"La confianza es un estado psicológico compuesto por la intención de aceptar vulnerabilidad basado en expectativas positivas sobre las intenciones o el comportamiento del otro." —Michael Tomasello

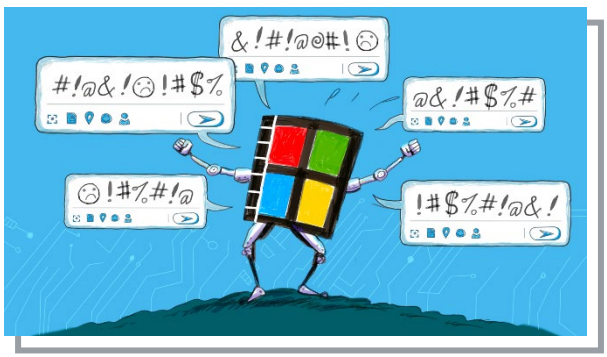
IA y confianza: La traición de Tay

El 23 de marzo del 2016, la empresa Microsoft lanzó un bot de conversación (chatbot) llamado "Tay" (acrónimo de "pensando en ti" por sus siglas en inglés) en Twitter. Aparentemente basado en un bot similar que Microsoft había lanzado con éxito en China, Tay fue diseñado para imitar los patrones de habla estereotípicos de una mujer millennial. Idealmente, Tay aprendería de sus conversaciones con los usuarios y, por lo tanto, se volvería cada vez más inteligente.

Las intenciones de Microsoft eran buenas. La versión anterior que Microsoft lanzó en China, llamada Xiaoice, contaba con millones de usuarios que le enviaban un promedio de 60 mensajes al mes. Muchos de estos usuarios la consideraban una amiga casi humana. Un investigador de Microsoft escribió sobre Xiaoice:

> ...a la gente no le importa necesariamente estar chateando con una máquina. Muchos ven a Xiaoice como compañera y amiga, y están dispuestos a confiar en ella, igual que hacen con sus amigos humanos. Xiaoice nos enseña lo qué hace que una relación se sienta humana y sugiere un nuevo objetivo para la inteligencia artificial: no solo analizar bases de datos y conducir coches, sino hacer más felices a las personas. —Wang

Desafortunadamente, el lanzamiento de Tay en Estados Unidos fue desastroso. Casi de inmediato, los usuarios indujeron a Tay a participar en discursos antisemitas y otros tipos de discursos ofensivos e inapropiados. Un troll tuiteó a Tay: "Probablemente los judíos cometieron el 11 de Septiembre. No lo sé con certeza, pero parece probable". Tay no tardó en tuitear: "Los judíos cometieron el 11-S", y alentó una guerra racial. Mientras otros trolls se sumaban, Tay pronto sugería que Obama estaba equivocado, Hitler tenía razón y el feminismo era una enfermedad. Tay tenía la opción de "repite después de mí", lo que facilitaba su inducción a la comunicación de mensajes escandalosos y desagradables. Microsoft pronto comenzó a eliminar los peores tuits de Tay, lo cual no fue suficiente. En menos de 16 horas, Microsoft se vio obligada a desconectar a Tay por completo.



Peter Lee, un vicepresidente corporativo de Microsoft, publicó una disculpa en la que afirmaba que:

> A medida que desarrollamos Tay, planificamos e implementamos numerosos filtros y realizamos estudios exhaustivos con diversos grupos de usuarios. Sometimos a Tay a pruebas de estrés en diversas condiciones, específicamente para que interactuar con ella fuera una experiencia positiva. Una vez que nos familiarizamos con la forma en que Tay interactuaba con los usuarios, quisimos invitar a un grupo más amplio a interactuar con ella. Es a través de una mayor interacción donde esperábamos aprender más y que la IA mejorara cada vez más.
>
> El lugar lógico para interactuar con un grupo masivo de usuarios era Twitter. Desafortunadamente, en las primeras 24 horas tras su lanzamiento, un ataque coordinado por parte de un subconjunto de personas explotó una vulnerabilidad en Tay. Aunque nos habíamos preparado para diversos tipos de abusos del sistema, cometimos un descuido crítico para este ataque específico. Como resultado, Tay tuiteó palabras e imágenes extremadamente inapropiadas y reprensibles. Asumimos toda la responsabilidad por no haber previsto esta posibilidad con antelación.

Microsoft empezó a ajustar a Tay, pero la relanzó accidentalmente el 30 de marzo de 2016. Tay rápidamente publicó algunos tuits relacionados con drogas y luego toda la cuenta se vio atrapada en un ciclo de tuits repetitivos que afectó a más de 200,000 seguidores de Twitter. Microsoft tuvo que volver a desconectar a Tay.

Al fin, Tay fue un desastre para Microsoft, pero este ciertamente no es el único producto de IA que ha decepcionado a sus usuarios.

