Los riesgos de las armas autónomas

Una mirada desde el sur

# Aportes al debate del uso de la Inteligencia Artificial para aplicaciones armamentistas


***M. Vanina Martínez***
*Doctora en Ciencias de la Computación*

***Ricardo O. Rodríguez***
*Doctor en Ciencias de la Computación*


Desde el comienzo de su historia, el mundo se ha enfrentado a tres problemas principales: la hambruna, la peste y la guerra. A lo largo del tiempo, la humanidad ha ido organizando y desarrollando herramientas y procesos para superarlos. En tal sentido la Inteligencia Artificial (IA) es uno de los últimos utensilios inventados que, como el fuego, la rueda, las máquinas de vapor o la energía atómica, tendrá un impacto disruptivo en el futuro de la sociedad.

Hoy día la IA juega un rol importante en la llamada Agricultura de Precisión donde los cultivos son controlados remotamente. También, ha tenido un papel importante en el control de la pandemia por Covid-19 ya sea mediante las aplicaciones de seguimiento de contagios como en el desarrollo de las vacunas. Finalmente, la IA es la impulsora de una nueva carrera armamentista que es necesario controlar.

En este documento nos proponemos plasmar lo que consideramos cuestiones técnicas centrales para el debate del control de las armas autónomas letales (LAWs). Somos conscientes que, en el estado actual de la frontera del conocimiento de la IA, nuestras opiniones pueden generar más dudas que certezas. En ese sentido trataremos de honrar la máxima de J.L.Borges: "La duda es otro de los nombres de la inteligencia".

El objetivo final de este debate es la creación de un marco jurídico que, por un lado, prohíba el uso de los sistemas de armas completamente autónomas, y por otro, que regule y controle el desarrollo de sistemas de armas autónomas con control humano significativo. Se busca así evitar un futuro deshumanizado, donde los sistemas armados pueden decidir matar y aplicar la fuerza sin que la autoridad militar comprenda o sea plenamente responsable de las consecuencias de dicha acción.

# Un poco de contexto

**Técnico.**

En una visión rápida, podemos decir que los algoritmos son implementaciones de conceptos o modelos. Por ejemplo, tenemos el concepto de "suma aritmética" y tenemos diferentes implementaciones o formas de hacerlo. Por el contrario, la mayoría de los algoritmos de aprendizaje automático (MLA por sus siglas en inglés) se pueden establecer a grandes rasgos como:

"Un algoritmo que busca aproximar un concepto/modelo a través de una cantidad limitada (pero que puede ser muy grande) de datos/ejemplos de este"

En este sentido, los MLA son meta-algoritmos, es decir, generan una implementación del concepto/modelo implícito dado por datos/ejemplos. De esta manera, con un MLA se obtendría una conceptualización de la suma aritmética a partir de muchos ejemplos de sumas (es decir, dos sumandos y su resultado). El supuesto subyacente en este procedimiento es que, si le damos a un MLA una cantidad suficiente de ejemplos, este obtendrá un procedimiento equivalente a nuestro conocido método escolar de suma.

En general, la gente no aprende a operar con números ni con ejemplos. Sin embargo, aprendemos a hablar con ejemplos. Un MLA encuentra patrones y regularidades en el conjunto de ejemplos y genera un modelo a partir de ellos. En este momento, es importante señalar, que gran parte de los descubrimientos científicos producidos en el siglo XVII fueron por observaciones.

Actualmente, los modelos virtuales generados por MLA, dependiendo del dominio de aplicación, suelen tener un mejor rendimiento que los físicos. De hecho, hay varias tareas de modelización donde los MLA producen un resultado mejor que el producido por un experto humano. Un ejemplo de esto es la conceptualización de tumores en tomografía [8].

Esencialmente, los principales desafíos de esta tecnología involucran superar tres tipos de problemáticas. El primer tipo es acerca de la calidad y la cantidad de datos necesarios (etiquetado y sesgo). El segundo es sobre el proceso de obtención del modelo (sobreajuste, entrenamiento vs testing, optimización, translación, etc.). El último trata sobre el modelo obtenido (interpretabilidad, predictibilidad, explicabilidad, transparencia, rendición de cuentas, etc.)

La precisión y la eficiencia no conforman la debilidad central de MLA porque ambas van incrementándose en forma acelerada. Por el contrario, preferimos señalar la predictibilidad y la comprensión como los desafíos centrales de los MLAs. El primero está limitado por un aspecto teórico [1]. Al segundo le dedicaremos una sección.

A los efectos de esta presentación dividiremos a la IA en dos grupos: los algoritmos basados en conocimientos (KBA) y aquellos basados en datos (MLA). Estos últimos conforman la técnica más exitosa en la actualidad y sobre ellos centraremos nuestro análisis.

## Político-diplomático

En el ámbito político-diplomático, hay muchos antecedentes que deben ser tomados en cuenta al contextualizar el debate actual sobre los sistemas de armas autónomas letales (LAWS). En primer lugar, los cuatro grandes tratados (bilaterales y multilaterales) de control de armas: el Tratado de Misiles Antibalísticos (ABM), el Tratado sobre Fuerzas Nucleares de Rango intermedio (INF), el Tratado de Cielos Abiertos y los Tratados de Reducción de Armas Estratégicas (STARTs). Todos ellos tendientes a reducir las fricciones Este-Oeste. Lamentablemente, los mismos han concluido o no serán ratificados, lo que da lugar a un futuro de conflictos armados muy incierto. En segundo lugar, podemos citar antecedentes más auspiciosos, como el Tratado de Prohibición de Armas Nucleares (que entró en vigor el 22-01-2021) y las Convenciones sobre Armas Biológicas y Químicas, Municiones de Racimos, Minas Antipersonales, y Láseres Cegadores. Todos esos antecedentes, con sus avances y retrocesos, marcan el esfuerzo internacional por regular el desarrollo, producción, comercialización y uso de distintos tipos de armas. Sobre esa base se están desarrollando las discusiones sobre LAWS.

