Riesgos de la concentración en los desarrollos cognitivos

Martín Daguerre

Es profesor y licenciado en Filosofía por la Universidad Nacional de La Plata (UNLP) y doctor en Sociología por la Universidad de Barcelona. Es Profesor Adjunto Ordinario de Ética en la Facultad de Humanidades y Ciencias de la Educación (FaHCE-UNLP) y de Lógica en Facultad de Psicología (UNLP). Es docente en la Maestría en Bioquímica Endocrinológica de la Facultad de Ciencias Exactas (UNLP). Es director del proyecto de investigación titulado "La relación entre emociones y razones desde una perspectiva naturalista metodológica: aspectos fisiológicos, psicológicos y sociales".

Comencemos con una cita para darle un entorno filosófico a lo que quiero discutir, puesto que se trata de un debate que ya se presentaba en la antigüedad griega, cuando Platón se preguntaba: "¿y no corresponde a la parte racional mandar, por el hecho de ser prudente y tener la misión de vigilar el alma entera?" (*República*).[1]

Mientras que para Platón el *auriga* racional conduce nuestra alma, para Aristóteles, en cambio, nuestra capacidad cognitiva racional tiene un lugar subordinado al logro de metas ya establecidas: "no deliberamos sobre los fines, sino sobre los medios que conducen a los fines" (*Ética a Nicómaco*).

Ese mismo debate se reproduce en la modernidad; en este caso, quien ocupa el lugar de Platón es Kant: "el verdadero cometido de la razón ha de ser producir una voluntad buena no acaso como medio en otro respecto, sino en sí misma" (*Fundamentación de la metafísica de las costumbres*). Es decir, la voluntad, para ser una buena voluntad, debe ser guiada por la razón y no por la inclinación. En la posición de Hume encontramos el planteo contrario: "la razón es, y sólo debe ser, esclava de las pasiones" (*Tratado de la naturaleza humana*).

> En este debate sobre el papel de la razón y las pasiones o inclinaciones humanas que pervive desde la antigüedad, me voy a inclinar por Aristóteles y Hume. Me parece que lo que estamos viviendo con el desarrollo de las inteligencias artificiales contribuye a afirmarnos en esa dirección.

Si identificamos la racionalidad con pensar lógicamente, una IA puede desempeñarse racionalmente, de una manera que nosotros no podríamos soñar lograr. Si relajamos la noción de racionalidad, para incorporar análisis estadísticos, nuevamente una IA se puede desempeñar mucho mejor que nosotros en ese terreno.

[1] El presente trabajo no pretende ofrecer una argumentación minuciosa de las tesis defendidas ni una exégesis fina de los filósofos citados, sino ofrecer ideas claras y disparadoras de un debate que considero indispensable.

Si uno no quiere limitar la racionalidad al plano de las decisiones individuales, y la piensa, entonces, vinculada a las decisiones estratégicas, como en el ajedrez, de nuevo, una IA puede ser más racional que nosotros, como demuestra el hecho de que ya nadie pueda ganarle en juegos de este tipo. Si queremos dar un paso más y adentrarnos en decisiones estratégicas en el marco de un juego no limitado a la aplicación de ciertas reglas, como puede ser el póker, volveremos a encontrarnos con IAs que nos superan también en ese terreno.

De manera que, identifiquemos con lo que identifiquemos a la racionalidad, es claro que, si no hoy, en breve, las inteligencias artificiales van a lograr mayor racionalidad que la que podamos alcanzar nosotros.

Para un enfoque racionalista, entonces, deberíamos ver la manera de poner esta racionalidad tan lograda al mando de nuestros asuntos. Sin embargo, parece que el desarrollo de IAs ha generado más temor que esperanza. El desarrollo de la racionalidad, en nuestro contexto, sería fuente de nuevos y más dramáticos problemas, antes que solución de los que teníamos.

