LA I.A., LA NUEVA CARRERA ARMAMENTISTA Y EL ROL DE LOS CIENTÍFICOS

María Vanina Martínez y Ricardo Oscar Rodríguez
Departamento de Computación. FCEyN-UBA
Instituto de Ciencias de la Computación. UBA-CONICET

Resumen
El desarrollo de nuevas tecnologías inteligentes para uso militar inaugura una nueva etapa en la carrera armamentista que puede significar una sofisticada pérdida de libertades para la ciudadanía global. Gobiernos, organismos internacionales, sociedad civil y la comunidad científica enfrentan el desafío de promover parámetros legales, éticos y morales para el desarrollo de I.A con fines militares. Este artículo explora conceptos y definiciones sobre el desarrollo de armas autónomas y reflexiona sobre la importancia de establecer cuerpos legales y culturales que guíen el desarrollo de la I.A.

No hay duda que el desarrollo de la Inteligencia Artificial (I.A.) tendrá un fuerte impacto cultural y social en la vida cotidiana de la humanidad. De hecho, ya lo tiene y existen muchas aplicaciones basadas en IA, que facilitan o mejoran la vida de las personas. Hay ejemplos tan simples como los sistemas anti-spam o los asistentes digitales que nos hablan y asesoran con total naturalidad. Otros, son un tanto más sofisticados, como el sistema de monitoreo remoto que permite predecir incendios forestales, o los sistemas de diagnóstico de cáncer o predicción de cegueras, o asistentes semiautomáticos de aterrizajes en grandes aeropuertos y hasta autos sin conductor. Y todos ellos son ejemplos nimios frente los grandes augurios que se vaticinan para esta tecnología. Pero, como con cualquier otra tecnología disruptiva, no todo es beneficio y prosperidad. Su uso indebido y falto de ética también existe. Clara muestra son los sistemas que predicen las preferencias de diferentes grupos sociales y tratan de influir/orientar/manipular sus opiniones y acciones. Un caso reciente, ha sido el escándalo de la "fuga" masiva de datos de Facebook a favor de Cambridge Analytica, y el uso de estos para definir campañas de atracción de votos para el referéndum del Brexit o la candidatura de Trump en la elección presidencial del 2016 en USA. Pero existen muchas formas de control de opinión más sutiles y veladas a los que la sociedad está diariamente expuesta, como lo son los sistemas de recomendaciones orientados por técnicas de I.A. Nuestras amistades en las redes sociales, nuestros consumos o búsquedas por internet, determinan un perfil que es reconocido por

estos sistemas y nos vuelve vulnerables. Sin embargo, es importante dejar claro que más allá del objetivo de manipulación con el cual pueden usarse estas tecnologías, esos mismos perfiles pueden servir para detectar pedófilos, depresivos con intenciones suicidas, identificar hechos de discriminación o bullying en redes sociales, etc.

Pero si los métodos sutiles de manipulación/persuasión/dominación social mencionados anteriormente, no fueran suficiente para la voracidad corporativa, ella siempre podrá recurrir al ancestral método de la violencia. Y nuevamente, aquí la I.A. también tendrá un fuerte impacto a través del desarrollo de una tecnología militar inteligente. Tanto es así, que las fuerzas armadas de los países desarrollados han iniciado una nueva carrera armamentista con el fin de mantener su poderío militar. Para los expertos estaríamos entrando en la tercera era de la tecnología armamentista después de las irrupciones de las armas de fuego y las bombas nucleares. Para los simples mortales estaríamos entrando en una sofisticada etapa de pérdida de libertades.

La carrera armamentística adquiere una nueva dimensión con las nuevas tecnologías inteligentes. La ciudadanía enfrenta una sofisticada etapa de pérdida de libertades

Para visualizar esa vulnerabilidad basta considerar el potencial destructivo que supone la utilización de técnicas de I.A., como los avances en el área de la robótica, en el desarrollo armamentista. Aún la más mínima posibilidad de dotar a la más simple arma con capacidades muy básicas de autonomía tales como movilidad, de percepción y comprensión del entorno (cómo las que permiten hoy en día implementar las técnicas basadas en aprendizaje automático y redes neuronales), rápidamente despierta nuestra inquietud en relación a cómo y con qué propósito pueden ser utilizados. Si extendemos el análisis con la posibilidad de incorporar capacidades de razonamiento y toma de decisiones autónomas, el escenario se vuelve aún más complejo y preocupante. Más aún, imagine uno de los modernos autos sin conductor cargado de explosivos y quedará claro de que estamos hablando.