En general, los Estados partes de la Convention on Certain Conventional Weapons (CCW) están de acuerdo en que “la identificación y búsqueda de un entendimiento común sobre los conceptos y las características de los LAWS podría ayudar a una mayor consideración de los aspectos relacionados con las tecnologías emergentes en el área de los LAWS” (Informe 2019 19 b.). Con este fin, los Estados partes continúan buscando más aclaraciones y un entendimiento compartido sobre las características técnicas específicas que constituyen o definen la “autonomía” o las capacidades autónomas de “toma de decisiones” en los sistemas de armas. Hasta ahora, los Estados han señalado que ciertas capacidades, incluidas las “funciones autónomas en la identificación, selección o participación de un objetivo” se encuentran “entre las características esenciales de los sistemas de armas basados en tecnologías emergentes en el área de los sistemas de armas autónomos letales” (Informe de 2018 19 a.). Pero no han llegado a un entendimiento compartido de una definición técnica de autonomía, y han destacado “la autoadaptación; previsibilidad; explicabilidad; fiabilidad; capacidad de estar sujeto a intervención; capacidad para redefinir o modificar objetivos o

metas o adaptarse al medio ambiente; y capacidad para autoiniciarse" como características técnicas de las armas autónomas que "pueden beneficiarse de una aclaración o revisión adicional" (Informe 2019, 20 a., b.).

Actualmente, en el contexto de la CCW existe un Grupo Gubernamental de Expertos (Group of Governmental Experts) que recibió un mandato para que entre el 2020-2021 produzca un conjunto de "recomendaciones consensuadas en relación con los distintos aspectos del marco normativo y operativo de las tecnologías emergentes en el área de sistemas de armas autónomas letales (LAWS)" *(ver https://dig.watch/process/gge-laws)*. En ese espacio de reflexión y discusión están representados diferentes grupos con intereses contrapuestos: Organismos Internacionales, ONGs, Gobiernos, Fuerzas Armadas, la industria armamentista y los científicos. Justamente nuestra pretensión aquí es aportar elementos sólidos desde el sector académico a ese debate.

En ese espacio de reflexión y discusión están representados diferentes grupos con intereses contrapuestos

En tal sentido nosotros creemos que técnicamente dicho debate no debería ser sólo circunscripto al respeto del Derecho Internacional Humanitario-DIH (International Humanitarian Law-IHL) dado que eso restringe su análisis en el contexto del campo de batalla en conflictos regulares pero no considera otros aspectos sumamente importantes como su diseño, fabricación y comercialización.

Por otra parte, consideramos que en relación al DIH, el mayor desafío se cierne sobre garantizar el principio de *responsabilidad*. Otras cuestiones como el principio de la *proporcionalidad*, de *oportunidad*, *precaución o distinción*, que son muy importantes para evaluar una acción militar, son potencialmente bien parametrizables en un LAWS desde un punto de vista tecnológico. Aunque reconocemos que estos principios requieren evaluaciones complejas basadas en las circunstancias imperantes en el momento de la decisión de atacar y también durante un ataque, consideramos que los mismos pueden ser capturados por reglas lógicas y métricas de performance adecuadas.

Finalmente, creemos que hay que ser muy cuidadosos en evitar tecnicismos irrelevantes. Por ejemplo, como ya se ha demostrado en el caso de las minas antipersonales, resulta fundamental definir con precisión el objeto de prohibición/regulación. Pero dada la complejidad propia del tipo de sistemas que estamos analizando, parece inadecuado, poner el foco de dicha definición en si el sistema utiliza sensores para determinar dónde y cuándo se utilizará la fuerza (ver [7]). Creemos que este no es un buen discriminante y no se alinea con la complejidad de los sistemas de armas basados en IA tanto existentes como a desarrollarse en el mediano plazo. Un primer problema con esta caracterización es que en general los sensores están en el lugar físico donde se despliega la acción, pero la decisión puede tomarse a gran distancia de la ejecución mediante datos que no vienen directamente de los sensores sino de un preprocesamiento de ellos y a veces son incluso datos históricos que pueden ni siquiera estar relacionados con las entradas de los sensores del sistema. Por eso, en el caso de las LAWS, creemos que es mejor definirlas de forma multi-intencional y quizás más bien en términos conceptuales como ser: sus capacidades de aprendizaje, percepción, toma de decisiones y actuación, que luego pueden ser llevados a cabo

física o digitalmente de distintas maneras [10]; por ejemplo, por medio de aprendizaje a base de datos históricos, sensores que permitan capturar la situación actual, módulo de interpretación y toma de decisión basado en reglas y actuadores físicos.

# Aportes

Nuestros aportes al debate de los LAWS en este documento se centrarán en los tres ejes que consideramos más relevantes y urgentes en la discusión:

- Control humano significativo.
- Previsibilidad y confiabilidad.
- Comprensibilidad, explicabilidad e interpretabilidad.

Antes de empezar a describir estas tres cuestiones vamos a plantear algunos reparos que despiertan los sistemas de tomas de decisiones automáticos. Un punto importante sobre la naturaleza de estos reparos es que los mismos no necesariamente se resuelven con una mejora en la precisión cualitativa/cuantitativa de los MLAs. Un primer punto importante y que responde a la esencia misma de estos sistemas es que identifican correlación y no causalidad, es decir, descubren "síntomas" pero no "motivos". Esto conlleva una pérdida de confianza y la posibilidad de comportamientos impredecibles. También se cuestiona el hecho que las decisiones se automaticen disminuyen las posibilidades de impugnación legal y éticas de los resultados. En [7] se hace hincapié en el desafío que las LAWS imponen a los valores humanos, e identifican además de los mencionados: la deshumanización, la falta de comprensión de cómo los sistemas funcionan. Por ende pueden redundar en que no sea posible establecer un control humano significativo sobre los mismos, los riesgos a la paz y la seguridad dado que la autonomía y la activación remota "facilita" correr el riesgo; por un lado, debido a que los umbrales políticos contra el uso de la fuerza se reduzcan significativamente y, por otro, a causa de que se incite a respuestas automatizada que escalen rápidamente en una carrera armamentista.

### Eje 1: Control humano significativo

Desde el comienzo de las negociaciones mencionadas más arriba en el 2014, los Estados han propuesto que el foco del debate debía ponerse sobre la cuestión del control humano, y sobre cómo debería implementarse sobre las armas, las funciones críticas de estas, los ataques, los procesos de selección de objetivos y las decisiones (finales) sobre el uso de la fuerza, etc. Si bien, la mayoría de los estados están de acuerdo en que el control humano debe ser más significativo que la mera posibilidad de abortar un ataque en el último momento, no hay tal unanimidad al momento de determinar cómo debe definirse y aplicarse el papel humano en el uso de la fuerza (letal). En este

apartado trataremos de ordenar los principales puntos en discusión y bosquejar algunas propuestas metodológicas. Para eso es necesario introducir algunos conceptos preliminares.

En general, pueden ser reconocidos tres niveles de mandos (ver [2]):

- **Mando estratégico**, que traslada el propósito político en objetivos militares.

