Ética y Riesgos en Ciberseguridad

Ingeniera Andrea Vera

Jornadas de Investigación IA, ¿moda pasajera o cambio permanente de paradigma?

MESA # 3
ÉTICA Y RIESGOS EN CIBERSEGURIDAD

Andrea Celeste Vera
Ingeniera en Sistemas de la Información
Universidad Tecnológica de Argentina

Entusiasta con más de 20 años de experiencia en Tecnología y Seguridad de la Información. Panelista en varios congresos y conferencias nacionales e internacionales.

Magíster en Ciberseguridad (Universidad de Valencia). Ingeniera en Sistemas de Información. Diplomatura en Seguridad de la Información. Certificaciones internacionales: CISA (Certified Information Systems Auditor) y CEH (Certified Ethical Hacking), ISO/IEC 27001:2022 Internal Auditor.

UNIVERSIDAD DE COSTA RICA

PROSIC Programa Institucional Sociedad de la Información y el Conocimiento

¿QUÉ ES LA ÉTICA DE LA INTELIGENCIA ARTIFICIAL?

Se considera La ética de la inteligencia artificial como una rama de la ética que analiza y evalúa los dilemas morales que se derivan del despliegue de esta tecnología en la sociedad.

Criterios de análisis de la ética

- Datos
- Algoritmos
- Prácticas

La ética en las distintas etapas de IA

PRINCIPALES RETOS ÉTICOS Y EFECTOS NO DESEADOS DE LA ÉTICA DE LA IA

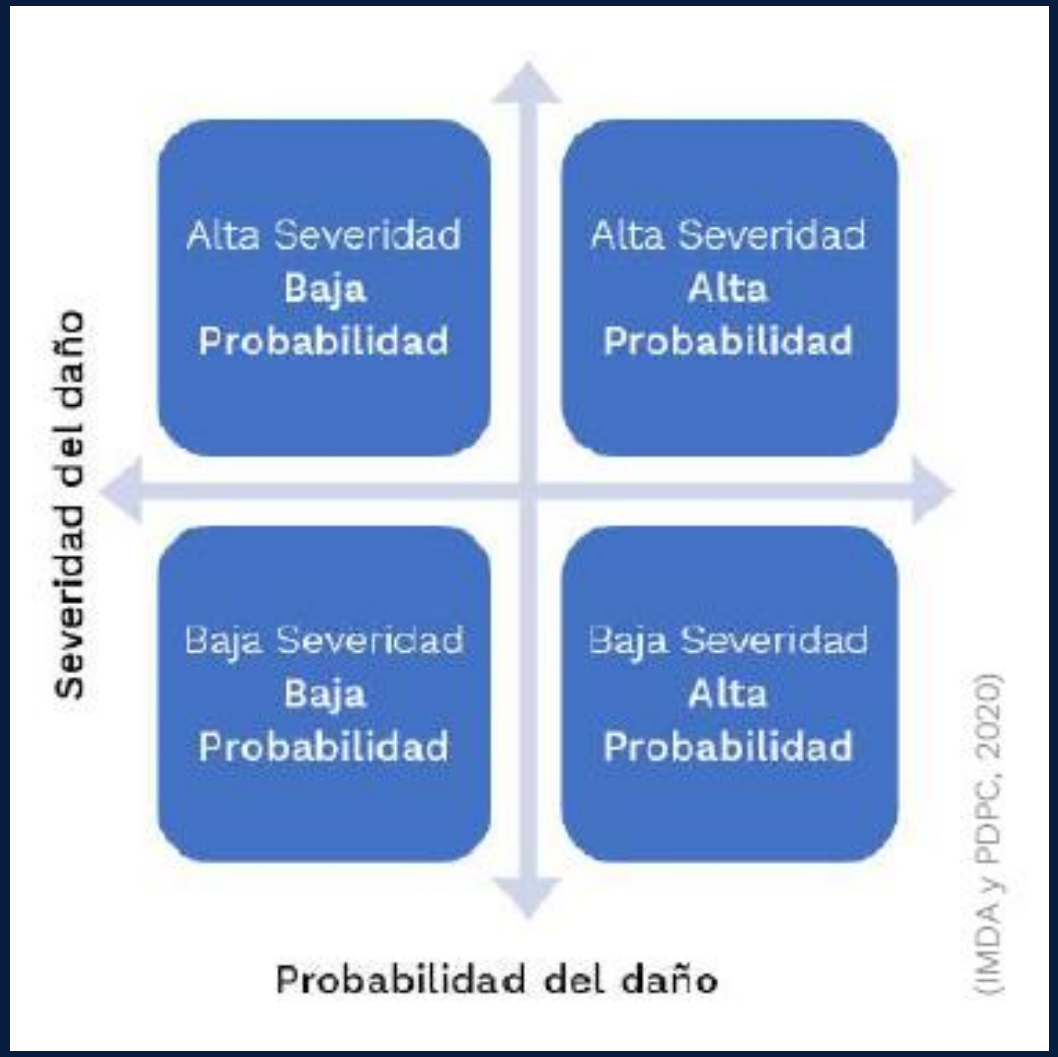
- Posibles amenazas a la libertad de voluntad y responsabilidad
- Sesgos, discriminación y exclusión
- Perfilamiento algorítmico: personalización versus beneficios colectivos
- Buscar un nuevo balance al prevenir bases de datos masivas mientras se aumenta la IA
- Calidad, cantidad y relevancia: el reto de los datos adaptados para IA
- Identidad humana frente al reto de la inteligencia artificial

