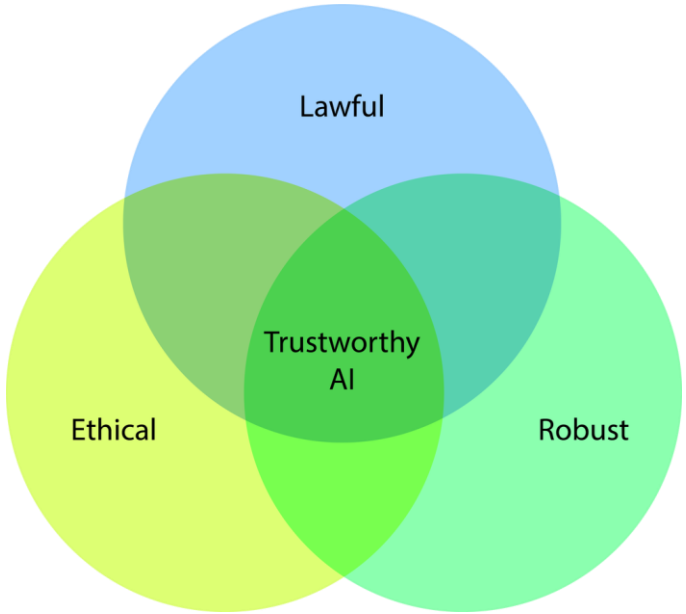
# Fundamentos

Desde la perspectiva teórica:

(1) legal - respetando todas las leyes y regulaciones aplicables

(2) ético - respetando principios y valores éticos

(3) robusto - tanto desde una perspectiva técnica como teniendo en cuenta su entorno social




Informe de Seguridad de la IA - AMETIC

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

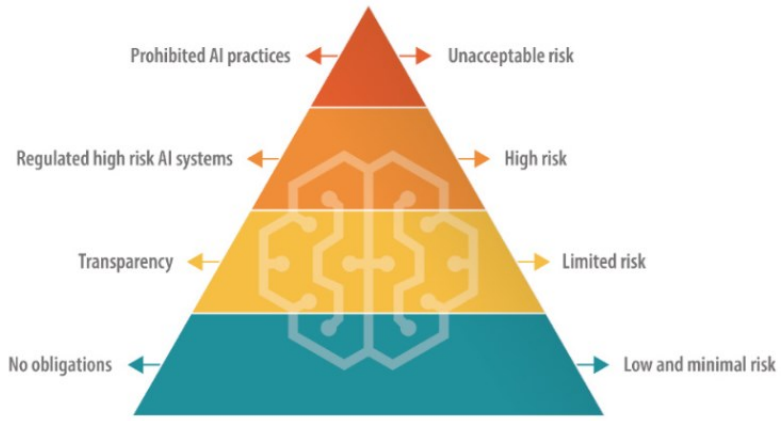
# Fundamentos

Desde la perspectiva de cumplimiento:

- perspectivas impulsadas por el mercado: Iniciativa Americana de IA (2019)
- enfoque impulsado por el ser humano: Estrategia Canadiense de Inteligencia Artificial (2017) o las iniciativas europeas, las Pautas Éticas para la Inteligencia Artificial Confiable (2019) y la Ley de Inteligencia Artificial (2023)

Ley de IA:

- Primer intento a la regulación compartida dentro de la Unión Europea.
- Prohibir y limitar los casos de uso de IA que podrían dañar a los ciudadanos.
- Enfoque basado en el riesgo.




Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

# Fundamentos

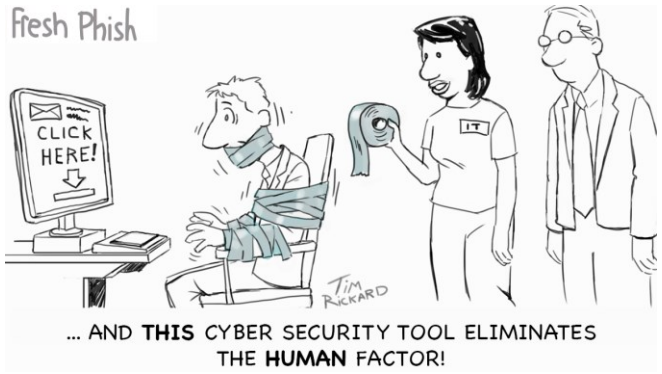
Desde la perspectiva ética:

