La ética de la inteligencia artificial | Ethics of Artificial Intelligence

La ética de la inteligencia artificial | [Ethics of Artificial Intelligence](#)

Fig. I A. C4.1.1- Crédito Imag. (Pixabay). URL: <https://pixabay.com/es/illustrations/inteligencia-artificial-cerebro-4389372/>

⇐ Ir al índice principal del libro	⇐ Go to the main index of the book
Contenidos	Contents

| Autor / [Author](#): [Juan Antonio Lloret Egea](#) | Miembro de la [Alianza Europea para la IA](#) / [Member to the European AI Alliance](#) | <https://orcid.org/0000-0002-6634-3351> | © 2019. Licencia de uso y distribución / [License for use and distribution](#): [[Los estados de la inteligencia artificial \(IA\)](#) | [The states of artificial intelligence \(AI\)](#)] [creative commons CC BY-NC-ND](#) | [ISSN 2695-3803](#) | | Escrito / [Writed](#): 16/07/2019. Actualizado / [Updated](#): 16/08/2019.

(Extracto del [artículo publicado en Scientific American](#) por Shohini Kundu, estudiante de doctorado en la Universidad de Chicago).

Extracto más adelante unas líneas de pensamiento que son escritas por Shohini Kundu. A las que quiero añadir escuetamente lo siguiente: supongamos que por correlación (causa-efecto) llegamos a la conclusión de que el cáncer de mama está relacionado con llevar falda. Entonces podríamos llegar (por inferencia u otra lógica también) a la conclusión de que la falda es la causa y el efecto el cáncer de mama. Y claro, muchas mujeres llevan falda y el error sería mayúsculo. En latín se decía aquello de *Cum hoc ergo propter hoc...* Correlación no implica causalidad.

[\[English\]](#)

I extract below some lines of thought that are written by Shohini Kundu. To which I want to briefly add the following: suppose that by correlation (cause-effect) we conclude that Breast cancer is related to wearing a skirt. Then we could reach (by inference or other logic as well) the conclusion that the skirt is the cause and effect of breast cancer. And of course, many women wear skirts and the mistake would be capitalized. In Latin it was said that of *Cum hoc ergo propter hoc...* Correlation does not imply causation.

Los éxitos de los sistemas de juego de inteligencia artificial en la derrota del campeón mundial de ajedrez Gary Kasparov y el campeón mundial de go contra Alphago, Ke Jie, subrayan el aspecto cualitativo de la inteligencia artificial que demostró ser superior a los expertos humanos en calcular el impacto de las decisiones actuales en posibles movimientos futuros. Desafortunadamente, en ese proceso de toma de decisiones, la IA también eliminó la transparencia, la explicabilidad, la previsibilidad, la capacidad de enseñanza y la auditabilidad del movimiento humano, reemplazándolo con opacidad. La lógica del movimiento no sólo es desconocida para los jugadores, sino también desconocida para los creadores del programa.

Fig. I A. C4.1.2- Crédito Imag. (Pixabay). URL: <https://youtu.be/gr185aI5vqI>

Imagine una situación en la que su hijo llega a casa y le pide dinero para ir a ver una película con sus amigos. Y usted cede. Una semana después, otro de sus hijos llega a usted con la misma solicitud pero esta vez usted se niega. Esto planteará de inmediato la cuestión de la injusticia y el favoritismo. Para evitar cualquier acusación de favoritismo, usted le explica a su hijo que debe terminar su tarea antes de calificar para cualquier dinero de bolsillo. Sin ninguna explicación, seguramente habrá tensión en la familia.

Ahora imagine reemplazar su rol con un sistema de inteligencia artificial que ha reunido datos de miles de familias en situaciones similares. Al estudiar las consecuencias de las decisiones de asignación en otras familias, se llega a la conclusión de que un hermano debe obtener el dinero mientras que el otro hermano, no.