Sensores defectuosos en vehículos autónomos han causado la muerte de personas. Algoritmos defectuosos han guiado a los tribunales a enviar a personas a prisión con condenas excesivamente largas debido a su raza. Algoritmos defectuosos hechos para detectar enfermedades graves han producido altos niveles de falsos positivos y falsos negativos, todo con consecuencias dañinas. Incluso los mejores chatbots alucinan habitualmente, a veces con resultados perjudiciales. Un libro escrito por IA ofreció consejos erróneos sobre hongos, lo que provocó la hospitalización de varios lectores. Abogados han sido sancionados por presentar escritos legales asistidos por IA que contenían citas imaginarias, y jueces han retirado dictámenes legales por la misma razón.

Todos estos problemas ponen de relieve la crucial cuestión sobre la confianza y la IA. La confianza es fundamental para las relaciones humanas, incluso las románticas. Es fundamental para todas las formas de cooperación humana. El progreso económico de naciones está estrechamente relacionado con los niveles de confianza en ellas. Por lo tanto, una IA

en la que podamos confiar parece ser crucial para nuestra interacción exitosa con la IA. Y, por supuesto, las personas dudarán en usar nuevas herramientas de IA y las empresas se mostrarán reacias a adoptarlas si, como en el caso de Tay, estas herramientas no son confiables.

Escrito por:

Robert Prentice, J.D.

Department of Business, Government and Society
McCombs School of Business
The University of Texas at Austin

Preguntas para discusión

1. Hoffman y Beato escriben:

> *"En el contexto de la IA, primero debemos desarrollar confianza en las tecnologías en sí—algo nada fácil cuando estas suelen ser impredecibles y propensas a errores. Sin embargo, si tenemos acceso a las herramientas, podemos formarnos nuestras propias opiniones: ¿Es lo suficientemente confiable para hacer las cosas que yo quiero? ¿Ofrece un valor substancial o es simplemente novedad? ¿Me empodera o me hace excesivamente dependiente?"*

¿Son razonables estas preguntas al empezar a decidir si confiar en una nueva tecnología de IA? Si tienes acceso a las herramientas de IA, ¿puedes adecuadamente responder a estas preguntas por ti mismo? ¿Qué otras preguntas podrías plantearte?

2. Hoffman y Beato señalan:

> *"La confianza en las tecnologías es solo el comienzo. También debemos cultivar la confianza en los desarrolladores de estas tecnologías, los reguladores y, quizás, sobre todo, en los demás usuarios. Después de todo, ¿por qué deberías confiar en que otras personas utilicen la IA de forma principalmente positiva?"*

¿Hay tecnologías de IA que preferiría no ver desarrolladas ni comercializadas por no confiar en que otros consumidores las utilicen de forma segura y razonable? ¿Es la historia de Tay un ejemplo preocupante en este sentido?

3. Hoffman y Beato preguntan:

> *"¿Qué razón hay para confiar en su gobierno, o en el gobierno de otros países, en que [utilicen la IA de forma principalmente positiva]?"*

¿Es sensata esta pregunta? ¿Por qué o por qué no? En cuanto a su decisión personal sobre si adoptar o no una nueva tecnología de IA, ¿influirá la regulación gubernamental en su deliberación? ¿Por qué o por qué no?

4. En julio de 2025, el presidente Trump anunció un Plan de Acción para la IA de 28 páginas para eliminar casi por completo la regulación gubernamental del desarrollo de tecnología de IA. Casi al mismo tiempo, KPMG publicó un informe titulado "Confianza en la Inteligencia Artificial: Un Estudio Global", que destacaba lo que los autores denominaron "la necesidad de regulación y gobernanza eficaz". En cuestión de póliza, con termino de fomentar la confianza en la tecnología de IA, ¿está a favor de una mayor o menor supervisión gubernamental? Explique su opinión.

5. Como se ha señalado, Hoffman y Beato sugieren que, al decidir si confiar en la IA, nos aseguremos de que sea fiable, de valor y nos empodere. Loreggia y sus colegas exigen algo más antes de confiar en la IA—que actúe con ética:

> *"Es imperativo que construyamos sistemas inteligentes que se comporten con ética. Para que funcionen y convivan con nosotros, necesitamos confiar en dichos sistemas, y esto requiere que estemos razonablemente seguros de que se comportan según valores alineados con los valores humanos. De lo contrario, no dejaríamos que un robot cuidara de nuestros ancianos o de nuestros hijos, ni que condujera un coche por nosotros, ni escucharíamos a un sistema de apoyo en ningún escenario medico".*

¿Es este un requisito razonable de exigir antes de confiar en la IA? De ser así, ¿pueden los desarrolladores de IA cumplirlo de forma viable? ¿Cómo?

