

MORAL MACHINE: DESCUBRIENDO EL DESAFÍO ÉTICO DE LAS IA

Juan Felipe Guevara

Universidad Autónoma Metropolitana-Cuajimalpa (México)

En el 2017, el director ejecutivo de Netflix, Reed Hastings, anunció que la competencia por parte de Amazon y HBO no le quitaba el sueño, pues la verdadera competencia de la plataforma era precisamente el sueño. La sentencia exuda una confianza desmesurada en el éxito de su producto, al tiempo que revela una fisura en su concepción: por más que los algoritmos de Netflix aprendan sobre los hábitos de consumo audiovisual de sus usuarios y lleguen a hacerles recomendaciones cada vez más atinadas e interesantes, o que la compañía misma aproveche ese aprendizaje para decidir qué tipo de producciones va a financiar y lanzar en el futuro, el sueño, aspecto vital de la condición humana, queda fuera del poder calculador de dichos algoritmos. Una situación similar ocurre con el experimento de la Moral Machine (MM) que ha mantenido el Instituto Tecnológico de Massachusetts (MIT, por su sigla en inglés) desde el 2016 (<https://www.moralmachine.net>). En este caso, lo que parece quedar fuera del cómputo de este experimento de inteligencia artificial (IA) es el problema de la vulnerabilidad y la responsabilidad ética.

La MM es presentada como un elemento clave en una serie de proyectos tecnológicos más amplios concernientes a la difusión y empleo masivo de artefactos autónomos –capaces de tomar decisiones por sí mismos–, con un especial énfasis en los vehículos de transporte. El ejercicio está encaminado a la programación de la IA que habrá de tomar las decisiones y su finalidad radica en alinear los algoritmos morales con los valores humanos mediante una «conversación global» acerca de cuáles son esos valores aceptados y consensuados

(Bonnefon, Shariff y Rawan, 2016). Para llevar a cabo una conversación de tal magnitud, el equipo del MIT puso en marcha la MM: una plataforma virtual en la cual se presentan diversos escenarios de dilemas éticos modelados según el consabido dilema del tranvía, a la que puede acceder cualquier persona alrededor del mundo que tenga una conexión a internet e interés en responder a los dilemas planteados¹. Además de consignar las elecciones realizadas por cada participante, también es posible indicar el género, la nacionalidad, la edad, entre otras. En términos de recolección de datos, el experimento fue un éxito rotundo: en total se obtuvieron 39.61 millones de decisiones de participantes provenientes de 231 entidades geopolíticas (Awad *et al.*, 2018). Derivado del análisis de los datos, el equipo del MIT identificó una serie de tendencias globales y locales que les permitiría afrontar el problema de la aceptación social de los vehículos autónomos según el contexto cultural y legal de cada territorio. Ciertamente, algunas de estas tendencias resultan llamativas y cabe señalar que una de las conclusiones principales del estudio muestra una marcada tendencia utilitarista al momento de tomar decisiones en los dilemas morales propuestos. Esta tendencia, no obstante, bien puede deberse a un sesgo introducido por el modelo empleado de dilema del tranvía, algo que señalan tanto académicos (ver Bigman y Gray, 2020) como muchos de los comentarios que dejaron en la página de la MM quienes participaron.

Aunque estos resultados son en sí mismos una rica fuente de discusiones éticas, la pregunta que involucra a las IA surge con el planeamiento mismo del experimento: ¿por qué es necesario alinear los algoritmos morales con los valores humanos? Con este interrogante en mente, el interés ético se desplaza de las decisiones individuales que se puedan tomar con relación a un dilema hacia la percepción humana de la autonomía de las máquinas (ver Bonnefon, Shariff y Rahwan, 2016; Shariff, Bonnefon y Rahwan, 2017; Noothigattu *et al.*,

1. El sitio de internet en cuestión, <https://www.moralmachine.net/>, continúa en funcionamiento y todavía se puede participar del experimento.

2018). Después de todo, de las decisiones que tomen estos vehículos dependerán vidas humanas, tanto las de quienes se transportan en ellos como las de quienes comparten los espacios viales en otras modalidades de transporte. Desde esta perspectiva, se empieza a perfilar la pregunta por la vulnerabilidad y la responsabilidad a partir de las implicaciones que conlleva caracterizar a estos vehículos como autónomos. ¿Qué sentido tiene hablar de autonomía en relación a máquinas y vehículos? Ronald Arkin, uno de los defensores más vocales y reconocidos del uso de robots en situaciones de guerra, establece la siguiente distinción: para la filosofía, «la autonomía añade agencia moral y libre albedrío a un sistema robótico, algo que no existe aún y que no existirá por un buen tiempo si es que alguna vez lo hace», mientras que para la robótica, simplemente se trata de delegar la toma de decisiones a una máquina que ha sido previamente programada por un humano (Arkin, 2017: 35). Desde esta perspectiva, la autonomía es simplificada a una respuesta automática en la cual no interviene un proceso de razonamiento humano en el momento mismo de tomar la decisión. La agencia humana, con todo y los problemas que ella misma acarrea, funge como modelo de programación, lo cual para el experimento de la MM se traduce en que las decisiones individuales recogidas quedan subsumidas en la decisión social programada en el algoritmo moral, al tiempo que la sociedad queda reducida al agregado de decisiones recolectadas y su estandarización. Al final del proceso, la pretendida autonomía de los vehículos es equivalente a un «perfil de preferencias anónimo» (Noothingattu *et al.*, 2018).