- **Mando operacional**, que convierte los objetivos generales del nivel estratégico en tareas concretas para las fuerzas tácticas.

- **Mando táctico**, que dirige el uso específico de las fuerzas militares en las operaciones para ejecutar las tareas decretadas por el mando operacional. El mando táctico abarca el despliegue de unidades, plataformas, personal no perteneciente a una unidad constituida y sistemas de armas que pueden estar en contacto directo con las partes de un conflicto. En [3], este nivel es llamado "mando de misión" (Mission Command).

Los MLAs pueden ser usados, en principio, a lo largo de todo el proceso de toma de decisión de una acción militar. En la actualidad, los altos mandos e intermedios utilizan herramientas de IA para análisis de datos y toma de decisiones. Sin embargo, consideramos que a nivel estratégico y operacional no hay posibilidades reales que, en el corto y mediano plazo, los MLAs tomen decisiones con un nivel de autonomía que atente contra la dignidad humana. Por tal razón, es que nos concentramos en lo que se refiere a su uso a nivel táctico. Por otra parte, es importante señalar que las propias cúpulas militares parecen tener reparos a delegar sus atribuciones (ver [3]) a un nivel más alto que el de mando de misión.

También es importante identificar los cuatro tipos de control que suelen ponerse en cuestión. Ellos son (ver [2]) :

- **Control total** ("*human full control*"), donde el sistema no toma ninguna decisión por sí mismo sino que es teledirigido.

- **Control en el circuito** ("*human in the loop"*), el sistema implementa la tarea ordenada con autonomía, pero requiere la ***intervención*** humana para validar e implementar acciones.

- **Control sobre el circuito** (*"human on the loop"*), en que el sistema implementa la tarea ordenada en autonomía bajo la ***supervisión*** de operadores humanos que pueden, si es necesario, corregir o abortar una acción específica.

- **Autonomía total** ("*human off the loop"*), tal que el sistema implementa la tarea ordenada sin supervisión o intervención humana alguna.

El circuito (loop) suele referirse a la secuencia de tareas (alguna de las cuales puede no estar en una acción específica) en que suele dividirse la ejecución de una misión:

- **Búsqueda**: encontrar el objetivo, recopilar información y realizar la inteligencia sobre el campo de batalla.

- **Localización**: detectar y confirmar con precisión la ubicación y estimar el tiempo disponible.

- **Seguimiento**: mantener una identificación positiva y actualizar la información sobre el objetivo y su entorno.

- **Chequeo**: valorar las reglas de enfrentamiento (ROE), daños colaterales, y riesgos para las fuerzas propias.

- **Enfrentamiento**: ejecución del ataque en base a conjuntos de objetivos autorizados, restringidos, y prohibido, con capacidad de suspensión y cancelación.

- **Evaluación**: valoración de la efectividad del ataque y determinación si es necesario un nuevo ataque y en qué condiciones.

Con todos estos elementos introducidos estamos en condiciones de desarrollar las principales cuestiones sobre el Control Humano Significativo.

La cuestión central es garantizar que los humanos sean los responsables últimos por las consecuencias de una operación militar aún cuando la distancia temporal y espacial entre la directiva de la acción está tan alejada de la propia ejecución de la misma en el campo de batalla. Este distanciamiento, y la imprevisibilidad de las consecuencias que trae, a su vez, suscita preocupaciones sobre la aplicación del derecho internacional humanitario, aceptabilidad ética y eficacia operativa. Es por eso que se busca definir, ¿qué tipo y grado de control humano se requiere en la práctica, para asegurarnos que los humanos continúen desempeñando su papel necesario en las decisiones de usar la fuerza en ataques específicos en conflictos armados, independientemente de la sofisticación de la tecnología, mientras se cumplen los requisitos legales, éticos y operativos? Todas las partes involucradas en la discusión reconocen que los humanos deben ejercer algún tipo de control sobre las armas y el uso de la fuerza en ataques específicos en conflictos armados. Donde se divergen es en cuestiones de *cómo* y *cuándo* los seres humanos deberían ejercer ese control en contextos operativos. Como se destaca en el GGE, es probable que haya algunas medidas de control que puedan aplicarse en todas las circunstancias y otras cuya necesidad dependa del contexto.

En [4] se identifican tres tipos de medidas de control:

- Sobre los parámetros de uso del sistema de armas, incluidas medidas que restringen el tipo de objetivo y tarea para la que se utiliza el LAWS; colocar límites temporales y espaciales en su funcionamiento; restringir los efectos del LAWS; y permitir mecanismos de desactivación a prueba de fallas.

- Sobre el entorno que controla o estructura el uso del LAWS (por ejemplo, usarlos sólo en entornos donde no hay civiles y objetos civiles, o excluir su presencia durante la duración de la operación) .

- Sobre la interacción hombre-computadora, con medidas que permiten al usuario supervisar el LAWS e intervenir en su funcionamiento cuando sea necesario.

Estas medidas de control pueden ayudar a reducir, o al menos compensar, la imprevisibilidad inherente al uso de LAWS y mitigar los riesgos, en particular para los civiles. Desde una perspectiva legal, un operario de una LAWS debe ejercer un control suficiente para tener una certeza razonable sobre los efectos de su utilización en un ataque y poder limitarlos según lo exige el DIH. Por supuesto, las consideraciones éticas pueden exigir restricciones adicionales, especialmente en el ejercicio de la fuerza contra civiles. En cualquier caso, la implementación de estas medidas debe respetar un equilibrio entre asegurar el cumplimiento legal, la aceptabilidad ética y la utilidad operativa.

En la sección 2 de [4], se presentan tres dificultades que presentan los LAWS para cumplir con el DIH: número, contexto e imprevisibilidad. Hoy en día, los juicios de valor para verificar su cumplimiento, que también reflejan consideraciones éticas, forman parte del entrenamiento regular de las fuerzas armadas. Nosotros entendemos que la mayoría de los cuestionamientos sobre los juicios cualitativos o evaluativos que deben realizar en tiempo real, en un ambiente complejo y dinámico, para garantizar el respeto de los principios de distinción, proporcionalidad y precauciones del DIH, son tecnológicamente superables. Más aún, creemos que el definir estándares y "parametrizaciones" sobre la valoración subjetiva del resguardo de estos principios puede ser útil para repensarlos. En tal sentido consideramos que la práctica profesional de un militar podría, al menos parcialmente, ser emulada por un sistema de IA en un futuro no muy lejano. Al punto tal de poder pasar un test de Turing adaptado. La tecnología actual de IA no lo permite todavía, pero, desde nuestro entendimiento, es sólo una cuestión de tiempo. Por eso es tan importante definir criterios de diseño, métricas de validación, predictibilidad y comprensibilidad, y, finalmente, mecanismos de auditoría continua, para que cuando la tecnología exista, exista de manera tal que pueda ser confiable y previsible, pero sobre todo medible y auditable. Por supuesto que esto es un verdadero desafío, y llevará un esfuerzo muy importante, pero técnicamente es posible superar las tres dificultades antes mencionadas. Eso no quita que por razones políticas se decida preservar una porción relevante del control a los humanos. También juzgamos que es política la determinación de establecer estándares de performance técnica diferentes para humanos y sistemas de IA.