Por todo eso, hacer tomar conciencia a la sociedad sobre la vulnerabilidad a la que nos exponen las nuevas tecnologías de I.A., es una responsabilidad que tenemos, especialmente, los científicos que desarrollamos y perfeccionamos estas técnicas. Esencialmente para no repetir el error cometido con el uso de la energía atómica. Justamente por esa razón, la comunidad científica de I.A. viene bregando desde hace varios años por el no desarrollo de armar letales autónomas. Durante la conferencia

IJCAI2015 en Buenos Aires, miles de investigadores en I.A. de la comunidad internacional hicieron un llamamiento a la no proliferación de armas letales autónomas a través de una carta abierta que tuvo repercusión a nivel mundial (ver https://futureoflife.org/open-letter-autonomous-weapons/). En línea con esta puesta en escena del problema, durante el desarrollo del IJCAI2017 se hizo pública una carta abierta dirigida las Naciones Unidas firmada por los presidentes de 116 empresas líderes en el uso de I.A. (de 26 países) alertando nuevamente sobre el peligro de las armas con I.A. y llamando a su prohibición. Un interesante efecto de estas campañas de concientización ha sido la negativa de empleados de Google en participar del llamado *Proyecto Marven* generado a partir de un contrato entre la compañía y el Pentágono para el desarrollo de "drones asesinos". De hecho, más de tres mil empleados de Google solicitaron no sólo que se abandone el proyecto, sino que se redacte, publique y aplique una política clara que establezca que ni la compañía ni sus contratistas participarán nunca en el desarrollo de tecnología de guerra.

En esa misma dirección, Stuart J. Russell, profesor de la Universidad de California en Berkeley, y el Future of Life Institute, crearon un video que da cuenta de la capacidad destructiva que tienen ese tipo de drones (ver en https://www.youtube.com/watch?v=9Pn17-Mr7wc&t=17s). Ese video fue expuesto durante un evento en la Asamblea de las Naciones Unidas sobre Armas Convencionales para hacer un llamado a realizar acciones que prohíban el desarrollo de este tipo de armamento.

Pero proponer guías y promulgar una legislación sobre el uso de las técnicas de I.A. en la construcción de armas no es la única solución a este problema que enfrenta la sociedad. A diferencia de otras tecnologías que los seres humanos hemos desarrollado a lo largo de nuestra historia, los sistemas de I.A. pueden ser construidos en base a códigos morales y/o éticos de manera que su comportamiento pueda ser informado y/o restringido de acuerdo a los valores que esos códigos definen, de la misma manera que sucede con el comportamiento humano. Como científicos también es parte de nuestra responsabilidad sentar las bases tanto para la discusión y el avance en torno de áreas de investigación que permitan entender e incorporar comportamiento ético en los sistemas de I.A. desde su desarrollo, como así también bregar por políticas de construcción y desarrollo de sistemas que aseguren que éstos van a comportarse de acuerdo a ciertos códigos y/o valores y que no podrán ser fácilmente manipulados por otros agentes para que falten a esos principios.

Es por todo esto que las actividades que se han estado desarrollando no se limitan tan sólo a alertar sobre los peligros del uso de la I.A. con fines bélicos, sino que también se ha

venido trabajando en diferentes foros y organizaciones internacionales para generar consenso acerca de políticas y leyes, no sólo contra la proliferación de armamento basado en técnica de I.A. sino también para establecer un uso adecuado de dicha tecnología en todos los ámbitos. En este sentido, el principal rol de los científicos ha sido, y es, clarificar los alcances de las nuevas tecnologías desarrolladas a partir de la I.A. para desarmar tanto argumentos minimalistas como fatalistas en que los debates políticos suelen caer. Sentar las bases para un debate constructivo que permita generar herramientas de control precisas que no impida el avance de la I.A.