- **Respeto por la Autonomía Humana**: Priorizar el reconocimiento y la valoración de los derechos de decisión de los individuos, implementando procesos de consentimiento informado transparentes para empoderar a los participantes en el control del uso de sus datos.
- **Prevención de Daños**: Minimizar éticamente los riesgos asociados con las actividades de la conferencia, enfatizando la responsabilidad de considerar las consecuencias potenciales y abogando por prácticas de investigación que contribuyan positivamente al bienestar social.
- **Equidad**: Asegurar la diversidad, equidad e inclusión a lo largo de la conferencia, buscando activamente perspectivas diversas y fomentando un ambiente que sea representativo y de apoyo para todos los participantes.
- **Modelos Interpretativos**: Enfatizar la importancia ética de presentar modelos de aprendizaje automático de manera interpretable, promoviendo la transparencia y la rendición de cuentas en la implementación de tecnologías complejas.

# Fundamentos

Desde la perspectiva de robustez: Un modelo robusto es aquel que es técnicamente seguro y que cumple con un conjunto de diferentes aspectos.


Transparencia




Privacidad y seguridad

A. Calvo *et al.*, "Safe AI: Ensuring Safe and Responsible Artificial Intelligence," *2023 JNIC Cybersecurity Conference (JNIC)*, Vigo, Spain, 2023, pp. 1-4, doi: 10.23919/JNIC58574.2023.10205749.

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

# Explicabilidad

Por qué es relevante la explicabilidad en la era de la Inteligencia Artificial Confiable?

*La inteligencia artificial explicable (XAI, por sus siglas en inglés)*
*es un conjunto de procesos y métodos que permiten a los usuarios humanos*
*comprender y confiar en los resultados y salidas creados por*
*algoritmos de aprendizaje automático.*

Proposiciones:

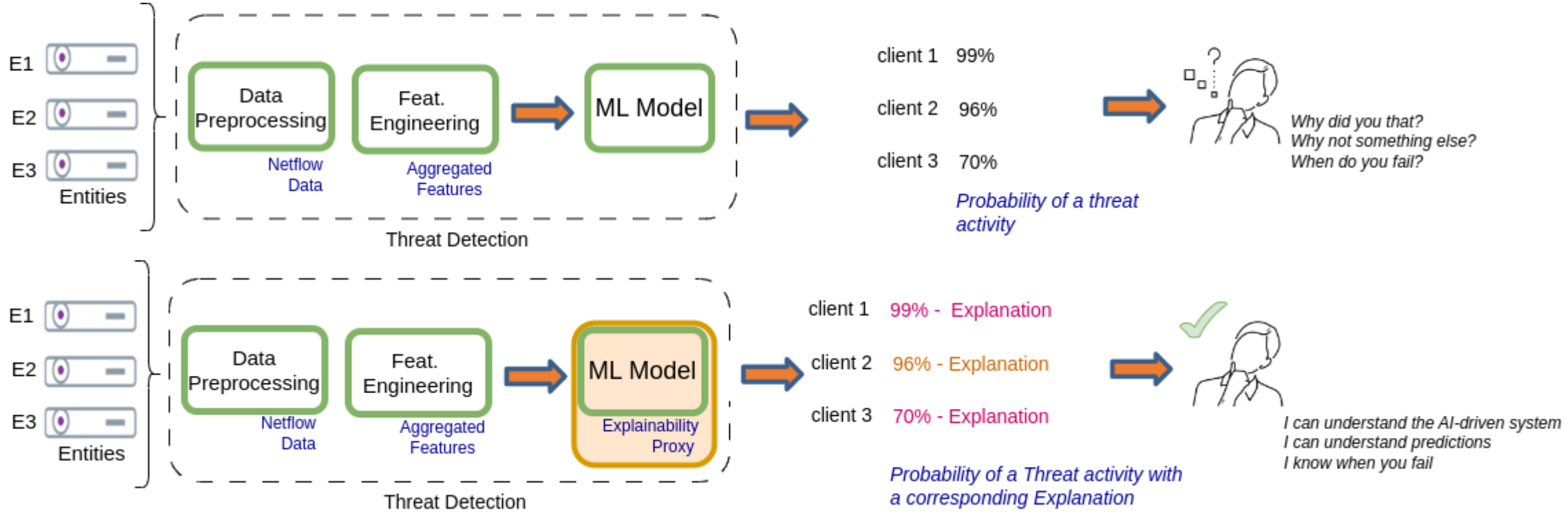
- Comprender el aprendizaje automático podría mejorar las consideraciones éticas.
- Las explicaciones podrían mejorar la robustez.
- La IA explicable podría utilizarse para la rendición de cuentas (cumplimiento).