Visto con cuidado, el modo en el que la autonomía de las máquinas aparece en estos planteamientos es similar al de una ficción útil para explorar y evaluar las respuestas morales humanas. Pese a ello y aunque el algoritmo moral que rige sus decisiones sólo pretenda ser una extensión de valores social y culturalmente sancionados bajo la ley, no es posible dejar de lado el hecho de que es la máquina la que está a cargo de tomar las decisiones. Lejos de ser un resultado anecdótico inesperado, es parte fundamental del proyecto que la máquina sea quien toma las decisiones. Las publicaciones del grupo del MIT exaltan constantemente las ventajas que ofrece la programación de la máquina comparada con los humanos en este respecto. En primer lugar, y quizás la más repetida, se encuentra la muy superior capacidad de cómputo de las máquinas, capaces de sopesar una inmensa cantidad de escenarios y sus resultados en fracciones de segundo: ahí donde al ser humano no le queda más que reconocer el impulso

del instinto para tomar una decisión que no da lugar a la cuidadosa reflexión consciente, la máquina ofrece un ensanchamiento temporal que transforma el instante del instinto en la pausa del cómputo. De hecho, Noothingattu *et al.* (2018: 20) consideran que una de las mayores limitaciones de la MM estaría en que las personas no emplearan el tiempo suficiente para tomar la decisión que consideran más apropiada en cada escenario propuesto. En segundo lugar, en tanto que las máquinas son programadas para cumplir no sólo con las expectativas morales de los humanos, sino también con las leyes que regulan el tránsito, eso en sí mismo constituye una característica deseable pues se asume que el cumplimiento de la ley es garante de una disminución en los accidentes. La programación de las máquinas sería incapaz de violar la ley, en otras palabras, sólo para las máquinas la ley sería efectivamente *ley*.

Ahora bien, la discusión que se puede tener en torno a la autonomía pareciera ser limitada cuando se introduce la distinción entre el ámbito de la filosofía y el de la robótica. Sin embargo, el acotamiento operacional que hace la robótica de la autonomía como delegación es suficiente para invitar a otro tipo de reflexión filosófica que no se centre exclusivamente en el espinoso tema de la agencia. Al contrario de lo que indica el prefijo *auto-*, la autonomía de las máquinas refiere constantemente a la autonomía humana. Sobre este punto podrían llover las acusaciones de antropocentrismo y desenmascarar la incapacidad de pensar la agencia de entidades diferentes a los seres humanos. Pero no hay que apresurarse. El lenguaje utilizado para referirse a esta relación entre máquinas y humanos es muy dicente. Por un lado, Bonnefon, Shariff y Rahwan (2016: 1576) hablan de «dotar (*endow*) millones de vehículos con autonomía», pues los algoritmos son programados por seres humanos (con todo y sus sesgos y prejuicios). Por otro lado, los mismos autores mencionan en el resumen de su artículo que un vehículo autónomo tiene la opción de sacrificarse a sí mismo junto con su(s) pasajero(s). Probablemente sólo se trate de un desliz de lengua, dado que en el cuerpo del artículo no vuelven a referirse al sacrificio de los vehículos y sólo hablan del sacrificio humano. Pero no deja de llamar la atención el sacrificio de las máquinas. Estos dos factores revelan el aspecto vulnerable de estas máquinas en su ineludible relación con los seres humanos. En tanto que la preocupación que articula a la MM es la percepción humana de la autonomía de las máquinas, hablar de la posibilidad de que se sacrifiquen es ya un tema de controversias. De igual forma, la manera en la que sean programados sus algoritmos morales condiciona-

rá el tipo de decisiones que tomen, la forma en la que eventualmente podrían aprender de esas decisiones y, por ende, su recepción por parte de sus potenciales usuarias y otros grupos de interés. La vulnerabilidad en este contexto se manifiesta como situaciones que no se pueden anticipar, para las cuales cualquier tipo de preparación es insuficiente y que demandan una respuesta a través de la cual se constituye una obligación ética (Butler, 2012). Esta noción ciertamente cabe en el escenario que concierne a las inteligencias artificiales y los vehículos autónomos.