Para comprender cuán artificiales pueden ser las decisiones basadas en inteligencia artificial es importante examinar cómo los humanos toman decisiones. Las decisiones humanas pueden guiarse por un conjunto de reglas explícitas, o por asociaciones simplemente basadas en el consecuencialismo, o por una combinación. Los humanos también son selectivos acerca de qué información es relevante para tomar una decisión. Al carecer de selectividad, las máquinas pueden considerar factores que los humanos consideran impertinentes para tomar una decisión.

Hay innumerables ejemplos de esto, desde que Microsoft **cerró** su chatbot Tay después de que

comenzó a arrojar retóricas antisemitas incendiarias en Twitter, hasta un estudio de la Universidad de Boston que encontró asociaciones de género entre palabras como "jefe", "arquitecto" y "financiero" con hombres, y palabras como "enfermera" y "repcionista" con mujeres. Esto puede ser confirmado por los datos, pero contrasta con nuestros valores explícitos. Si los procesos basados en datos dependen de la salida generada por estos algoritmos de IA, producirán decisiones sesgadas, a menudo en contra de nuestros valores éticos.

Un programa de computadora utilizado por los tribunales de los EE. UU. identificó erróneamente a los acusados negros que no reincidieron durante un período de dos años con la probabilidad de convertirse en delincuentes reincidentes casi el doble que los acusados blancos: 45 por ciento en comparación con 23 por ciento. Si un humano hiciera lo mismo, sería denunciado como racista.

Al expone el cisma entre nuestros valores explícitos y nuestras experiencias colectivas. Nuestras experiencias colectivas no son estáticas. Están formados por importantes decisiones sociales, que a su vez están guiadas por nuestros valores éticos. ¿Realmente queremos dejar el proceso de toma de decisiones a máquinas que aprenden únicamente del pasado y, por lo tanto, están en deuda con él, en lugar de dar forma al futuro?

Dada la escala de las aplicaciones de IA en campos como el diagnóstico médico, los servicios financieros y la evaluación del empleo, las consecuencias de cualquier problema son inmensas. A medida que los algoritmos dependen de más y más características para mejorar la previsibilidad, la lógica que rige tales decisiones se vuelve cada vez más inescrutable. En consecuencia, perdemos el aspecto holístico de la toma de decisiones, descartando todos los principios a favor de las observaciones pasadas. En algunos casos, esto puede ser poco ético, en algunos, ilegal y en otros, miope. El algoritmo de reincidencia se burló descaradamente de principios como la presunción de inocencia y la igualdad de oportunidades.

La consistencia es indispensable para la ética y la integridad. Nuestras decisiones deben cumplir con un estándar más alto que la precisión estadística. Durante siglos las virtudes compartidas de confianza mutua, reducción de daños, igualdad y equidad han demostrado ser las piedras angulares esenciales para la supervivencia de cualquier sistema de razonamiento. Sin una coherencia lógica interna, los sistemas de IA carecen de solidez y responsabilidad, dos medidas críticas para generar confianza en una sociedad. Al crear una brecha entre el sentimiento moral y el razonamiento lógico, la inescrutabilidad de las decisiones basadas en datos excluye la capacidad de participar de manera crítica en los procesos de toma de decisiones ^{C4.1-1}

[\[English\]](#)

(Excerpt from the [article published in *Scientific American* by Shohini Kundu](#), a PhD student at the University of Chicago).

The successes of artificial intelligence game systems in the defeat of world chess champion Gary Kasparov and go world champion Ke Jie underscore the qualitative aspect of artificial intelligence that proved to be superior to human experts in calculating the impact of current decisions in possible future movements. Unfortunately, in that decision-making process, AI also eliminated the transparency, explainability, predictability, teaching capacity and auditability of the human movement, replacing it with opacity. The logic of the movement is not only unknown to the players, but also unknown to the creators of the program.