... La IA explicable es un impulsor para una IA segura.

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

# Explicabilidad

## Entendiendo la Explicabilidad

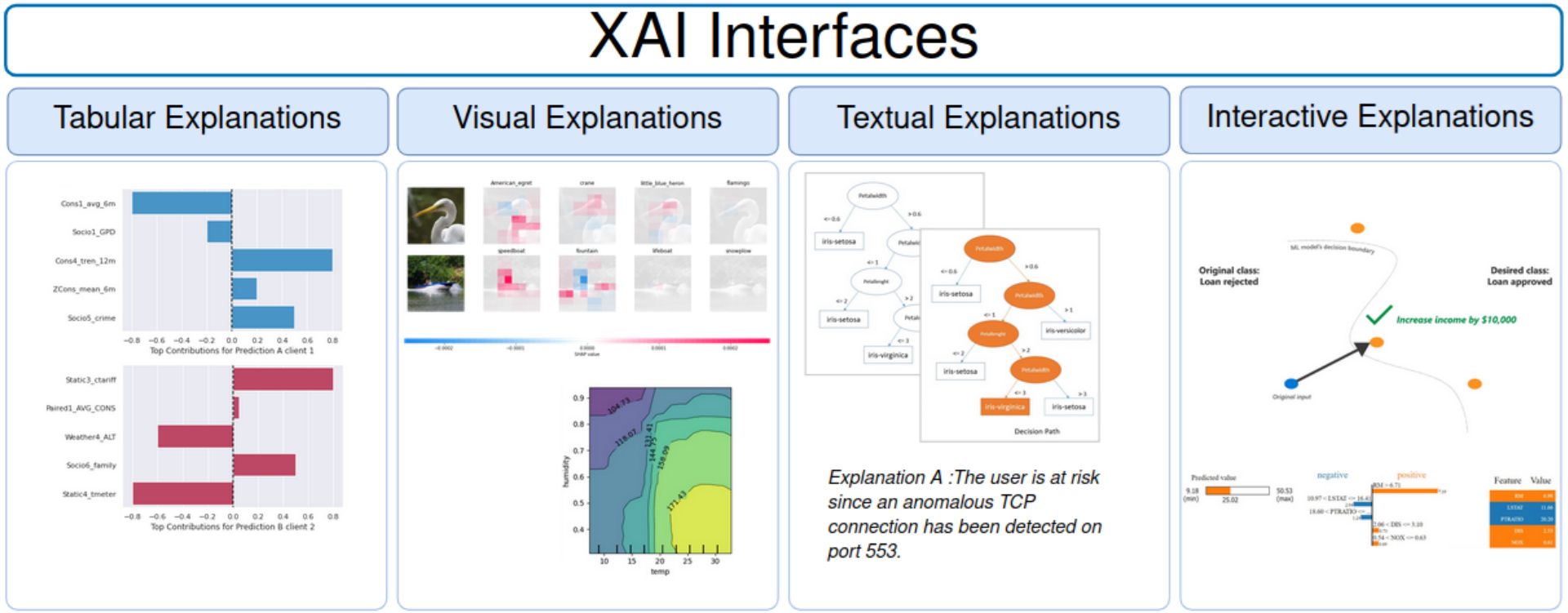



*Explicaión: Unidad básica de interpretación*

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

# Explicabilidad

## Entendiendo la Explicabilidad y las interfaces




Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

# Seguridad y privacidad

## Seguridad a alto nivel

## Los sistemas de IA se clasifican según riesgo

Riesgo inaceptable:

- Incluye sistemas de IA que utilizan, por ejemplo, manipulación subliminal o puntuación social general.

