**El problema de la responsabilidad en la IA**

**IECISA**



Este documento contiene información y material confidencial propiedad de Informática El Corte Inglés S.A. Los materiales, ideas y conceptos contenidos en esta propuesta serán utilizados exclusivamente para evaluar las capacidades de Informática El Corte Inglés y no deberán ser divulgados fuera de su organización o utilizados con propósitos distintos a los mencionados. No está permitido su reproducción total o parcial ni su uso con otras organizaciones para ningún otro propósito, excepto autorización previa por escrito.

La incorporación de la tecnología a nuestro entorno personal y productivo ha suscitado, inevitablemente, un encendido debate en su aplicación práctica. La Inteligencia Artificial, como tecnología aplicada no podía estar exenta de dicho debate, más aún, en la medida que se propone simular el comportamiento humano, incluyendo los procesos de toma de decisiones.

En Enero de 2.017 el *Future of Life Institute* (LFI) organizó la conferencia de Asimilar entre los que se encontraban alguno de los principales referentes tanto a nivel de investigación como empresarial del mundo.

Fruto de esta conferencia se desarrollaron los 23 principios de Asimilar que representan el compromiso voluntario por parte de los firmantes para “la investigación, desarrollo y aplicación de la Inteligencia Artificial”[[1]](#footnote-1)

En estos principios se aborda la problemática inducida de la Inteligencia Artificial desde varias ópticas.

* La necesidad de la cooperación y la transparencia en todas las fases del desarrollo tecnológico.
* La necesidad de gestionar la trazabilidad y seguridad de los mecanismos de decisión de la IA.
* Preservar todos los valores humanos (libertad, dignidad, derecho a decidir)
* Control del desarrollo, incluyendo el auto-inducido por la Inteligencia Artificial

Si bien todas ellas requieren una atención singular vamos a desarrollar uno de los puntos que menos atención ha suscitado en este debate: el problema de la responsabilidad.

El problema de la responsabilidad.

Hasta ahora la tecnología ha sido predominantemente “***actuadora***”, es decir, en determinadas condiciones y, generalmente con el ***concurso y dirección*** de las ***personas***, realizaba tareas de tal forma que se multiplicaba la productividad natural de una persona a una fracción del coste del mismo.

Normalmente, las consecuencias económicas de la incorporación de la tecnología han supuesto un desplazamiento de la masa productiva del modelo anterior al nuevo modelo, la destrucción de puestos de trabajo fue compensada por nuevos servicios o productos que absorbieron dicho impacto. Esto ha ocurrido con todas las revoluciones industriales: agrario->fábrica->servicios y, esperamos, lo haga en el futuro sobre un nuevo paradigma (¿economía verde, *health economy*?)

La inteligencia artificial está cambiando drásticamente este panorama en la medida que un sistema es/será capaz de ***tomar decisiones*** y, si atendemos a nuestra memoria histórica, hacerlo con el ***mismo grado de eficacia*** de una persona, pero de ***forma más eficiente***. Es decir, somos/seremos capaces de tener ***sistemas que “deciden”*** y ***sistemas que “actúen”*** de forma totalmente ***integrada***, ***minimizando el factor humano*** y aportando mejoras significativas a la productividad de la economía en general.

Hasta ahora, los sistemas que eran capaces de decidir y actuar de forma independiente atendían a una lógica de la que conocíamos todos sus detalles y de la que podíamos trazar los mecanismos y, en esencia, entender las motivaciones. Dicho de otra forma, la ***delegación que se hacía de la autoridad*** era un ***acto respaldable*** y donde la ***responsabilidad*** quedaba perfectamente establecida.

La IA, y en particular, la denominada IA fuerte[[2]](#footnote-2) se postula como un mecanismo donde esta responsabilidad puede llegar a diluirse. Consideremos este escenario.

* El fabricante provee de un sistema de inteligencia artificial.
* La organización entrena a dicho sistema.
* El sistema se despliega en un contexto que es cambiante y para el que nuestra aplicación funciona como una caja negra que, ante un estímulo puede dar un abanico de soluciones con un determinado grado de certeza.

Consideremos un caso de uso clásico. Hemos construido un sistema de priorización de mensajes el cual, a la recepción de un e-mail, procede a su clasificación en diferentes áreas temáticas, previamente identificados y lo prioriza.

Para resolver este problema utilizamos dos algoritmos que aplican en cascada:

* Contamos con un sistema de tratamiento de lenguaje natural basado en un grafo de conocimiento donde podemos explicitar las reglas de contexto y desambiguación lo que nos permitirá saber de qué se está hablando y establecer la temática basándonos en reglas.
* La priorización, sin embargo, se basa en el conocimiento de las personas que reciben estos mensajes para lo cual se ha utilizado una red neuronal para su modelización. Esta red recibe la temática, una serie de palabras clave, origen, destinatario y la prioridad que tuvo en su momento para el entrenamiento.

El primero es trazable, las reglas a utilizar son claras y, en una buena práxis IA, queda claramente identificada la responsabilidad. El segundo, por su propia naturaleza, es una caja negra en la que la configuración interna de la red neuronal expresa la inteligencia del sistema, pero no tenemos una teoría consistente que explique su funcionamiento y mucho menos exprese la trazabilidad.

¿En este contexto y ante un fallo producido donde queda la responsabilidad?

¿Puedo sustituir el algoritmo “caja negra” por otra solución y seguir siendo igual de eficaz o competitivo?