Y esto nos trae a lo que se llama, en inteligencia artificial, el problema de la alineación de valores. Lo que ha vuelto a las inteligencias artificiales tan competentes es que aprenden (con *deep learning*, con aprendizaje por refuerzo). Cuando uno genera una IA, tiene que establecer qué tiene valor y qué no, porque esa será la meta que dirija el aprendizaje de la IA. Por ejemplo, si quiero que la IA aprenda a jugar bien al ajedrez, le asignaré un puntaje positivo al logro de llegar a un jaque mate, y un puntaje negativo a sufrir un jaque mate; luego, tendrá un puntaje positivo, aunque menor, aquella jugada previa a hacer jaque mate, y un puntaje negativo, aunque menor, aquella jugada previa a sufrir un jaque mate; y así sucesivamente, de manera que el objetivo de la IA sea lograr el puntaje positivo más alto. De esta manera, a medida que la IA aprenda, siempre seleccionará la jugada que maximice el valor positivo.

Planteada la meta en estos términos, podemos encontrarnos con un problema: la IA no buscará la partida más directa que lleve al triunfo, sino la que otorgue más puntos, por lo que quizá seleccione jugadas que lleven al triunfo, pero de la manera más larga posible, para sumar la mayor cantidad de puntos posibles.

Stuart Jonathan Russell y Peter Norvig ofrecen otro ejemplo muy claro.[2] Supongamos que intento programar una IA para una aspiradora. Puede que piense que la meta obvia que debo asignarle es aspirar la mayor cantidad de polvo posible. Pero esto daría lugar a que la aspiradora junte polvo y lo vuelva a tirar, para tener más polvo que juntar y aumentar, entonces, la cantidad de polvo aspirada.

Este es el problema de la alineación de valores: cómo lograr que la IA haga lo que efectivamente deseamos que haga. Dada la eficiencia de una IA, cualquier error en la especificación del valor, de la meta última, puede tener como resultado un despegue marcado de los efectos secundarios no deseados.

Pero desde el punto de vista del usuario de la IA, ya no de quien la diseña, la situación es más preocupante, porque siempre desconoce cuál es la meta de la IA con la que trabaja. Es cierto que, al utilizar una IA, vamos encontrando lo que buscamos, pero también es cierto que nuestra interacción con la misma revelará un conjunto de información sobre nosotros que, muy probablemente, no querríamos revelar y que, por lo demás, puede ser lo que persigue conocer el propietario de la IA. Ya navegando por internet estamos ofreciendo un cúmulo de información que, leída por los programas informáticos correspondientes, sale fácilmente a la luz (incluso si no hemos visitado páginas relacionadas con condiciones propias que no queremos que se descubran): si estamos emba-

[2] Russell, S. y Norvig, P. *Artificial Intelligence. A Modern Approach.* (Cuarta edición), Pearson, 2021.

razadas, si somos hetero u homosexuales, si padecemos estrés postraumático, si nuestro sueldo es bajo, si hemos perdido a un ser querido recientemente, etc. Estas condiciones, luego, permiten clasificarnos como potenciales consumidores de distintos productos, información que es vendida luego a quienes comercian esos productos y saben cómo explotar tales condiciones.

Una universidad privada puede pagar equis cantidad de dólares por un listado de personas que padezcan estrés postraumático. Luego llamará a esas personas, les contará sobre todas las oportunidades que se les abrirán por estudiar en esa universidad y les ofrecerá hacer todo el papelerío para que logren un crédito estatal para estudiar allí. Dada la condición vulnerable de quienes reciben la llamada, posiblemente se logre que accedan, con lo cual, si luego ocurre lo más probable, esto es, que no terminen sus estudios por los problemas que arrastran, la universidad terminará con dinero en sus arcas y ellas con una deuda impagable.[3]

Hoy por hoy, los críticos suelen centrarse en los errores de la IA, en sus alucinaciones, en los sesgos que tiene, en la opacidad, es decir, en el desconocimiento de cómo el algoritmo llegó a equis resultado. Sin embargo, cabe esperar que estos problemas sean paulatinamente superados. El problema realmente importante es que nosotros nunca sabemos cuál es la meta (y, estrechamente vinculado con ello, qué es lo que es considerado un costo) que se le asigna a la IA.