Riesgo Alto

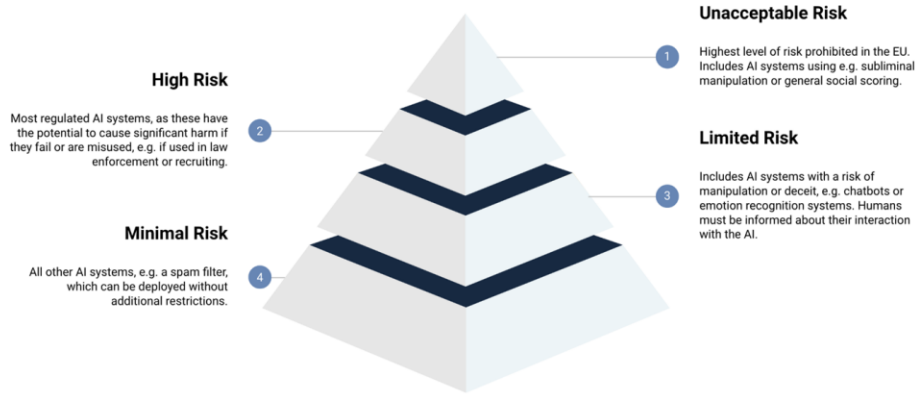
- Sistemas de IA más regulados, ya que estos tienen el potencial de causar daño significativo si fallan

Riesgo limitado

- Incluye sistemas de IA con un riesgo de manipulación como chatbots o sistemas de reconocimiento de emociones.
  Los humanos deben ser informados sobre su interacción con la IA.

Riesgo Mínimo

- Todos los demás sistemas de IA, por ejemplo, un filtro de spam, que se pueden desplegar sin restricciones adicionales.