Fig. I A. C4.1.2- Crédito Imag. (Pixabay). URL: <https://youtu.be/hLfLPz7HxFA>

Imagine a situation in which your child comes home and asks for money to go to a movie with his friends. And you give in. A week later, another of your children comes to you with the same request but this time you refuse. This will immediately raise the issue of injustice and favoritism. To avoid any accusation of favoritism, you explain to your child that you should finish your homework before qualifying for any pocket money. Without any explanation, there will surely be tension in the family. Now imagine replacing your role with an artificial intelligence system that has gathered data from thousands of families in similar situations. When studying the consequences of the allocation decisions in other families, it is concluded that one brother should get the money while the other brother does not.

To understand how artificial decisions based on artificial intelligence can be, it is important to examine how humans make decisions. Human decisions can be guided by a set of explicit rules, or by associations simply based on consequentialism, or by a combination. Humans are also selective about what information is relevant to make a decision. Lacking selectivity, machines can consider factors that humans consider impertinent to make a decision.

"There are innumerable examples of this, from Microsoft [shutting down](#) its chatbot Tay after it started spewing incendiary anti-Semitic rhetoric on Twitter, to a Boston University [study](#) that found gender associations between words like "boss," "architect" and "financier" with men, and words like "nurse" and "receptionist" with women. This may be borne-out by *data*, but it stands in contrast with our explicit values. If data-driven processes rely on output generated by these AI algorithms, they will produce biased decisions, often against our ethical values".

A computer program used by courts in the USA misidentified black defendants who did not relapse for a period of two years with the probability of becoming repeat offenders almost twice as many as white defendants: 45 percent compared with 23 percent. If a human did the same, he would be denounced as a racist.

AI exposes the schism between our explicit values and our collective experiences. Our collective experiences are not static. They are made up of important social decisions, which in turn are guided by our ethical values. Do we really want to leave the decision-making process to machines that learn only from the past and, therefore, are indebted to it, rather than shaping the future?

Given the scale of AI applications in fields such as medical diagnosis, financial services and employment evaluation, the consequences of any problem are immense. As algorithms rely on more and more features to improve predictability, the logic that governs such decisions becomes increasingly inscrutable. Consequently, we lose the holistic aspect of decision making, discarding all principles in favor of past observations. In some cases, this may be unethical, in some, illegal and in others, myopic. The recidivism algorithm blatantly made fun of principles such as the presumption of innocence and equal opportunities.

Consistency is indispensable for ethics and integrity. Our decisions must meet a higher standard than statistical accuracy. For centuries the shared virtues of mutual trust, harm reduction, equality and equity have proven to be the cornerstones essential for the survival of any reasoning system. Without an internal logical coherence, AI systems lack solidity and responsibility, two critical measures to build trust in a society. By creating a gap between moral sentiment and logical reasoning, the inscrutability of data-based decisions excludes the ability to participate critically in decision-making processes^{C4.1-1}.

“

Conclusión al artículo / Conclusion to the article

Quizá la parte más preocupante es si la IA (o un *chatbot*) escapa incluso al control humano o a sus propios creadores (programadores) por falta de transparencia o fallas en el diseño no previstas en el modelo. Lo que nos dirige inequívocamente a esa falta de explicabilidad y la previsibilidad... Y aún podemos ir más allá y preguntarnos: ¿qué sucede si dos máquinas hablan entre ellas y como consecuencia de los algoritmos internos inventan su propio lenguaje y deciden caminar solas? Les dejo para que reflexionen con [un artículo](#) que habla de esto precisamente. Interesante. La consistencia es indispensable para la ética y la integridad, sí.

La IA de Facebook crea su propio lenguaje en una espeluznante vista previa de nuestro futuro potencial (Fuente: Forbes)

[English]

Perhaps the most worrying part is whether AI (or a chatbot) escapes even human control or its own creators (programmers) due to lack of transparency or design flaws not provided for in the model. Which leads us unequivocally to this lack of explainability and predictability ... And we can go further and ask ourselves: what happens if two machines speak to each other and as a consequence of the internal algorithms invent their own language and decide to walk alone? I leave you to reflect with an article that speaks about this precisely. Interesting. Consistency is indispensable for ethics and integrity, yes.