Volviendo sobre nuestro punto de partida, los desarrollos de la IA pueden verse como desarrollos de la racionalidad. Ahora bien, como la racionalidad es esclava de las pasiones, que son las que otorgan valor a algo, una IA puede estar al servicio de nuestras metas o, todo lo contrario.

[3] Encontramos muchos ejemplos de este tipo en el libro de Cathy O´Neil, *Weapons of Math Destruction*, Crown, 2017.

Analicemos ahora, brevemente, qué quiere decir que las pasiones son las que otorgan valor a algo. Dada nuestra naturaleza, estamos "programados" para valorar ciertas cosas: cierto tipo de alimentación, ciertas temperaturas, cierto tipo de vínculos sociales, etc. Para sobrevivir y reproducirnos debemos mantener un conjunto de variables dentro de ciertos rangos, y lo logramos por medio de procesos homeostáticos. Nuestra capacidad cognitiva, consciente, puede contribuir a la consecución más eficiente de esta meta, pero no pone la meta. Como sostiene el neurocientífico Antonio Damasio, nuestros procesos conscientes dependen de la corteza cerebral dedicada a mapear los estados corporales.

La teoría polivagal del psicólogo Stephen Porges va en el mismo sentido, y ahora quiero detenerme un poco en ella. En este caso, veremos de qué manera nuestro comportamiento está en función de lo que ocurre en nuestro sistema nervioso autónomo.

Según Porges, el sistema nervioso autónomo se puede dividir en tres partes: el sistema *vagal* ventral, el sistema simpático y el sistema *vagal* dorsal. El último es el más antiguo en términos evolutivos. El primero es el más reciente y propio de mamíferos como nosotros, que dependemos de largos procesos de aprendizaje, a diferencia de lo que ocurre, por ejemplo, con un reptil. El hecho de nacer con una gran plasticidad neuronal, fundamental para incorporar conocimientos del entorno, implica, por otra parte, la necesidad del cerebro de utilizar mucha energía para tal tarea.

El sistema nervioso simpático es aquel que dispara las conductas típicas de estrés (que consumen, también, mucha energía), de lucha y huida, que muchos animales tienen ante la presencia de otros animales. En animales cooperativos, como nosotros, hace falta que el complejo *vagal* ventral inhiba la activación del sistema nervioso simpático, relaje al individuo, y éste pueda, entonces, utilizar mucha energía para el logro de un desarrollo normal.

¿Cómo el complejo *vagal* ventral pone freno al sistema nervioso simpático? Cuando una madre habla en cierto tono o cuando presenta una mirada amable o cuando acaricia a su bebé, genera *inputs* que ponen freno a la potencial amenaza que el bebé podría sentir frente a otro individuo, esto es, pone freno a la activación del sistema nervioso simpático. El *output* podrá ser una mirada acorde a la de la madre, una sonrisa, etc. Todo ello deriva en un vínculo relajado que permite al bebé dedicar gran parte de su energía a consolidar aprendizajes.

Ahora, supongamos que el bebé no está relajado ni protegido, porque está en un orfanato. La ausencia de los *inputs* que recién destacamos llevará a que esté activo su sistema simpático de lucha y huida. Podremos encontrarlo gritando, llorando, que es una manera de luchar por lograr comida, abrigo, etc. Supongamos que nadie le hace caso, ahí pasamos a lo que se llama el complejo vagal dorsal, el sistema de inmovilización. El cuerpo entiende que ya no puede gastar más energía, de manera que para sobrevivir pasará a un estado de aparente calma, de desconexión con la situación, comparable a la reacción de una comadreja que queda inmovilizada cuando ya no puede escapar del peligro. Si bien esta estrategia del bebé puede ser inteligente, dada su situación, lo cierto es que como consecuencia sufrirá un deterioro cognitivo porque también evitará consumir energía para nutrir a su cerebro.