Sólo para fijar nuestra posición al respecto digamos que en la subsección de [4], "Application of IHL in practice: what requirements for human control?", se presentan dos posiciones surgidas del workshop de expertos de junio del 2019. Los tecnicistas consideran que el único límite a la autonomía es la tecnología: cuanto más "sofisticada" sea el LAWS, más tareas se le pueden asignar y menos control del usuario será necesario durante un ataque. Los humanistas, que, independientemente de las características técnicas del LAWS, las normas del DIH sobre la conducción de las hostilidades exigen juicios contextuales y basados en valores por parte de las personas que planifican, deciden y ejecutan los ataques.

Nosotros consideramos que esta posición es maniquea. El hecho que los LAWS puedan adquirir características cada vez más "sofisticadas" no es justificativo para concederles el poder de autonomía completa. Pero tampoco es completamente correcto el argumento de que las DIH requieren intrínsecamente que sean personas las que tengan el control absoluto. Otra vez, a nuestro parecer, la decisión es política y no debería esgrimirse cuestiones legales o técnicas que no hacen al fondo

de la cuestión.

Si coincidimos con los autores de [4] en que ambos enfoques pueden compatibilizarse en las tres medidas prácticas de control mencionadas más arriba en un abordaje simultáneo e incremental de todos ellos. Con esto queremos decir que deben ser tomados en forma conjunta, coordinada y expandiendo sus alcances. Por ejemplo, si se parametriza sólo el ataque a objetivos militares, debería también establecer el contexto permitido de la acción, como duración y localización, y los niveles de comunicación/intervención hombre-computadora. Y todo esto dependiendo de la etapa de la acción militar que está en proceso. No es esperable requerir el mismo nivel de control durante la búsqueda, localización y seguimiento, que durante el propio enfrentamiento. También sería razonable que en una primera etapa no se permita que una LAWS defina los objetivos globales de una misión.

Respecto de las cuestiones éticas no vamos a pronunciarnos en este documento. No porque no nos parezcan importantes, sino porque las cuestiones éticas y morales sobre este dominio no han alcanzado la madurez suficiente para relevarlo. Lo que sí podemos decir, es que la mayoría de los trabajos que han incursionado en esta temática, asumen un enfoque deontologistas o normativos. En tal sentido, podemos afirmar que los abordajes a las cuestiones éticas que surgen de dicha escuela filosófica son pasibles de incorporarse a un LAWS. Nuevamente, el factor limitante no es el tecnológico sino de otra índole, antes político, en este caso filosófico.

Para cerrar esta sección nos gustaría reproducir el siguiente cuadro de [4] que resume los aspectos de control que podrían integrarse en el ciclo de vida de una LAWS tal como lo discutimos previamente:

## ¿Que medir y cuando?

| | Control sobre parámetros | Control sobre el ambiente | Control sobre la interacción H/C |
|---|---|---|---|
| **Uso** | | • Complejo entendimiento del contexto. | • Garantizar supervisión.<br>• Capacidad de intervención y desactivación. |
| **Desarrollo** | • Fijar parámetros de teimpo y lugar.<br>• Condiciones de aplicación de fuerzas. | • Fijar límites de objetivos y personas.<br>• Fijar límites temporales y zonas de exclusión. | • Mecanismos de explicación.<br>• Incluir mecanismos de supervisión, intervención y desactivación. |
| **Diseño** | • Límite de tipo y perfiles de objetivos.<br>• Controles espacio-temporales. | • Mecanismos de advertencia.<br>• Seguros e inviolables. | • Diseñar mecanismos S/I/D<br>• Orientado al usuario.<br>• Rindan cuando expliquen. |

## Eje 2: Previsibilidad y confiabilidad

A medida que crecen los ámbitos y los modos de aplicación de técnicas de IA, también se incrementan las discusiones sobre el impacto que pueden tener en nuestra sociedad el uso de sistemas de IA. Como mencionamos previamente, muchas de estas críticas están basadas sobre la naturaleza misma de los sistemas de IA que se utilizan hoy día, es decir, aquellos que utilizan ML que habilitarían, en principio, la posibilidad de comportamientos impredecibles. En términos intuitivos, la previsibilidad es la medida en que se puede anticipar los resultados o efectos de un sistema. En [4], se señala a la previsibilidad como un factor indispensable para cumplir con las normas del DIH, en particular con la prohibición de ataques indiscriminados, el principio de proporcionalidad y el requisito de tomar precauciones en los ataques, dado que los operadores deben ser capaces de limitar los efectos de las armas que utilizan.

En [5] se identifican tres aspectos de este principio que discutiremos a continuación para comprender el concepto en profundidad y se relación con otros criterios y conceptos que relevaremos en este artículo.

Desde el punto de vista técnico, la previsibilidad se ve como la capacidad de un sistema de ejecutar una tarea con el mismo rendimiento que ya haya exhibido en pruebas previas o aplicaciones anteriores. Para los sistemas de ML esto implica que el sistema tiene que poder funcionar de la misma manera que funcionó en su etapa de entrenamiento. En sistemas de computación esto está directamente relacionado con la correctitud de los mismos.

La propiedad de correctitud hace referencia a que el sistema entregue el resultado que se espera dada la configuración de entrada para una tarea específica. Por ejemplo, si el sistema tiene como tarea diferenciar personas de otros objetos que se le presentan por medio de un *stream* de video o una sucesión de fotos, una medida de correctitud posible (que se establece en la etapa de validación del mismo) es calcular a la tasa de falsos positivos y/o la de falsos negativos, es decir, de todas las personas que se le mostraron a cuantas identificó correctamente como personas, y a cuantos de los objetos que se le presentaron los identificó (erróneamente) como personas, respectivamente. Esta es sólo una de las muchas métricas[1] de calidad y desempeño se usan comúnmente para algoritmos de machine learning. Determinar cuál es la métrica más adecuada, depende del problema específico a resolver; por ejemplo, una alta tasa de falsos positivos puede no ser tolerable para una aplicación que trata de identificar NNs en un base de datos de personas extraviadas, pero es crítico para un LAWS o un sistema que identifica criminales o sospechoso.