Foto de la Campaña Internacional #stoplillerrobots (www.stopkillerrobots.org)

Al respecto parece importante clarificar y/o fijar posición sobre algunos aspectos o conceptos que aparecen reiteradamente en distintos documentos oficiales (a veces sin un unificado significado):

1) Sistemas de Armas Autónomas Letales, SAAL, (Lethal Autonomous Weapons Systems). Esencialmente se refiere a cualquier sistema bélico que puede perpetrar un ataque letal contra seres humanos de forma autónoma, es decir, que pueda planificar el modo de abordar su objetivo y tomar la decisión de asesinar sin

supervisión humana. En términos generales se refiere cualquier dispositivo diseñado para matar personas ejecutando la acción con criterio propio. Esto no incluye misiles teledirigidos o drones pilotados a distancia para los cuales los humanos toman todas las decisiones de ataque.

2) Derecho Internacional Humanitario (DIH). Conjunto de normas que buscan limitar los efectos de los conflictos armados. Descansan sobre la base de cinco principios: Humanidad (prioridad al respeto de la persona por sobre las necesidades militares), Necesidad Militar (prohibición de realizar acciones militares innecesarias), Distinción (determina la necesidad de diferenciar en todo momento entre civiles y combatientes, así como entre bienes civiles y objetivos militares), Limitación (prohibición de ciertos métodos y armas de combate tales como químicas, bacteriológicas, nucleares e incendiarias y minas antipersona), y Proporcionalidad (reglas para valorar como lícitos o ilícitos los daños causados a personas y bienes que no participan en las hostilidades por un ataque dirigido contra un objetivo militar).
3) La "autonomía" es un término que no debe ser entendido de manera unidimensional. Por el contrario, es un concepto que se deriva de dos palabras griegas ("auto" –self– y "nomo" –governance-) y que posee dos sentidos propios: por un lado "autosuficiencia" (self-sufficiency), referido a la capacidad de cuidarse a sí mismo, o lo que es igual, a la condición o estado de quien se basta a sí mismo. Por otro está "autodirección" (self-directedness), entendido como el atributo de estar libre de todo control externo. Un rasgo esencial de la autonomía es el autoaprendizaje que le permite adaptarse a los cambios y aumentar su impredecibilidad. En la actualidad la autonomía total no ha sido implementada en ningún sistema artificial. Existen sí dispositivos con autonomías parciales como los automóviles sin conductor o los drones sin piloto. En los casos conocidos, la autonomía es específica a una tarea. Un SAAL requiere tener autosuficiencia en abordar múltiples tareas combinando y coordinando las acciones necesarias para logra un objetivo. La autonomía es una característica ortogonal o aditiva de un sistema armamentístico. Es decir, todo tipo de arma convencional es potencialmente pasible de alcanzar total autonomía con las tecnologías de I.A.

   Para ser más precisos un SAAL sería un tipo de arma que puede seleccionar (es decir, buscar, detectar, identificar y localizar) y atacar (usar la fuerza en contra, neutralizar, dañar o destruir) objetivos sin intervención humana. Este tipo de armas tendría la habilidad de aprender y/o adaptar su funcionamiento en respuesta a las circunstancias cambiantes del entorno en el que se despliegan, por

lo que su uso podría reflejar un cambio cualitativo de los paradigmas en la conducción de las hostilidades.

A pesar de lo expresado en los párrafos precedentes, debe dejarse bien en claro que es muy difícil definir con precisión el concepto de autonomía. La pretensión de hacerlo como paso previo de estipular la prohibición de sistemas que busquen alcanzarla, aunque parezca sensato, conduce a la inacción.

4) Tratados internacionales. Existen tratados dedicados a tipos específicos de armas para municiones en racimo, minas antipersonales, láseres cegadores, armas químicas y armas biológicas, y por supuesto armamento nuclear.