LOS PRINCIPIOS ÉTICOS PROPUESTOS PARA LA IA

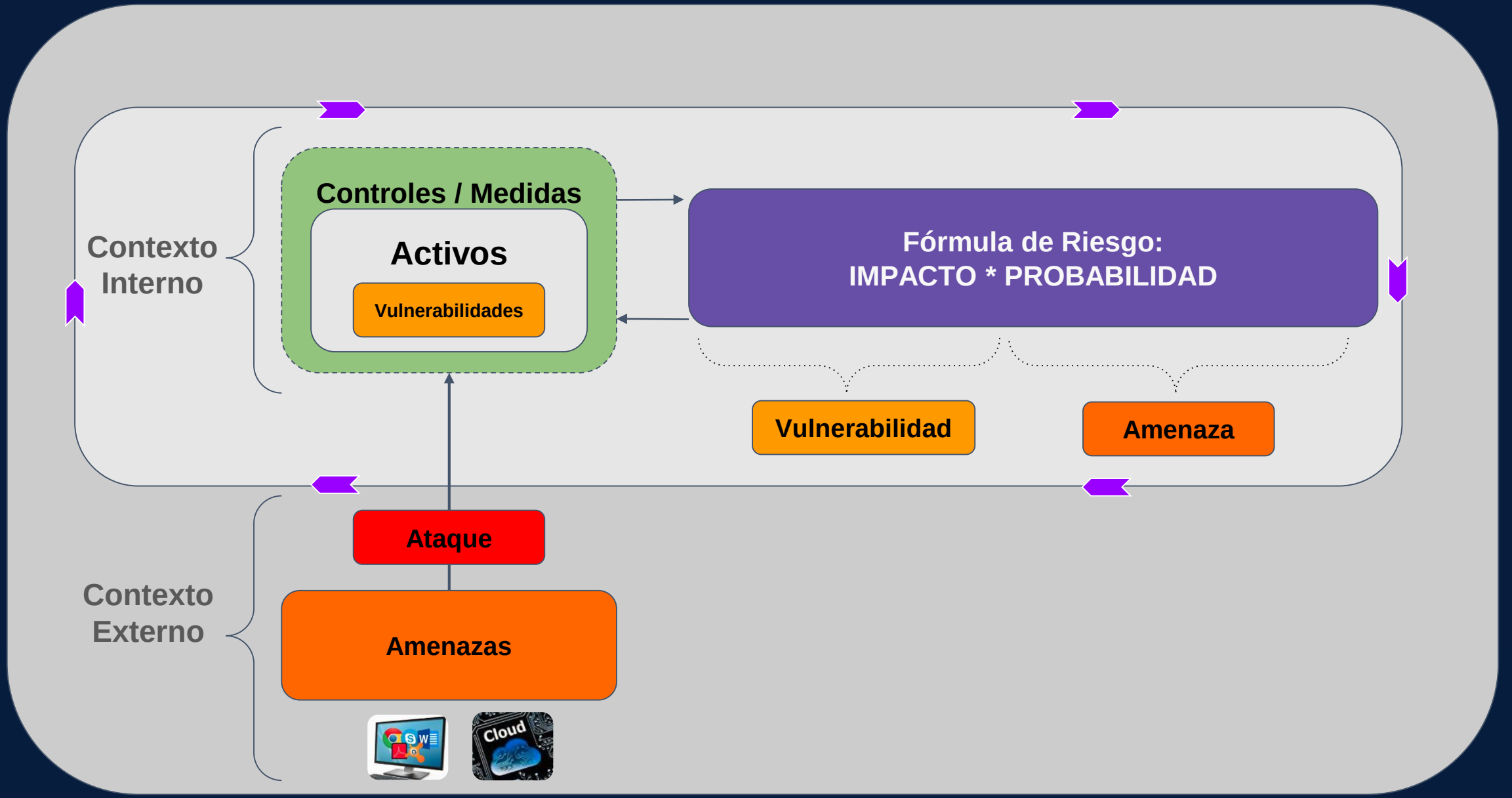
Algunos principios propuestos por la UNESCO