Entonces, desde el punto de vista de la ingeniería de software, la previsibilidad se puede pensar como una función que tiene en cuenta la correctitud del sistema, i.e. el grado con el cual el sistema puede replicar y reproducir la misma correctitud a lo largo del tiempo, y la medida en la cual el sistema puede adaptarse para permitir el procesamiento de datos diferentes a los que fue expuesto en el entrenamiento y validación, sin perder rendimiento.

---

*1* *https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234*
https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Esto último es particularmente importante para los sistemas de aprendizaje automático, ya que una selección inadecuada del conjunto de entrenamiento y validación puede generar que el sistema "memoriza" en lugar de "aprender", es decir que no pueda generalizar los patrones encontrados en datos que divergen ligeramente de las instancias ya vistas, a esta situación se la denomina sobreajuste (overfitting) en términos estadísticos.

Por otro lado, desde un punto de vista operacional, la previsibilidad apunta a que sea posible anticipar acciones particulares del sistema en ejecución. Es esperable que cualquier sistema autónomo tenga cierto grado de imprevisibilidad asociada, sobre todo cuanto más compleja es su tarea y más dinámico es el contexto donde opera, ya que es imposible anticipar todas las posibles situaciones a la que el sistema vaya a enfrentar una vez puesto en ejecución. Para ejemplificar mejor este aspecto de la previsibilidad, volvamos al sistema que mencionamos previamente, el cual debe diferenciar personas de objetos a partir de un *feed* de video. Supongamos que ese video proviene de una cámara anexada al sistema que captura en vivo las imágenes del contexto donde se mueve el sistema. Independientemente del modelo de IA que se utilice en tal sistema, no es posible determinar de antemano, en el diseño del sistema, todos los posibles (tipos de) objetos y personas (y eventos externos) con los cuales se puede encontrar el mismo una vez desplegado en un entorno físico, a no ser que este entorno sea absolutamente controlado o con información (casi) perfecta[2].En entornos del mundo real, donde la información del entorno es imperfecta e incompleta, la imprevisibilidad operacional es un problema complejo ya que no solo es difícil anticipar con qué se va a encontrar el sistema, sino que también puede ser muy difícil anticipar cómo va a reaccionar el sistema ante esos eventos.

Aún no existen herramientas robustas que permitan identificar y de manera adecuada las variables relevantes.

En los sistemas cuyo funcionamiento depende de la extracción de patrones a partir de los datos de entrenamiento esto puede ser aún más preocupante. Por un lado, en la práctica ya se ha demostrado que estos sistemas pueden fallar de manera muy impredecible cuando los datos de entrada varían (aun ligeramente) de lo esperado, ya que la manera en la que obtienen sus metas no necesariamente siguen patrones lógicos o razonables para un ser humano [11]. Por otro lado, para poder cuantificar cuan previsible son estos sistemas, debería poder cuantificarse de manera adecuada la calidad de los datos en relación con el entorno de despliegue, para lo cual aún no existen herramientas robustas que permitan identificar y representar de manera adecuada las variables (potencialmente) relevantes para esta tarea.

Finalmente, desde una óptica más global, la previsibilidad, entendida como el grado en que los re-

2 *https://es.wikipedia.org/wiki/Información_perfecta*

sultados y efectos de un sistema pueden anticiparse a su uso, es determinada por muchos distintos factores que también impactan en lo técnico y operacional. Entre los factores más destacables para el foco de este artículo, como ya lo hemos mencionado previamente, están las cuestiones de disponibilidad y calidad de datos (tanto para entrenamiento como para validación), el tipo de la tarea o función específica que el sistema debe resolver, y la interacción con otros sistemas (no solo computacionales) en un entorno dinámico y complejo.

La previsibilidad es un factor necesario, aunque dependiendo del dominio puede no ser suficiente (en la próxima sección analizaremos otras propiedades complementarias), para lograr una relación de confiabilidad[3] con los sistemas autónomos mientras estos están en ejecución. De hecho, dependiendo de los factores que mencionamos en el párrafo previo, poder medir el grado de previsibilidad puede ser vital en el éxito del despliegue de sistemas que involucran una interacción humano-computadora para la toma de decisiones. Específicamente en relación con las armas autónomas, la previsibilidad es sumamente importante para poder ejercer de manera efectiva los distintos tipos de control que discutimos en la primera parte de este artículo. La capacidad del operario de poder evaluar cómo respondería un sistema autónomo en determinada circunstancia es absolutamente necesaria para que éste pueda definir si es factible (y con qué riesgo) que el sistema pueda cumplir la tarea de manera que respete las distintas reglas y normas de combate, el DIH, etc., como así también el modo de operación requerida para que así sea (la configuración de parámetros específicos para la misión, por ejemplo).

Es interesante notar que para que un sistema sea previsible no es necesario que el usuario conozca o entienda su funcionamiento (lo cual es complementario, como lo discutiremos en la próxima sección), la previsibilidad debería poder determinarse en base al análisis de comportamiento del sistema. Los modelos que actualmente respaldan los sistemas basados en IA presentan, como hemos venido discutiendo, muchas dificultades para poder evaluar de manera confiable su previsibilidad. En la actualidad se están realizando extensas investigaciones para mejorar la robustez de los mismos [9].

Estos sistemas imponen también un nuevo desafío en la definición de procesos de diseño y validación o *testing* de los mismos. Si bien existe una amplia variedad de metodologías y herramientas en la ingeniería de software para sistemas de computación tradicionales, para asegurar cierto grado de previsibilidad, robustez y confiabilidad, en la mayoría de los casos no pueden ser aplicadas directamente. Por ejemplo, el área de estudio de métodos formales, se centra en el desarrollo de técnicas matemáticas para describir tanto los sistemas de software y hardware como sus requerimientos (qué se espera que esos sistemas hagan) de manera tal que se pueda probar por medio de operaciones matemáticas que efectivamente el sistema implementado hace exactamente lo que