La comunidad científica tiene la responsabilidad de sentar las bases para la discusión y el avance en la investigación de los sistemas de I.A de acuerdo a parámetros éticos y valores que no puedan ser fácilmente manipulados.

5) Inteligencia Artificial. Sistemas informáticos que llevan a cabo tareas que usualmente realizan los humanos haciendo uso de inteligencia. Dichas tareas esencialmente involucran razonamiento y aprendizaje. Lo cual implican otras habilidades cognitivas como representar información/conocimiento, reconocer imágenes y sonidos, etc. En la actualidad los sistemas de I.A. tienen muchas limitaciones.

6) Ética en un sistema informático. Dotar al sistema con la capacidad para distinguir/discriminar lo que es correcto/bueno de los que no lo es, ya sea en un sentido cultural, social o legal.

Planteadas estas cuestiones definicionales y conceptuales, cabría hoy preguntarse: ¿hasta qué punto se pueden medir los riesgos de algo que aún no se sabe si podrá crearse?; ¿será posible que algún día existan niveles de "inteligencia artificial" tan sofisticados que generen sistemas armamentistas completamente autónomos?; y sin llegar muy lejos, ¿cómo podría un humano programar a un sistema autónomo para que logre diferenciar a un civil de un combatiente?; ¿qué fórmulas se pueden aplicar para programar en el arma un estándar de proporcionalidad en el uso de la fuerza letal en zonas de guerra que por definición son bastante imprevisibles?; si existiere un error de diseño y de programación ¿quién debería rendir cuentas acerca del daño que produzca ese dispositivo bélico?, ¿el comandante de la misión, el operador humano que la activó, el programador?; más aún, si estos sistemas son completamente autónomos ¿podrían desobedecer órdenes? O en su defecto ¿podemos dotar a estos sistemas de bloqueadores éticos que les impida cometer fechorías?; pero sobre todo ¿en qué situación se encontraría la dignidad humana de una

persona que llegue a sentir terror, pánico, desolación o impotencia al verse afectada por un daño cometido en su contra por una máquina y producto de un error técnico?
Todas estas cuestiones válidas resaltan el hecho que la discusión principal del uso apropiado de una tecnología potencialmente tan poderosa debe ser abordado en lo inmediato. En los tiempos modernos, generalmente el desarrollo tecnológico avanza en forma mucho más acelerada que el análisis de las controversias éticas que produce. Pero en este caso particular creemos que el abordaje de las implicaciones éticas del uso de I.A. debe realizarse a priori o al menos en paralelo. Quede claro que no estamos propiciando volver a prácticas medievales donde la sociedad y el estado se arrogaban el derecho de impedir el desarrollo de la astronomía. En cambio, bregamos por el establecimiento de cuerpos legales y culturales que guíen el desarrollo de la I.A.

En tal sentido parece necesario promover las siguientes acciones:

1) La promulgación de códigos éticos para la aplicación de técnicas de I.A. en cualquier dispositivo. Según estos principios, los sistemas inteligentes no podrían cometer ilícitos, ni atentar contra la integridad física o psicológica de las personas.
2) Definir criterios de legislación vinculante y no vinculante (derecho duro y blando) en el uso de la I.A.
3) La prohibición del desarrollo de los SAALs antes que las mismas sean tecnológicamente factibles de ser construidas. En la actualidad los conocimientos científicos y tecnológicos no han alcanzado la envergadura y desarrollo suficiente para la construcción efectiva de este tipo de armamentos, pero se estima que los mismos serán alcanzados en las próximas décadas. El establecimiento anticipado de normativas internacionales que prohíban su desarrollo busca evitar las condiciones especiales que se dieron en el Tratado de No-proliferación de Armas Nucleares donde cinco países se arrogaron el derecho de poseer armas nucleares por el simple hecho de haber realizado ensayos previos a la firma del mismo.
4) Promover más proyectos de investigación interdisciplinarios para estudiar los efectos socio-culturales producto del desarrollo de sistemas de I.A. y su implicancia ética y moral en la sociedad.
5) Incentivar la conformación de foros multidisciplinarios (que incluyan, entre otros, psicólogos, sociólogos, politólogos, filósofos, computadores científicos, economistas, legisladores, etc.) para discutir y brindar orientación sobre temas emergentes relacionados con el impacto de la I.A. en la sociedad.
6) La enseñanza de aspectos éticos en las carreras de formación de profesionales que desarrollarán estas nuevas tecnologías.