Definición y gestión de riesgos

Clasificación del Riesgo



Clasificación del Riesgo



❖ **RIESGOS PARA LA CONTINUIDAD Y LA CALIDAD DE LAS OPERACIONES**:

» ***Para la seguridad física:*** Algunos sistemas con componentes de IA manejan aspectos asociados a la seguridad física (safety), no sólo de bienes tangibles como las mercancías y las instalaciones, sino incluso de las personas y otros seres vivos. Un error en este ámbito puede llegar a poner en peligro la vida.

» ***Para las operaciones:*** La mala calidad de la información puede llevar a malas decisiones operativas, así los peligros de una información sesgada o incompleta o inventada (riesgo de alucinación) proporcionada por una IA , puede presentar un grave riesgo a las operaciones e incluso a la continuidad del Negocio.

❖ **RIESGOS DE REPUTACIÓN Y CUMPLIMIENTO LEGAL Y ÉTICO**:

» *Riesgo de discriminación:* Daños representativos y de asignación que pueden influir en que se perpetúen los estereotipos y prejuicios sociales.

» ***Automatización y daños ambientales****:* La capacitación y operación de LLM requiere mucha capacidad de cómputo, lo que comporta altos costos ambientales derivados del consumo de energía.

» *Riesgos legales:* Como los derivados de la utilización de datos de carácter personal fuera de las finalidades, o en número desproporcionado o sin respetar la duración acordada, entre otros.

» ***Riesgo de toma de decisiones no éticas:*** Si a la IA generativa se da la posibilidad de tomar decisiones aparece el riesgo de que las mismas sean inadecuadas o no éticas.

❖ **RIESGO DE CONFIDENCIALIDAD:** pueden comprometer la confidencialidad al filtrar información clasificada e inferir información confidencial, esto aplica tanto a datos personales como a cualquier otro tipo de información clasificada.

Frameworks para la evaluación de Riesgos en Ciberseguridad e IA

ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) is a globally accessible, living knowledge base of adversary tactics and techniques against AI-enabled systems based on real-world attack observations and realistic demonstrations from AI red teams and security groups.

https://atlas.mitre.org/

Táctica	Nº Técnicas	Qué intenta el adversario
Reconnaissance	5	Recopilar información sobre el sistema de ML para planificar operaciones futuras (p.ej. buscando todo tipo de información pública de la víctima y lanzando escaneos externos)
Resource Development	7	Establecer recursos que pueda utilizar en sus operaciones (p.ej.: buscar información pública sobre artefactos ML, infraestructura, credenciales, ...)
Initial Access	4	Ganar acceso al sistema ML (p.ej: comprometiendo porciones del sistema ML de la cadena de suministro, abusando credenciales, ...)
ML Model Access	4	Ganar algún tipo de acceso a un modelo ML (p.ej. accediendo a la API de inferencia, influyendo en el modelo mediante el acceso a la ubicación física donde se recopilan los datos y alterándolos en origen, ...)
Execution	2	Ejecutar código malicioso embebido en sistema ML (p.ej. provocando que usuario autorizado lo ejecute sin darse cuenta)
Persistence	2	Lograr mantener la intrusión de forma continuada en el tiempo (p.ej. desplegando un "backdoor" en el modelo)
Defense Evasion	1	Evitar ser detectado por las defensas del sistema (p.ej. creando datos adversarios)

https://owasp.org/www-project-ai-security-and-privacy-guide/

This page is the OWASP AI security & privacy guide. It has two parts:

1. How to address AI security
2. How to address AI privacy

Overview

Welcome to the repository for the OWASP Machine Learning Security Top 10 project! The primary aim of the OWASP Machine Learning Security Top 10 project is to deliver an overview of the top 10 security issues of machine learning systems. More information on the project scope and target audience is available in our project working group charter

Top 10 Machine Learning Security Risks

- **ML01:2023 Input Manipulation Attack**
- **ML02:2023 Data Poisoning Attack**
- **ML03:2023 Model Inversion Attack**
- **ML04:2023 Membership Inference Attack**
- **ML05:2023 Model Theft**
- **ML06:2023 AI Supply Chain Attacks**
- **ML07:2023 Transfer Learning Attack**
- **ML08:2023 Model Skewing**
- **ML09:2023 Output Integrity Attack**
- **ML10:2023 Model Poisoning**

Recomendaciones

1. Usar cualquier marco de gestión de riesgos y aplicar medidas de control de riesgo para:
(i) evaluar y gestionar los riesgos de desplegar IA, incluyendo cualquier potencial impacto adverso para los individuos;
(ii) decidir sobre el nivel apropiado de involucramiento en la toma de decisiones ayudada por IA, y
(iii) manejar el modelo de entrenamiento de IA y el proceso de selección.

2. Hacer mantenimiento, monitoreo, documentación y revisión de los modelos de IA que han sido desplegados, con miras a tomar remedios en caso de ser necesarios.

3. Revisar canales de comunicación e interacciones con los stakeholders para brindar divulgación y canales de retroalimentación efectivos .

4. Asegurar que el personal relevante que lidia con sistemas de IA esté adecuadamente entrenado. Otros miembros del personal cuyo trabajo requiera la interacción con el sistema de IA debe estar entrenado para al menos estar alerta de y sensible sobre los beneficios, riesgos y limitaciones al utilizar IA, para que sepan cuándo alertar a los expertos en la materia dentro de sus organizaciones.

Referencias

- Cabrol, M., González, N., Pombo, C., & Sanchez, R. (2020, Enero). Adopción ética y responsable de la Intelidencia Artificial en América Latina y el Caribe. Retrieved from Interamerican Development Bank: https://publications.iadb.org/publications/spanish/document/fAIr_LAC_Adopci%C3%B3n_%C3%A9tica_y_responsable_de_la_inteligencia_artificial_en_Am%C3%A9rica_Latina_y_el_Caribe_es.pdf

- CNIL. (2018, Mayo 25). Algorithms and artificial intelligence: CNIL's report on the ethical issues. Retrieved from CNIL: https://www.cnil.fr/en/algorithms-and-artificial-intelligence-cnils-reportethical-issues
- Comisión Europea. (2020, Febrero 19). White Paper On Artificial Intelligence - A European approachto excellence and trust. Retrieved from European Commission: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
- Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., . . . Whitt. (2019, Diciembre). AI Now Institute. Retrieved from AI Now 2019 Report: https://ainowinstitute.org/AI_Now_2019_Report.pdf
- Doshi-Velez, F., & Kortz, M. (2017). Accountability of AI Under the Law: The Role of Explanation. Retrieved from Berkman Klein Center Working Group on Explanation and the Law: http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584
- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020, Enero). Principled Artificial Intelligence:
- Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Retrievedfrom Berkman Klein Center for Internet & Society: https://dash.harvard.edu/handle/1/42160420
- World Wide Web Foundation. (2018). HOW ARE GOVERNMENTS IN LATIN AMERICA USING ARTIFICIAL INTELLIGENCE? A proposal for effective and legitimate implementations of AI systems in the public sector. Retrieved from World Wide Web Foundation: http://webfoundation.org/docs/2018/07/AI-in-Latin-America_Overview.pdf
- World Wide Web Foundation. (2018, Septiembre). World Wide Web Foundation. Retrieved from ALGORITHMS AND ARTIFICIAL INTELLIGENCE IN LATIN AMERICA A Study of Implementation by Governments in Argentina and Uruguay: http://webfoundation.org/docs/2018/09/WF_AIin-LA_Report_Screen_AW.pdf

Gracias

CONTACT US


@AndreCelesteV


www.linkedin.com/in/andrea-celeste-vera