La separación entre la noción de autonomía para la filosofía y para la robótica pone de manifiesto que el rasgo al cual no se quiere renunciar, el mismo que genera una buena parte de los problemas que acechan a la agencia, es la individualidad. No obstante, partiendo incluso de la definición operacional de autonomía de la robótica ha sido posible llegar a una visión relacional de la misma. En ese sentido, las éticas con pretensiones universalistas basadas en la soberanía de un agente racional individual o en la suma de los mismos, como es el caso del utilitarismo, es insuficiente para encarar el desafío ético de la IA. Una situación similar se puede constatar con la actual contingencia sanitaria y social por la COVID-19 y la escasez de ventiladores en hospitales y clínicas, lo cual ha llevado a la producción de las llamadas «Guías bioéticas de

triaje» que ofrecen criterios para decidir a quiénes se les conceden estos recursos y, con ello, justificar la manera de proceder del personal de salud². El problema es que para estas guías sólo existen individuos, ya sean las personas afectadas por el virus o el personal de salud, y se pierde de vista la importancia de considerar estos factores en escenarios más complejos y problemáticos, como se ha intentado hacer recientemente con la recomendación de cambiar el estatus de pandemia a sindemia (ver Horton, 2020) o con una consideración ética y filosófica de la COVID-19 que no se reduzca a números y estadísticas (ver González Valerio y Martínez Ruíz, 2020). Tanto los singulares interrogantes que despiertan las IA como el desconcierto que ha generado la COVID-19 requieren de una aproximación relacional, de una ética especulativa (Puig de la Bellacasa, 2017), que disloque la figura del individuo y permita pensar desde los intersticios y las superficies de contacto.

AGRADECIMIENTOS

Este texto se ha nutrido de las discusiones sostenidas en el Seminario de Vulnerabilidad, Tecnologías y Ética (UAM-Cuajimalpa). Le agradezco especialmente a José Agustín Mercado Reyes por compartir sus conocimientos sobre algoritmos. Este texto fue realizado con el apoyo del proyecto ReNACE, CONACyT A1-S-21700: *Exploración de Creatividad Relacional No-Antropocéntrica*.

- 2 Curiosamente, el sitio de internet de la MM cuenta también con una encuesta de este estilo en la cual se le pide a las participantes definir el peso que tienen diversos factores, tales como condiciones médicas preexistentes, edad, estatus socioeconómico, entre otros, al momento de decidir a quién se le asignan los recursos médicos escasos o a quién se le pudieran quitar para concedérselos a otra persona.

REFERENCIAS

- Awad, Edmond et al. (2018). The Moral Machine Experiment. *Nature*, 563 (7729): 59–64. <https://doi.org/10.1038/s41586-018-0637-6>
- Arkin, R. C. (2017). A Robotocist's Perspective on Lethal Autonomous Weapon Systems. En: United Nations, *UNODA Occasional Papers No. 30*. Nueva York: UN Publications, pp. 35–47. <https://doi.org/10.18356/7748aa31-en>
- Bigman, Yochanan E. y Gray, Kurt. (2020). Life and Death Decisions of Autonomous Vehicles. *Nature*, 579 (7797): E1–E2. <https://doi.org/10.1038/s41586-020-1987-4>
- Bonnefon, Jean-François, Shariff, Azim y Rahwan, Iyad. (2016). The Social Dilemma of Autonomous Vehicles. *Science*, 352 (6293): 1573–1576. <https://doi.org/10.1126/science.aaf2654>
- Butler, J. (2012). Precarious Life, Vulnerability, and the Ethics of Cohabitation. *The Journal of Speculative Philosophy*, 26 (2): 134–151. <https://doi.org/10.5325/jspecphil.26.2.0134>
- González Valerio, María Antonia y Martínez Ruíz, Rosaura. 22 de abril de 2020. *Triage. Decidir lo éticamente indecidible*. El Universal [online]. <https://www.eluniversal.com.mx/opinion/maria-antonio-gonzalez-valerio-y-rosaura-martinez-ruiz/triaje-decidir-lo-eticamente>
- Horton, Richard. (2020). Offline: COVID-19 Is Not a Pandemic. *The Lancet*, 396 (10255): 874. [https://doi.org/10.1016/S0140-6736\(20\)32000-6](https://doi.org/10.1016/S0140-6736(20)32000-6)
- Noothigattu, Ritesh et al. (2018). A Voting-Based System for Ethical Decision Making. *ArXiv:1709.06692 [Cs]*: 1–25.
- Puig de la Bellacasa, María. (2017). *Matters of Care: Speculative Ethics in More Than Human Worlds*. Minneapolis: University of Minnesota Press.
- Shariff, Azim, Bonnefon, Jean-François y Rahwan, Iyad. (2017). Psychological Roadblocks to the Adoption of Self-driving Vehicles. *Nature Human Behaviour*, 1 (10): 694–696. <https://doi.org/10.1038/s41562-017-0202-6>