7) Impulsar un tratado internacional de no-desarrollo de SAALs (A Treaty of Non-Development of LethalAutonomousWeaponsSystems), que debería basarse en al menos dos pilares: el no-desarrollo y el uso de la Inteligencia Artificial sólo para fines pacíficos. Para lo cual debería conformarse una Agencia Internacional de Inteligencia Artificial (en el marco de la ONU) que sea el ente de control de aplicación de las normas que surjan del tratado.
8) Desarrollar un protocolo de impacto social que permita evaluar la pertinencia o no de lanzar al mercado un nuevo producto que utilice técnicas de I.A.

Algunas de estos aspectos ya están siendo resueltos aisladamente por empresas como Google, cuyos empleados han promovido siete principios que los sistemas de I.A. deberían seguir: 1) Ser socialmente beneficiosos; 2) Evitar crear o reforzar sesgos injustos; 3) Construir sistemas socialmente seguros; 4) Ser responsable ante las personas; 5) Incorporar principios de diseño de privacidad; 6) Mantener altos estándares de excelencia científica; 7) Estar disponible para usos que estén de acuerdo con estos principios.(detalles en https://www.blog.google/technology/ai/ai-principles/).
Aún así, consideramos que las acciones no pueden limitarse a la iniciativa privada o unilateral de un grupo, más allá de su buena voluntad. Por el contrario, pensamos que debe ser la sociedad toda la que se comprometa en este debate.

Para cerrar esta presentación nos gustaría resaltar que los argumentos y conflictos éticos esgrimidos hasta ahora en relación con la construcción y uso de armas inteligentes demuestran claramente la importancia de la discusión y la necesidad de comenzar desde ya en la definición de planes de acción que ayuden a mitigar los problemas relacionados. Sin embargo, algunos científicos y otros actores interesados en el área, apuntan a una problemática que pareciera en principio ser más urgente que la discusión a largo plazo sobre los robots asesinos, y son las implicancias de la utilización de sistemas de I.A. llamados de "caja negra" para la toma de decisiones en situaciones cotidianas o civiles. Los algoritmos basados en técnicas de aprendizaje automático infieren patrones estadísticamente relevantes a partir del análisis de una gran cantidad de datos. Existen dos problemas relacionados con estos algoritmos, uno es que, si son alimentados con datos sesgados, los resultados y por ende las decisiones que se tomen en base a ellos pueden estar sesgadas. Esto es un problema porque en muchos casos estos algoritmos son usados por personas sin el entrenamiento necesario para proporcionar datos de entrada libre de sesgos. El segundo problema es la opacidad de estos algoritmos en términos de su funcionamiento. Algunas de las técnicas usadas son realmente complejas lo que dificulta poder entender y auditar los sistemas que los usan.

Para dejar clara la controversia, propondremos un par de ejemplos (pero hay muchísimos más):