Con todo esto quiero destacar lo siguiente. En principio, valoramos cierto tipo de relaciones no estresantes, de contención. En ese entorno nos desarrollamos de manera óptima, lo cual se refleja, a nivel consciente, como un sentimiento de bienestar. Si el entorno es peligroso, amenazante, abandonaremos las disposiciones propias para consolidar lazos solidarios, y nos dispondremos a competir o huir, en función de nuestras posibilidades. Si bien ya no se dará el correlato consciente de bienestar, no deja de ser cierto que nos sentiremos mejor si superamos el peligro, que si terminamos sometidos de alguna manera. En última instancia, si nuestra lucha o huida no resulta exitosa de ninguna manera, nos veremos condenados a una muy precaria existencia, con mínimas probabilidades de recuperación.

Ahora quiero especular un poco sobre un correlato social del vínculo entre madre e hijo que presenté hasta aquí. Hay dos modos en los que las personas se representan a la sociedad, y para verlos con claridad pueden analogarse a dos prácticas deportivas diferentes: la sociedad es el equivalente social de un equipo de básquet, o es el marco en el que debemos correr una carrera de natación.

Cuando se piensa como un equipo de básquet, se supone que el éxito depende de la actuación colectiva, de hacer lo mejor posible con las capacidades de los miembros. Si uno de los jugadores carece de todo talento, se elaborará la mejor estrategia posible, de manera tal que su falta de capacidad tenga el menor impacto posible sobre las probabilidades de éxito. Toda dificultad individual requiere soluciones colectivas, y el éxito y la derrota son del colectivo, no del individuo.

En cambio, si pensamos a la sociedad como el entorno en el que correremos una carrera de natación, entenderemos el éxito y el fracaso como dependientes de la actuación individual. En este caso, las dificultades individuales requieren soluciones individuales, y toda dificultad de los demás será una ventaja para uno, no algo que requiera soluciones colectivas.

Quienes quieren vivir en una sociedad "basquetbolista" entienden que la sociedad nos debe resguardar de sufrir las consecuencias más perjudiciales de poseer cualquier tipo de vulnerabilidad. Independientemente de la edad, el estado físico, la capacidad cognitiva, etc., todos forman parte del equipo, y una sociedad exitosa es aquella en la que todos cuentan con las condiciones para vivir bien, más allá de cuál sea su aporte individual.

Quienes quieren vivir en una sociedad "de nadadores" valoran el éxito individual, que depende de ser mejores que los demás, razón por la cual carece de sentido verse obligado a hacerse cargo de los problemas de otros.

¿Cómo se resuelve la convivencia de basquetbolistas y nadadores? Supongamos, primero, que nos encontramos en una sociedad regida por las normas "basquetbolistas". ¿Cuál sería el comportamiento del "nadador"? Lo que uno puede esperar es que quiera destacar individualmente, obtener algún tipo de privilegio por su mejor desempeño, etc. Pero posiblemente este tipo de conducta no tendrá buena recepción entre los "basquetbolistas", quienes tomarán medidas para evitar que este comportamiento individualista afecte las perspectivas de éxito del equipo. Tal podría decirse que es lo que ocurría durante nuestra vida como cazadores recolectores. De los últimos 100.000 años de nuestra existencia, alrededor del 90% lo vivimos de esta manera. Cada uno dependía del colectivo, y quien pretendía de alguna manera erigirse en dominador corría el riesgo de ganarse un rechazo generalizado, ser expulsado del colectivo y verse condenado a la muerte en soledad.

Posiblemente en ese período esté el origen de nuestras emociones morales, prosociales, que dieron lugar a que el bienestar de uno no está disociado del bienestar del otro.

Ahora bien, con el surgimiento de la agricultura y la consiguiente generación de excedentes acumulables, estrategias de carácter más egoísta pudieron proliferar, porque se volvió posible no depender de la cooperación social, en la medida en que se contara con la capacidad para adueñarse de la riqueza.

Si la sociedad es una sociedad de "nadadores", ¿cómo reaccionará un basquetbolista en ese entorno? Por un lado, dadas las condiciones, se verá obligado a "nadar" lo mejor que pueda, a desarrollar las capacidades que le den una ventaja; por otro, sentirá que debe solidarizarse con quienes no triunfan, ya sea enseñándoles a nadar mejor, o pugnando por que todos reciban una medalla por participar.