La respuesta, como en todo aquello que se relaciona con la tecnología, ***no puede quedar exclusivamente al arbitrio de las leyes de mercado***. El ***regulador*** debe desarrollar aquellos ***aspectos legislativos*** que den amparo y delimiten claramente las ***esferas de imputabilidad*** de cada uno de los actores.

Esto, en nuestro contexto de ***aldea global***, con ***múltiples sistemas legislativos***, es en sí, un proyecto titánico pero imprescindible sino queremos que los esfuerzos que se están haciendo de ***autorregulación*** queden en papel mojado o nos lleven a ***soluciones no-competitivas***.

***Europa***, retrasada en los aspectos tecnológicos, es, sin embargo, la única entidad ***supra***-***nacional*** que puede llevar a un marco regulatorio consistente y de amplia aplicabilidad que pueda garantizar, a largo plazo, un uso “humano” de la Inteligencia Artificial.

A este respecto ya existen iniciativas como la “***AI Ethics Guidelines produced by the Europea Commission’s High-Level Expert Group on Artificial Intelligence (AI HLEG)”*** y un compromiso firme de dotación económica para el despliegue de la Inteligencia Artificial.

Los principios de Asimilar

1. ***Meta de la investigación***: el objetivo de la investigación de la IA no debería ser crear inteligencia sin dirigir, sino inteligencia beneficiosa.
2. ***Financiación de la investigación***: la inversión en IA debería ir acompañada de fondos para investigar en asegurar su uso beneficioso, incluyendo cuestiones espinosas sobre ciencias de la computación, economía, legislación, ética y estudios sociales.
3. ***Enlace entre ciencia y política***: debería haber un intercambio constructivo y sano entre los investigadores de IA y los legisladores.
4. ***Cultura de la investigación***: una cultura de cooperación, confianza y transparencia debería ser fomentada entre los investigadores y desarrolladores de IA
5. ***Evitar las carreras***: los equipos que estén desarrollando sistemas de IA deberían cooperar activamente para evitar chapuzas en los estándares de seguridad.
6. ***Seguridad***: los sistemas de IA deberían ser seguros a lo largo de su vida operativa, y verificables donde sea aplicable y posible.
7. ***Transparencia en los fallos***: si un sistema de IA causa daño debería ser posible determinar por qué.
8. ***Transparencia judicial:*** *cualquier intervención de un sistema autónomo en una decisión debería ir acompañada de una explicación satisfactoria y auditable por parte de una autoridad humana competente*.
9. ***Responsabilidad***: los diseñadores y desarrolladores de sistemas avanzados de IA son depositarios de las implicaciones morales de su uso, mal uso y acciones, con la responsabilidad y oportunidad de dar forma a dichas implicaciones.
10. ***Alineación de valores***: los sistemas de IA altamente autónomos deberían ser diseñados para que sus metas y comportamientos puedan alinearse con los valores humanos a lo largo de sus operaciones.
11. ***Valores humanos***: los sistemas de IA deberían ser diseñados y operados para que sean compatibles con los ideales de dignidad humana, derechos, libertades y diversidad cultural.
12. ***Privacidad personal***: la gente debería tener el derecho de acceder, gestionar y controlar los datos que generan, dando a los sistemas de IA el poder de analizar y utilizar esa información.
13. ***Libertad y privacidad***: la aplicación de la IA a los datos personales no puede restringir de forma poco razonable la libertad, real o sentida, de las personas.
14. ***Beneficio compartido***: las tecnologías de IA deberían beneficiar y fortalecer a tanta gente como sea posible.
15. ***Prosperidad compartida***: la prosperidad económica creada por la IA debería ser compartida ampliamente, para el beneficio de toda la Humanidad.
16. **Control humano:** los seres humanos deberían escoger cómo y si delegan decisiones a los sistemas de IA para completar objetivos escogidos previamente**.**
17. ***Sin subversión***: el poder conferido por el control de sistemas de IA altamente avanzados debería respetar y mejorar, más que subvertir, los procesos sociales y cívicos de los que depende la salud de la sociedad.
18. **Carrera armamentística:** debería ser evitada cualquier carrera armamentística de armas autónomas letales**.**
19. ***Capacidad de precaución***: al no haber consenso, deberíamos evitar las asunciones sobre los límites superiores de las futuras capacidades de la IA.
20. ***Importancia***: la IA avanzada podría representar un profundo cambio en la historia de la vida en la Tierra, y debería ser planificada y gestionada con el cuidado y los recursos adecuados.
21. ***Riesgos***: los riesgos asociados a los sistemas de IA, especialmente los catastróficos o existenciales, deben estar sujetos a planificación y esfuerzos de mitigación equiparables a su impacto esperado.
22. Automejora recursiva: los sistemas de IA diseñados para automejorarse recursivamente o autorreplicarse de una forma que pudiera llevar al rápido incremento en su calidad o cantidad deben estar sujetos a unas estrictas medidas de control y seguridad.
23. **Bien común: la superinteligencia debería ser desarrollada sólo en servicio de unos ideales éticos ampliamente compartidos y para beneficio de toda la Humanidad, más que para un Estado u organización.**

****

1. Informe sobre los principios de Asimilar en Inteligencia Artificial. VDW (Federación de Científicos Alemanes) [↑](#footnote-ref-1)
2. Mids, Brain and programas (Searle, 1980) [↑](#footnote-ref-2)