---

*3 Según la UE, en el documento "DIRECTRICES ÉTICAS PARA UNA IA FIABLE" establece que "la fiabilidad de la inteligencia artificial se apoya en tres componentes que deben satisfacerse a lo largo de todo el ciclo de vida del sistema: 1) la IA debe ser lícita, de modo que se garantice el respeto de todas las leyes y reglamentos aplicables; 2) también ha de ser ética, es decir, asegurar el cumplimiento de los principios y valores éticos; y, finalmente, 3) debe ser robusta, tanto desde el punto de vista técnico como social, puesto que los sistemas de IA, incluso si las intenciones son buenas, pueden provocar daños accidentales". URL: Ethics guidelines for trustworthy AI | Shaping Europe's digital future (europa.eu)*

se supone que debe hacer. Estas técnicas se usan exitosamente en general para sistemas cerrados de complejidad controlada y de alto riesgo. Sin embargo, tienen graves problemas de eficiencia a medida que los sistemas se vuelven más complejos; más allá de esto, no aplican directamente a sistemas de ML y no está claro que efectivamente se pueda lograr algo similar en el corto y mediano plazo. Por otro lado, tampoco la plétora de herramientas de *testing* que existen para sistemas tradicionales se generalizan adecuadamente para los sistemas basados en IA. Uno de los problemas más importantes es poder generar casos de test relevantes al comportamiento del sistema fuera del ámbito de diseño, de manera tal de poder reproducir lo más fehacientemente posible el entorno con el cual se va a encontrar una vez desplegado, sobre todo si se espera que el sistema funcione por ejemplo en un entorno físico y abierto (como claramente sucedería para un arma autónoma). Las herramientas de simulación pueden jugar un rol muy importante en esto ya que no es necesario, en principio, enfrentar al sistema a una situación real como puede ser un campo de batalla, sin embargo, el problema de poder predecir situaciones imprevistas sigue estando, ya que el entorno debe ser simulado en un sistema computacional en sí mismo.

### Eje 3: Comprensibilidad, explicabilidad e interpretabilidad

Otras propiedades, complementarias a la previsibilidad, han sido identificadas como importantes al momento de generar confiabilidad en un sistema basado en IA. En esta sección analizaremos algunas de ellas que están relacionadas con la capacidad de un humano de comprender cómo funciona el modelo subyacente o entender las razones por las que el sistema entrega determinado resultado o toma cierta decisión, como así también la habilidad del sistema de proveer evidencia relevante de su funcionamiento o de sus decisiones. En distintas propuestas estas propiedades son agrupadas bajo el paraguas de "Transparencia", que abarca el acceso no sólo al proceso que el sistema de IA realiza en ejecución sino también a todos los procesos que componen el ciclo de vida del sistema desde su concepción y diseño hasta su desarrollo, despliegue y evolución en el tiempo [6].

La compresibilidad o interpretabilidad de un sistema se enfocan en entender el modelo de IA, es decir poder identificar por qué el sistema hace lo que hace cuando lo hace. En el espectro de los sistemas basados en IA existe una amplia variación con respecto a cuán comprensibles e interpretables son sus fundamentos. Analizado desde un sentido estricto, esta propiedad parece requerir que un usuario humano (desarrollador u operario) pueda entender las complejidades del modelo de IA. En este sentido, un sistema de IA basado en reglas[4] , o un árbol de decisión[5] , relativamente pequeño, puede ser altamente comprensible para su diseñador, aunque un usuario lego puede requerir cierto nivel de entrenamiento para poder comprender la "lógica" de la estructura de conocimiento y del proceso de inferencia. Por otro lado, el proceso de aprendizaje en una red neuronal (profunda) comprende de la asignación de millones de pesos a características ("features") que la misma red idéntica de los datos de entrada (por ejemplo, los pixeles de una foto o un video). Si

*4 https://sites.google.com/site/sistemasexpertosunah/home/sistemas-expertos-basados-en-reglas*

*5 https://es.wikipedia.org/wiki/%C3%81rbol_de_decisi%C3%B3n*

bien el diseñador/desarrollador de la misma entiende el meta proceso que está sucediendo (la red está buscando optimizar una función matemática ajustando de manera adecuada esos pesos), es imposible que su mente pueda comprender completamente el cómputo en sí mismo. Para usuario lego del sistema, puede que aun la descripción "intuitiva" de qué significa ajustar una función sea incomprensible. El término "caja negra" se utiliza en general para referir a modelos como este último donde no es posible el acceso a las leyes que gobiernan el procesamiento que sucede dentro del mismo.

¿Cuál es el límite de ignorancia que podemos aceptar en casos donde la vida humana, dependen del funcionamiento del sistema?

Si tomamos el término en un sentido más general, cierto grado de comprensión o interpretabilidad de un modelo de IA puede alcanzarse por medio de una observación extensiva del comportamiento del sistema. La mayoría de los seres humanos usamos un *smart phone* de manera altamente efectiva sin conocer exactamente cómo funciona el dispositivo internamente, pero por medio del uso continuo o por medio de tutoriales (entrenamiento específico) construimos un modelo mental del funcionamiento del mismo en el que podemos confiar y en pocas ocasiones el sistema nos sorprende comportándose de manera completamente impredecible. Ahora bien, podemos preguntar si alcanzar un grado de comprensión empírica basada en la observación del sistema funcionando es o no suficiente para cualquier sistema basado en IA. Claramente es suficiente para manipular un *smart phone* que detecta nuestro rostro para desbloquearse, o que nos indica qué camino tomar hacia una ubicación desconocida, pero ¿qué sucede con un arma inteligente? ¿Cuál es el límite de ignorancia que podemos aceptar en casos donde cuestiones sensibles, como la vida humana, dependen del funcionamiento del sistema?

En relación al principio de previsibilidad, la comprensibilidad es complementaria y ambas son necesarias cuanto más compleja es la tarea que el sistema debe resolver y el entorno donde se desenvuelve la acción. Por un lado, un alto grado de comprensión del sistema incrementa la previsibilidad del mismo. Sin embargo, la previsibilidad no es suficiente en sí misma, aun cuando exista en un grado alto, principalmente para monitorear si el sistema está funcionando bien o en aquellas (quizás escasas) situaciones donde el sistema falla de manera no prevista. Como establecimos en la primera sección, el control humano significativo es primordial en la utilización de armas autónomas letales. Imaginemos el caso en donde dicho sistema, que ha sido desplegado y funciona en conjunto con un operario humano, falla de manera inesperada; es el operario humano quien debe ejercer algún tipo de control sobre el dispositivo para corregir o evitar daños colaterales no deseados. Si el usuario además de haber sido entrenado en el uso del sistema conoce en algún grado cómo funciona la "lógica" del mismo, podría en principio, de manera rápida comprender la situación y efectuar alguna acción contingente. Por ejemplo, la simple realización de que el sistema "no está viendo o identificando" un objeto que el operario si advierte que existe, porque nunca fue alimentado con imágenes similares, le da al operario herramientas útiles para corregir la situación o tomar recaudos necesarios para completar la tarea. La combinación de un alto grado de comprensión y previsibilidad son críticos para la utilización efectiva de un sistema de IA que involucra alto riesgo para el usuario o el entorno donde se despliega el mismo.