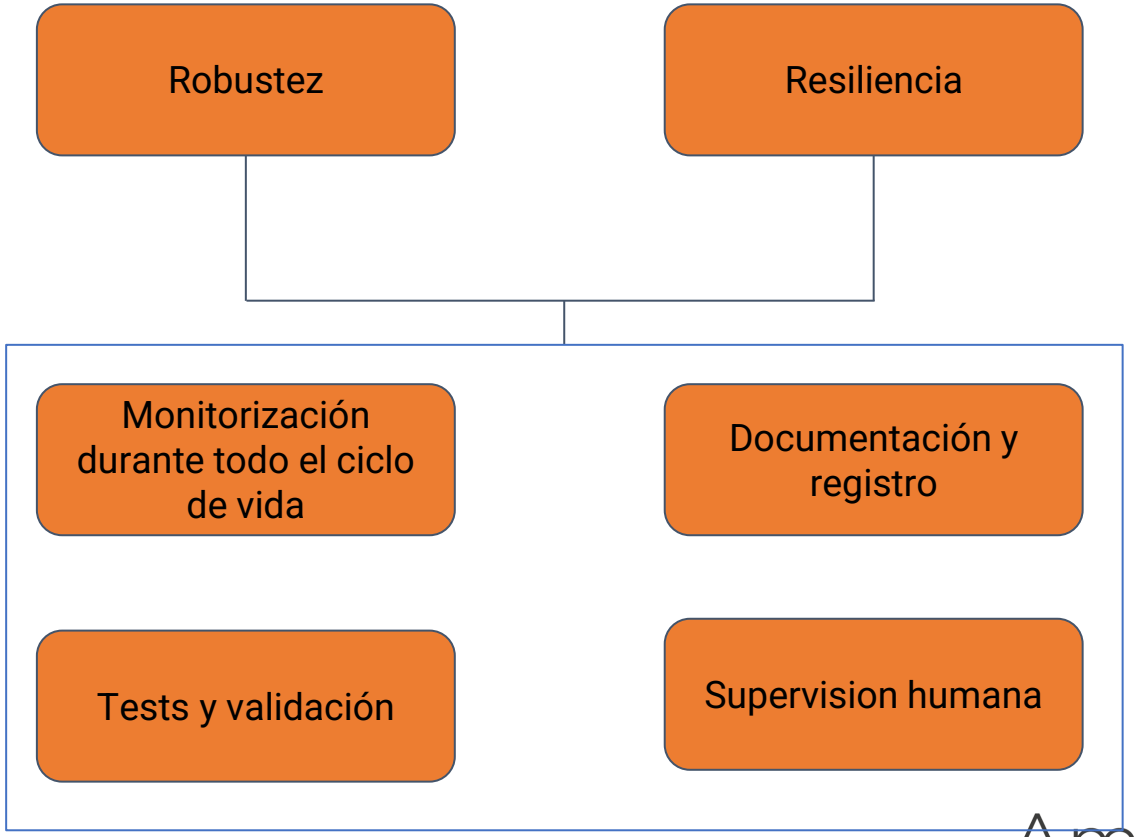

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

# Seguridad y privacidad

## Sistemas de riesgo Alto

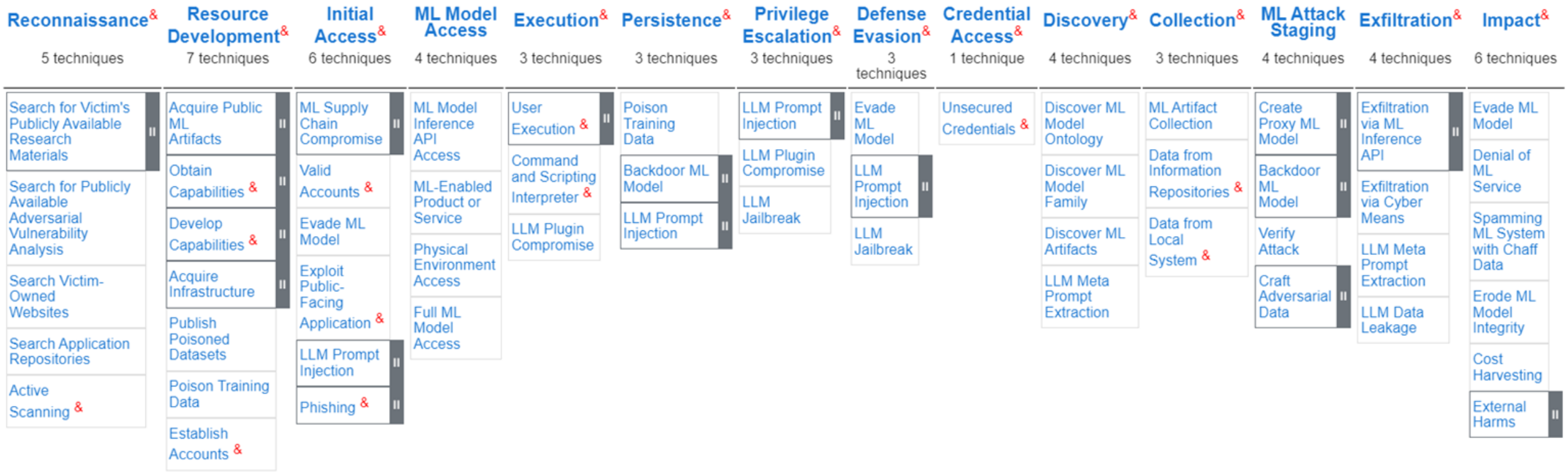
Estos son los sistemas más regulados actualmente

- Salud y Dispositivos Médicos
- Transporte
- Aplicación de la Ley
- Educación y Empleo
- Servicios de Administración Pública
- Procedimientos Legales y de Asilo
- Servicios Financieros



# Seguridad y privacidad

ATLAS MITRE




Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

# Seguridad y privacidad

PRIVACIDAD

Para entrenar grandes modelos necesitamos muchos datos

GPT-2 : ~40GB GPT-3: ~2 TB GPT-4: ~10TB

¿Todos los modelos se entrenan con datos legítimos?

¿Se puede saber si se han usado nuestros datos?

En Europa tenemos el GDPR, ¿nos salva?

La regulación aún está muy verde en este tópico

# Seguridad y privacidad

PRIVACIDAD

Qué datos se usan normalmente?

- Conjuntos de datos públicos
- Web Scraping
- Datos licenciados
- Datos de usuarios
- Datos sinteticos
- Datos abiertos del gobierno
- Datos de dispositivos IoT


Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

# Seguridad y privacidad

## PRIVACIDAD

Casos mas recientes

facebook

OpenAI

Microsoft

NETFLIX

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

# Para ir cerrando…


**W**hole-**B**rain **E**mulation

https://qntm.org/**mmacevedo**

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

# ¡Muchas gracias!

Para cualquier consulta o más información:


wwww.ametic.es

ametic@ametic.es

@AMETIC_es #AMETIC