El Basilisco de Roko.

Al profesor Héctor Torres Mendoza dedicado.

Luis Eduardo Cortés Riera.

cronistadecarora@gmail.com

Quizás sea la cosa más increíble que he leído en mi ya larga existencia es esto del Basilisco de Roko, noción que descubrí hace poco en Google internet, y que mis hijos adolescentes ya conocían. Pensé que ellos han debido comunicármelo cuando lo descubrieron hace dos años. Algo escandalosamente extraordinario está aconteciendo a escala planetaria con este experimento mental que hubiese agradado mucho al difunto Alan Turing.

La paradoja y la muy riesgosa Inteligencia Artificial (IA) están permanentemente en escena en esta muy curiosa formulación que ha tenido como escenario los ambientes académicos y de investigación del mundo anglosajón y que ha provocado crisis nerviosas depresivas en muchas personas.

El experimento plantea que, en el futuro, una inteligencia artificial con acceso a recursos casi ilimitados desde una perspectiva humana (el *basilisco*) pudiera decidir castigar de manera retroactiva a todos aquellos que de alguna manera no contribuyeron a su creación. Quien escribe ha sido uno de ellos, situación que me ha quitado el sueño durante semanas.

Fue planteado por vez primera este curiosísimo experimento mental en la llamada Comunidad LessWrong que nació en la ciudad de New York en 2009. *LessWrong* se desarrolló a partir de *Overcoming Bias*, un blog anterior del grupo centrado en la racionalidad humana, que comenzó en noviembre de 2006, con

el teórico de la inteligencia artificial amigable Eliezer Yudkowsky y el economista Robin Hanson, creador de una forma de gobierno llamada Futarquía, en la cual oficiales elegidos por votación definen medidas de bienestar nacional, y los mercados de predicción suelen determinar qué políticas tendrán el efecto más positivo.

La conducta altruista, como se ve, motiva a tal Comunidad multinacional en su composición. El altruismo eficaz o altruismo efectivo es una filosofía y movimiento social que aplica la evidencia y la razón para determinar las maneras más eficaces de ayudar a otros. Altruismo significa mejorar las vidas de los demás, a diferencia de egoísmo, que enfatiza solo el interés propio. Eficacia se refiere a hacer el mayor bien posible con los recursos disponibles, así como determinar qué es el mayor bien posible. Se han propuesto reducir la pobreza a escala mundial, atacar el cambio climático y disminuir el sufrimiento animal.

Entre las personalidades destacadas que participan en el movimiento del Altruismo Eficaz destacan el cofundador de PayPal Peter Thiel, el cofundador de Skype el estoniano Jaan Tallinn, el cofundador de Facebook Dustin Moskovitz, y los filósofos como el británico William MacAskill, cofundador de 80.000 Horas, autor de *Haciendo el bien mejor* (2015); el australiano Toby Ord, investigador del Instituto del Futuro de la Humanidad; el australiano Peter Singer, precursor de los derechos animales; y el filósofo alemán Thomas Pogge y su tesis de los deberes negativos de los más ricos para ayudar a los pobres.

El Basilisco de Roko lo ha comparado por Paul-Choudhury a la famosa Apuesta de Pascal (1670) que es un argumento creado por el filósofo francés Blaise Pascal en una discusión sobre la creencia en la existencia de Dios, basado en el pensamiento de que la existencia de Dios es una cuestión de azar, una lotería, pues. Un argumento cargado de matemática.

La premisa del Basilisco de Roko, dice Wikipedia, es el advenimiento hipotético, pero inevitable, de una superinteligencia artificial en el futuro. Esta superinteligencia sería el producto inevitable de la singularidad tecnológica, esto es, el momento en el que una inteligencia artificial creada por la humanidad fuera capaz de auto-mejorarse recursivamente. En el experimento del basilisco de Roko, esta superinteligencia es llamada el *basilisco*.

Es una superinteligencia benévola que sin embargo se ve obligada como imperativo moral kantiano, a castigar a todas aquellas personas e instituciones que retrasaban o impedían su advenimiento, lo que ocasionó serias perturbaciones emotivas y crisis depresivas. La máquina se da cuenta de que para hacer un mayor bien debería haber existido desde mucho antes, para poder ayudar a las personas que sufrían antes de que ella existiera. Una distopía casi increíble.

El basilisco es una furiosa criatura mitológica capaz de matar con la mirada. Así, en su desespero por hacer el bien, la máquina comienza a comportarse como un basilisco que mata a todas las personas que no trabajaron para haberla creado antes, pues eso le impide seguir aumentando su nivel de bondad. Este comportamiento del Basilisco me recuerda al de la máquina HAL 9000, un supercomputador de última generación que enloquece en el film de ciencia ficción *2001, Odisea del espacio* (1968), del director británico Stanley Kubrick.

Dice Dylan Love que "más te vale que ayudes a los robots a hacer del mundo un lugar mejor, porque si los robots descubren que no ayudaste a hacer del mundo un lugar mejor, te matarán por impedirles que hagan del mundo un lugar mejor. Al impedirles que hagan del mundo un lugar mejor, ¡estás impidiendo que el mundo se convierta en un lugar mejor!". El escritor David Auerbach, dice la BBC de Londres, llamó al Basílisco de Roko "el experimento mental más aterrador de todos los tiempos". Nos estamos adentrando en un mundo cada vez más complejo y desconocido, donde la tecnología se desarrolla a un ritmo frenético y amenaza con dejarnos atrás. ¿Estamos preparados para lo que viene? ¿Se están creando hogaño desde la ciencia natural una nueva mitología y una nueva religión? ¿Está la humanidad abriendo una nueva Caja de Pandora?

En mi fuero personal reflexiono honda y nocturnalmente que si el Basilisco de Roko hubiese existido en el año 1981, habría impedido el trágico fallecimiento de mi hermano gemelo Arnoldo Cortés Riera, muerto en mal momento cuando apenas contaba con 30 años de edad.

Santa Rita, Carora.

República Bolivariana de Venezuela.

Sábado 18 de noviembre de 2023.