Y acá es donde vienen los problemas de los desarrollos que se concentran exclusivamente en el plano racional. Tomemos a los científicos que trabajan en la edición genética con CRSPR-Cas9 o que hacen desarrollos dentro de la optogenética. Si son "nadadores", los incentivos son grandes, porque cualquier éxito en este terreno promete grandes premios individuales. Si, en cambio, son "basquetbolistas", aun así, tendrán grandes incentivos en lograr desarrollos, en la medida en que los mismos pueden ser aplicables al tratamiento de diversas enfermedades. Pero más allá de que todos estén incentivados para desarrollar estas potentísimas tecnologías, lo relevante es saber para qué fines, efectivamente, se van a utilizar.

El "basquetbolista" que, en una sociedad "nadadora", enseña a nadar a los que menos saben, ¿está contribuyendo a superar la sociedad nadadora en dirección de una sociedad basquetbolista, o está contribuyendo a que la competencia entre nadadores sea más pareja, pero sin dejar atrás a la sociedad nadadora? El científico "basquetbolista", preocupado por quienes padecen ciertas enfermedades, ¿contribuirá con sus desarrollos al logro de esa sociedad en la que quienes padecen enfermedades siempre contarán con un respaldo social, o dará herramientas más poderosas a quienes ya venían ganando la carrera?

Lo mismo podemos pensar con respecto a los desarrollos en IA. Claramente puede pensarse en fines que nos parecen dignos y que pueden alcanzarse mediante una IA. Un algoritmo puede localizar a personas altamente vulnerables, lo cual nos ayudaría a centralizar la ayuda donde realmente es necesaria. Pero en el contexto dominado por perfiles "nadadores" no será ese el destino

que tendrá todo desarrollo racional. Recordemos que la razón no puede sino ser esclava de las pasiones, y las pasiones dominantes parecen ser las individualistas, de manera que todo nuevo desarrollo se vuelve enormemente peligroso para los "basquetbolistas".

Dicho esto, puedo plantear lo que me preocupa de enfoques como los de Christopher Osterhaus. Dado que partimos de una sociedad competitiva, ¿el desarrollo cognitivo contribuirá únicamente a mejorar las capacidades para la competencia? Si logramos que todos tengan una teoría de la mente bien desarrollada, habremos logrado personas capaces de detectar en qué lugar de la jerarquía social están, quién es el alfa del grupo, cómo ser más eficiente para engañar, etc. ¿En qué medida todo esto mejorará la sociedad? Quizá lo haga en el sentido de hacer la competencia más equitativa ¿pero es ese nuestro objetivo, en el caso de que seamos "basquetbolistas"?

Parece que lo que necesitamos, previamente, es una mejor visión de nuestra meta. Lo que busca un basquetbolista es una sociedad de cuidado mutuo, en donde el complejo *vagal* ventral logra poner freno al sistema nervioso simpático, quiere rodearse de relaciones de reciprocidad positiva, mantenerse relajado y gozar de los logros que puedan darse dentro de ese entorno. Para un basquetbolista es prioritario que no exista el *bullying*, a que se tenga un buen dominio en lengua y matemática.

En nuestras sociedades de nadadores lo que está fuertemente activo es nuestro sistema nervioso simpático. Ello nos lleva a pensar que lo que debemos hacer es luchar o huir. Quienes logran ganar, pueden activar el complejo *vagal* ventral para poner freno al sistema nervioso simpático. Pero no deben confundirse, pensando que de lo que se trata es de darles a todos las herramientas que les permitieron a ellos ganar, y mucho menos generar herramientas más potentes. Por definición, en una sociedad de nadadores ganan los mejores, sea lo que sea lo que se considere mejor, pero nunca pueden ganar todos. Los basquetbolistas ganadores en el juego de la natación deben abogar por cambiar de deporte, para lo cual se requiere, ante todo, claridad en las metas. Sin esta claridad, los desarrollos cognitivos tenderán a ser usados para agravar el problema, antes que para resolverlo.