1) En 1991, la Dra. Diane F. Halpern (de la Universidad Estatal de California, San Bernardino) y el Dr. Stanley Coren (de la Universidad de Columbia Británica) publicaron un trabajo cuya conclusión era alarmante. Estos investigadores tomaron una muestra de 987 individuos que habían muerto y les preguntaron a sus familiares más cercanos, si eran zurdos o diestros. El hallazgo fue muy perturbador, los zurdos se morían nueve años antes que los diestros. El trabajo fue publicado en el New EnglandJournal of Medicine, que es una de las revistas médicas más prestigiosas del mundo. Si las conclusiones eran correctas, esto significaba que ser zurdo era tan malo como fumar 120 cigarrillos por día. Posteriormente se comprobó que la muestra que se había tomado era sesgada y no se correspondían con la realidad. Sin embargo, suponga que una compañía de seguro utiliza esa información para fijar la prima de sus pólizas. Eso implicaría que los seguros de vida para zurdos serían significativamente más caros.
2) De todas maneras, la cuestión del sesgo también es humana. El sistema de concesión de la libertad condicional para los presos en los Estados Unidos proporciona un ejemplo sorprendente. "Se ha demostrado que es mucho más probable que a los convictos se les conceda libertad condicional si comparecen ante el juez inmediatamente después del almuerzo en lugar de justo antes", (ver https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3084045/). Los algoritmos, que, por supuesto son inmunes al síndrome del estómago vacío, son fáciles de programar porque en los EE. UU. La libertad condicional depende esencialmente de un sólo parámetro: el riesgo de huir o reincidir en base a los antecedentes. Cabría preguntarse si entonces es posible una verdaderamente 'justicia ciega' basada completamente en hechos objetivos. La respuesta no parece ser contundente. Hoy día los juzgados de los Estados Unidos utilizan una “caja negra” para la toma de estas decisiones. Dicho sistema está basado en datos históricos y reproducen los sesgos de los datos con los que son entrenados.

Recientemente, varios científicos prominentes en el área de I.A. han declarado su postura frente a estos problemas y apuntan a cuán difícil es poder identificar comportamientos sesgados en los sistemas más comúnmente usados. Los algoritmos actuales simplemente no están diseñados para explicar las decisiones que toman, lo cual hace imposible que un usuario de una aplicación pueda entender, mucho menos cuestionar, cómo son usadas sus preferencias o por qué la información se le presenta en un determinado orden. Y la solución a esto no se alcanza con solicitarles a los proveedores del servicio que publiquen

los detalles de los datos con los que fueron entrenados los algoritmos o los algoritmos en sí mismos. Muchas de estas herramientas son demasiado complejas y requieren el control de muchos parámetros, como para poder ser examinadas meticulosamente en pos de entender su funcionamiento. Por lo cual, se están explorando modelos alternativos y complementarios que permitan, por ejemplo, que el sistema en sí mismo ofrezca justificaciones sobre las determinaciones que asume durante los procesos de toma de decisión.

Las técnicas modernas de aprendizaje automático han permitido que los sistemas de I.A. salieran de los laboratorios y superaran aplicaciones de "juguetes" para estar inmersos en el mundo real, permitiendo que el entorno pueda ser censado y comprendido. Los problemas de opacidad y potencial sesgo provenientes del manejo de los datos son dos de las limitaciones de los sistemas actuales de I.A. más urgentes que deben ser atendidos antes que comprometan aún más la privacidad y otras cuestiones éticas de nuestra sociedad.

Sobre los Autores

María Vanina Martínez es Dra. en Ciencias de la Computación, investigadora adjunta del CONICET, trabajando en el Instituto de Ciencias de la Computación (CONICET - UBA) y profesora en el Departamento de Computación de la UBA. Sus intereses de investigación se enfocan en el desarrollo de modelos para la representación de conocimiento en sistemas de apoyo a toma de decisiones basados en Inteligencia artificial.

Ricardo Oscar Rodríguez es Dr. 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.
Ha sido co-chair y finacial chair de IJCAI2015.

Sobre los organizadores

SEHLAC es una red de trabajo que se surge en el año 2008 impulsada por **Action on Armed Violence** con el fin de continuar la exitosa dinámica y trabajo logrado durante el **Proceso de Oslo sobre Municiones en Racimo** y para aprovechar la sinergia entre sus miembros y ampliar el trabajo hacia otros temas de carácter humanitario que afectan gravemente a Latinoamérica y el Caribe. Esta red está conformada por personas individuales y representantes de organizaciones de la sociedad civil. El grupo ofrece un espacio para

compartir ideas y debatir respecto al tema de la violencia armada y desarrollo; y para coordinar acciones y estrategias comunes.

Campaña para detener los robots asesinos - #Stopkillerrobots es una coalición internacional de ONG´s que trabaja preventivamente en la prohibición del desarrollo, producción y uso de armas totalmente autónomas y para lograr una protección humanitaria y control legal efectivo sobre los Robots Asesinos.