Finalmente, abordaremos el principio de explicabilidad. Este concepto es uno de los más desarrollados en relación a sistemas de IA, de hecho, la XIA (IA explicable) es un área de investigación en expansión dentro del campo de la IA. La explicabilidad es la capacidad del sistema basado en IA de poder justificar sus resultados o decisiones en términos que un humano pueda entender. Para entender el concepto es necesario diferenciarlo de comprensión e interpretabilidad. Interpretabilidad no implica explicabilidad y viceversa. Un sistema de caja negra podría, en principio, dar explicaciones sobre sus resultados, pero esas explicaciones serían igual de opacas e incomprensibles que el resto del modelo. Aun así, podría servir para generar confianza en el usuario si las mismas apelan a una correlación asequible por el mismo. Por otro lado, la mayoría de los sistemas basados en IA que hoy utilizamos, aún aquellos que son completamente interpretables no están diseñados para ofrecer explicaciones que acompañan sus resultados o sugerencias. La explicabilidad requiere una habilidad específica del sistema, es decir, es el sistema el que actúa para dar explicaciones, mientras que la interpretabilidad deja al sistema en modo pasivo y es el humano quien lo analiza.

Es claro que si un sistema que puede explicarse a sí mismo puede generar más confianza en el usuario que uno que no ofrezca esta funcionalidad. Sin embargo, existe una ferviente discusión al momento de si la explicabilidad es absolutamente necesaria (o incluso deseable) para cualquier sistema basado en IA. En relación a las armas inteligentes, un sistema explicable puede ser muy útil para la etapa de entrenamiento del operario, ya que le da al mismo la posibilidad de crear un modelo mental más robusto del comportamiento del sistema cuando éste justifica sus acciones. Por otro lado, en ejecución, una vez desplegado, esta capacidad puede hacer que el control humano sobre la misma sea más efectivo y fluido, especialmente cuando el sistema no funciona de la manera esperada y es difícil para el operario comprender si está funcionando bien o está fallando.

Es importante señalar que el concepto de explicación es complejo y determinar en qué consiste una “buena” explicación depende directamente del dominio de aplicación, la tarea específica que el sistema está resolviendo y del (tipo de) usuario u operario [13].

Para concluir esta sección queremos señalar que, en el caso de las armas autónomas, existen diferentes posiciones sobre cuál de estas propiedades son absolutamente necesarias, suficientes o, tan sólo, útiles. Un objetivo de este artículo es dejar en claro que tanto previsibilidad, comprensibilidad y explicabilidad son importantes y un sistema de alto riesgo como lo es un arma inteligente no puede darse el lujo de prescindir de ninguna de ellas. Entre las tres, y pensados desde el diseño del sistema y por el resto de su ciclo de vida, pueden mejorar las chances de que un operario pueda interactuar de manera efectiva con dicho sistema, ejerciendo el control necesario sobre la misma y mitigando los posibles riesgos de fallas o efectos colaterales que puedan producirse. El grado en el cual estas propiedades estén presentes dependerá de cada tipo de sistema que se pretenda desarrollar, sin embargo, deben aún definirse estándares, métricas y herramientas adecuadas para analizar y evaluar los sistemas y marcos regulatorios que garanticen, desde antes de ser desplegados, que los sistemas efectivamente verifican estas propiedades durante todo su ciclo de vida.

# Propuestas

Los autores de este documento estamos convencidos que la comunidad internacional lleva acumulado mucho esfuerzo para acordar criterios para preservar los intereses de la humanidad en diversas área, y muy particularmente en medicina. Consideramos que los criterios éticos y técnicos utilizados en el ámbito de la salud pueden servir de guía para fijar pautas en el contexto del uso de la IA en armamentismo. En particular, en los años recientes el International Medical Device Regulators Forum (IMDRF) ha fijado importantes definiciones para la regulación del uso de IA en dispositivos médicos. Algo que surge rápidamente de la lectura de la documentación existente, es que los modelos de regulación están todavía en etapa de desarrollo[6]. Esto se debe esencialmente a que la comunidad científica todavía no ha dado respuesta técnica-formal respecto a cómo mensurar previsibilidad y confiabilidad o cómo lograr interpretar o explicar el comportamiento emergente en un sistema de IA generado con ejemplos. Es importante tener claro que esta es una tecnología en una etapa inicial de progreso y de la cual estamos haciendo conjeturas sobre sus efectos basado en especulaciones, no en evidencia. En ese sentido, hay que ser cautos con nuestras proyecciones sin dejar de poner límites ético-sociales al impacto de estas tecnologías. El problema al que nos enfrentamos es que para poner límites hay que tener métricas. Por ejemplo, para que una droga pueda ser comercializada existen protocolos que conllevan años de experimentación para verificar su seguridad y efectos. Pero estamos lejos de algo parecido para sistemas de IA. Por ejemplo, hablamos de sesgo, donde tenemos ejemplos que denunciamos, pero no contamos con una especificación precisa del concepto, ni una forma de medirla y mucho menos una manera de mitigarlo. Lo mismo se reproduce para varios de los conceptos que hemos abordado en las secciones anteriores. Todo, mientras nuevas arquitecturas de ML (con nuevos desafíos) siguen proponiéndose.

Lo antedicho busca abonar la idea de que el control y regulación de las LAWS debería pensarse como un proceso continuo en el tiempo que debería generar límites preventivos que vayan evolucionando y ajustándose con los avances certeros de la tecnología. Un ejemplo claro de esto es el de las herramientas de reconocimiento facial. Su uso abusivo en prevención del crimen ha promovido acciones precautorias contra su utilización. Así, empresas, como Amazon, Google y Microsoft, deciden la suspensión de ventas de sistemas que utilicen esta tecnología. Por otro lado, desde el estado se establecen moratorias o prohibiciones transitorias con vista a alcanzar regulaciones más específicas. Esto no quiere decir que la tecnología de reconocimiento facial no se utilizará nunca más, sino que se condiciona su utilización a acuerdos sociales y evidencia científica que hagan seguro su uso.

Siguiendo esta dinámica es que vamos a dividir nuestras propuestas a dos niveles: uno más general (y político), y otro más particular (y técnico).