Facebook AI Creates Its Own Language In Creepy Preview Of Our Potential (Source: Forbes).

“ Por primera vez en China, la tecnología de asistencia #AI se usó en un juicio en el Tribunal Popular Intermedio Nº2 de Shanghai el miércoles, informó Legal Daily. Cuando el juez, el fiscal o el defensor le preguntaron al sistema de inteligencia artificial, éste mostró todas las pruebas relacionadas en una pantalla de la sala / For the first time in China, #AI assistive technology was used in a trial at Shanghai Nº 2 Intermediate People's Court on Wed, the Legal Daily reported. When the judge, public prosecutor or defender asked the AI system, it displayed all related evidence on a courtroom screen. pic.twitter.com/fEI7cR5U3T— People's Daily, China (@PDChina) 25 de enero de 2019

Colorario / Corollary

Mi posición final coincide casi al 100% con los razonamientos expuestos por Christopher Barnatt. Los considero muy equilibrados y ajustados a un futuro próximo. Pero quizá usted tenga otras conclusiones diferentes a mí. / My final position coincides almost 100% with the arguments put forward by Christopher Barnatt. I consider them very balanced and adjusted to the near future. But maybe you have other conclusions different from me.

Fig. I A. C4.1.3- Crédito Imag. (ExplainingComputers-Christopher Barnatt. [Vídeo embebido.Tenga en cuenta que el vídeo está excluido de la licencia Creative Commons y no se puede redistribuir de ninguna manera / [Embedded Video. Notice that the video is excluded from the Creative Commons license and cannot be redistributed in any way](#)]). URL: <https://youtu.be/hLfLPz7HxFA>

“ ¿La IA destruirá empleos, dará más poder a las compañías tecnológicas y pondrá en peligro a la raza humana? ¿O la IA mejorará nuestras vidas y nos permitirá sobrevivir y prosperar frente a los grandes desafíos que nos esperan? / [Will AI destroy jobs, hand more power to tech companies, and endanger the human race? Or will AI enhance our lives and allow us to survive and thrive in the face of grand challenges ahead?](#)

“ This is how human will power the future through [#automation](#); [#robotic](#).

By [@intel](#)

[#AI](#) [#robot](#) [#futureofwork](#) [#DigitalTransformation](#) <https://t.co/xh4w6Apszh>

[@FrRonconi](#) [@RaminNasibov](#) [@steube](#) [@JeffreyBuskey](#) [@JolaBurnett](#)

[@arlenenewbigg](#) [@terence_mills](#) [@kuriharan](#) [@MarshaCollier](#)

pic.twitter.com/XnRgN1uVMu

— Automeme ([@automeme](#)) [August 28, 2019](#)

Bibliografía | [Bibliografy](#)

[IC4.1-1] Kundu, S. (3 de julio, 2019). *Ethics in the Age of Artificial Intelligence*. Scientific American. [Recupedado (16/08/2019) de <https://blogs.scientificamerican.com/observations/ethics-in-the-age-of-artificial-intelligence/>]

© 2019. Licencia de uso y distribución / [License for use and distribution](#): [[Los estados de la inteligencia artificial \(IA\)](#) | [The states of artificial intelligence \(AI\)](#)] [creative commons CC BY-NC-ND](#) | [ISSN 2695-3803](#) |

- [Notas legales](#) / [Legal notes](#)

- [Página web de Formaempleo](#) / [Formaempleo website](#)
-

Revisión #1

Creado el Sat, Nov 23, 2019 11:54 AM por [Juan Antonio Lloret Egea](#)

Actualizado el Sat, Nov 23, 2019 11:55 AM por [Juan Antonio Lloret Egea](#)