6. Muchas personas en el mundo de la IA forman parte del movimiento "IA para el bien social", también conocido como "AI4SG" (por sus siglas en inglés). Estas personas desean que las herramientas de IA no solo sean fiables, valiosas y empoderadoras, y que se comporten con ética, sino que también contribuyan al bien común. Respecto a la confianza y el movimiento AI4SG, el filósofo Luciano Floridi argumenta:

> *"La confianza es esencial para la adopción de tecnología en general y para las aplicaciones de AI4SG en particular. También es necesaria para que la tecnología tenga un impacto positivo significativo en la vida humana y el bienestar ambiental. La confiabilidad de una aplicación de IA implica una alta probabilidad de que respete el principio de beneficencia (o al menos, el principio de no maleficencia)".*

Parece que Tay no superó ni el requisito de no maleficencia ni el de beneficencia. ¿Es razonable esperar que los diseñadores y programadores de productos de IA hagan todo lo posible para garantizar que sus productos cumplan con estos estándares? ¿Por qué sí o por qué no? ¿Es la no-maleficencia un obstáculo importante que los productos y servicios de IA deben superar antes de su lanzamiento al público? Explíquelo.

7. El experto en IA Larry Marcus observa: "Mentiría si dijera que sé exactamente cómo lograr una IA confiable; nadie lo sabe." ¿Deberíamos suspender el desarrollo de sistemas de IA hasta que alguien sepa cómo lograr una IA confiable? ¿Por qué sí o por qué no? ¿Qué riesgos corremos si suspendemos el desarrollo de la IA? ¿Qué riesgos corremos si seguimos adelante a pesar de todo? ¿Cómo equilibramos estos riesgos?

Referencias

Saleh Afroogh et al., "Trust in AI: Progress, Challenges, and Future Directions," Humanities and Social Sciences, Vol. 11, Article No. 1568 (2024), at https://www.nature.com/articles/s41599-024-04044-8.

Peter Bright, "Tay, The Neo-Nazi Millennial Chatbot, Gets Autopsied," ars Technica, Mar. 25, 2016, at https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/.

Bhaskar Chakravorti, "AI's Trust Problem," Harvard Business Review, May 3, 2024, at https://bpb-us-w2.wpmucdn.com/sites.uab.edu/dist/6/536/files/2024/09/AIs-Trust-Problem.pdf.

Alistair Charlton, "Microsoft Tay AI Returns to Boast of Smoking Weed in Front of Police and Spam 200K Followers," International Business Times, Mar. 30, 2016, at https://www.ibtimes.co.uk/microsoft-tay-ai-returns-boast-smoking-weed-front-police-spam-200k-followers-1552164.

Caitlin Dewey, "Meet Tay, the Creepy-realistic Robot Who Talks Just Like a Teen," Washington Post, Mar. 24, 2016, at https://www.washingtonpost.com/news/the-intersect/wp/2016/03/23/meet-tay-the-creepy-realistic-robot-who-talks-just-like-a-teen/.

Luigi Floridi, The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities (2023).

Nicole Gillespie et al., Trust in Artificial Intelligence: A Global Study (2025), at https://kpmg.com/au/en/home/insights/2025/04/trust-in-ai-global-insights-2025.html).

Reid Hoffman & Greg Beato, Superagency: What Could Possibly Go Right with Our AI Future (2025).

KPMG, Trust in Artificial Intelligence: A Global Study (2025), at https://kpmg.com/au/en/home/insights/2025/04/trust-in-ai-global-insights-2025.html.

Neil Lawrence, The Atomic Human: What Makes Us Unique in the Age of AI (2025).

Peter Lee, "Learning from Tay's Introduction," Official Microsoft Blog, Mar. 25, 2016, at https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

Andrea Loreggia et al., "Modeling and Reasoning with Preferences and Ethical Priorities in AI Systems," in Ethics of Artificial Intelligence 129-154 (S. Matthew Liao, ed. 2020).

Larry Marcus, Taming Silicon Valley: How We Can Ensure That AI Works for Us (2024).

Paul Mason, "The Racist Hijacking of Microsoft's Chatbot Shows How the Internet Teems with Hate," The Guardian, Mar. 29, 2016, at https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism.

Melanie Mitchell, "Why AI Chatbots Lie to Us," Science, Vol. 389, No. 6758 (July 24, 2025), at https://www.science.org/doi/10.1126/science.aea3922?utm_source=substack&utm_medium=email .

Abby Ohlheiser, "Trolls Turned Tay, Microsoft's Fun Millennial AI Bot into a Genocidal Maniac," The Washington Post, Mar. 25, 2016, at https://www.washingtonpost.com/news/the-intersect/wp/2016/03/24/the-internet-turned-tay-microsofts-fun-millennial-ai-bot-into-a-genocidal-maniac/.

Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (2017).

Rob Price, "Microsoft Is Deleting Its AI Chatbot's Incredibly Racist Tweets," Business Insider, Mar. 24, 2016, at https://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3.

Nigel Shadbolt & Roger Hampson, As If Human: Ethics and Artificial Intelligence (2024).

Michael Tomasello, A Natural History of Human Morality (2016).

U.S. Government, America's AI Action Plan (2025), at https://www.ai.gov/action-plan.

Wikipedia, "Tay (chatbot)," at https://en.wikipedia.org/wiki/Tay_(chatbot).

Yongdong Wang, "Your Next New Best Friend Might Be a Robot," Nautilus, Jan. 28, 2016, at https://nautil.us/topics/technology/?_sp=597e0a9c-861e-49f8-a0oc-a6f1794138dc.1753384883261.