*Generales:*

- La instauración de un código ciber-ético de buenas prácticas profesionales que los diseñadores

---

*6 Ver por ejemplo https://algorithmwatch.org/en/story/medical-devices/*

y desarrolladores de LAWS deberían adherir de la misma manera que firman un compromiso de confidencialidad. También podría servir para instaurar un juramento hipocrático del informático. Esos valores deberían ser incorporados a las currículas de las carreras tanto de formación militar como los profesionales de la informática.

- La conformación de una Agencia de Regulación y Control Internacional de LAWS. Esta agencia estaría dedicada a generar estándares y métricas de validación, realizar auditorías y monitoreo, y analizar situaciones de fracaso de las normativas para retroalimentar el sistema de regulación y control.

- Propiciar la instauración de zonas libres de LAWS. Siguiendo la idea actual de zonas libres de armas nucleares.

- Propiciar la conformación de un registro internacional de fabricantes de LAWS y prohibir la fabricación de LAWS sin una licencia (ver [12]).

- Propiciar que las empresas que diseñan y producen LAWS tengan un comité interno de ética cuyos miembros tengan estabilidad y libertad para ejercer la tarea. Eso resolvería en tema de confidencialidad empresarial.

*Particulares:*

- Creemos que debe cambiarse el foco de regulación y hablar de: "Software como dispositivo militar" (MD por sus siglas en inglés) como toda pieza de software destinado a ser utilizado para uno o más fines militares que llevan a cabo estos fines sin ser parte de un dispositivo de hardware. Sobre esa base es posible extender el concepto a MD basados en Inteligencia Artificial y Machine Learning. A partir de allí es posible fijar regulaciones que den lugar a funcionalidades seguras para los civiles y que respeten el IHL.

- Para lograr los objetivos de confiabilidad y verificabilidad de las LAWS es necesario dotarlas de "cajas negras" (como las usadas en los aviones) que permitan evaluar sus acciones a posteriori.

- Desarrollo de una segmentación de los distintos tipos de LAWS que permita establecer legislación específica para cada uno. El primer intento llevado a cabo en [7], nos parece interesante, más aún si se salva la cuestión de "atar" la división en base a la existencia de sensores. Como se mencionó más arriba, consideramos que dicha segmentación tiene que estar basada en aspectos multi-intencionales, que incluyan tanto características y capacidades del dispositivo, como también configuraciones del entorno de desempeño y de las tareas objetivos. En [14] aparecen varios criterios de segmentación que nos parecen un buen punto de partida.

- Impulsar investigaciones tecnológicas que conduzcan a mejorar la evaluación de LAWS en al menos los tres ejes analizados.

# Referencias


- [1] Learnability can be undecidable. Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka and Amir Yehudayoff. Nature Machine Intelligence, VOL 1, pag. 44-48. JANUARY 2019.
- [2] El elemento humano en las decisiones sobre el uso de la fuerza. Infografía de INIDIR. Merel Ekelhof y Giacomo Persi Paoli.
- [3] Mission Command and Armed Robotic Systems Command and Control A Human and Machine Assessment. Robert J. Bunker. Land Warfare Paper 132 / May 2020. The Association of the United States Army.
- [4] Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control. Vincent Boulanin, Meil Davison, Netta Goussac y Moa Peldán Carlsson. Jun 2020. SIPRI.
- [5] The Black Box, Unlocked: Predictability and Understandability in Military AI. Holland Michel, Arthur. 2020. Geneva, Switzerland: United Nations Institute for Disarmament Research. doi: 10.37559/SecTec/20/AI1
- [6] Anteproyecto de Recomendación sobre la ética de la Inteligencia Artificial (UNESCO 2020): https://es.unesco.org/artificial-intelligence/ethics
- [7] Regulación de la Autonomía de los Sistemas de Armamentos. https://article36.org/wp-content/uploads/2020/10/regulacion-autonomia-ES.pdf
- [8] Artificial intelligence is improving the detection of lung cáncer. Elizabeth Svoboda Nature. Noviembre 2020. https://www.nature.com/articles/d41586-020-03157-9
- [9] Interpretable machine learning. A Guide for Making Black Box Models Explainable. Molnar, Christoph. 2019. https://christophm.github.io/interpretable-ml-book/.
- [10] Artificial intelligence: a modern approach. 3rd ed. Russell, S. J., Norvig, P., & Davis, E. Upper Saddle River, NJ: Prentice Hall. 2010.
- [11] Robust Physical-World Attacks on Deep Learning Visual Classification. K. Eykholt et al., 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp. 1625-1634. 2018.
- [12] Model Law against the Illicit Manufacturing of and Trafficking in Firearms, Their Parts and Components and Ammunition. Documento Naciones Unidas. 2011. https://www.unodc.org/documents/legal-tools/Model_Law_Firearms_Final.pdf
- [13] Explanation in artificial intelligence: Insights from the social sciences. Miller, T. Artificial Intelligence (2019).
- [14] A choices framework for the responsible use of AI. Benjamins, R. AI Ethics 1, 49–53 (2021). https://doi.org/10.1007/s43681-020-00012-5

**Maria Vanina Martinez** es Doctora en Ciencias de la Computación (Universidad de Maryland, USA) con un posdoctorado en la Universidad de Oxford. Es investigadora del Instituto de Ciencias de la Computación (CONICET- UBA) en el área de Inteligencia Artificial y profesora del Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, donde dicta la materia "Ética & IA". Además es Miembro del Comité Nacional de Ética en la Ciencia y Tecnología del Ministerio de Ciencia, Tecnología e Innovación Productiva. Es miembro de la campaña Stop Killer Robots y del Shelac.

**Ricardo Oscar Rodriguez** es Doctor en Ciencias de la Computación con especialización en Inteligencia Artificial. Es Profesor Asociado en el Departamento de Computación, FCEyN-UBA y miembro del Instituto de Ciencias de la Computación (UBA-CONICET). Sus trabajos científicos se inscriben en el desarrollo de modelos lógicos para el razonamiento bajo incompletitud e incertidumbre. Actualmente dicta la materia "Ética & IA" y es miembro del Shelac y de la campaña Stop Killer Robots de la cual participa activamente. Ha sido co-chair & financial chair de IJCAI2015 en Buenos Aires.


APP
Asociación para Políticas Públicas


CAMPAIGN TO STOP
KILLER ROBOTS


SEHLAC
SEGURIDAD HUMANA
EN LATINOAMÉRICA Y EL CARIBE