TAREA 50

Aclaración terminológica: Por **productos AI** habrá de entenderse cualquier producto que incluya cualquier desarrollo de inteligencia artificial, sea material (un electrodoméstico inteligente) o virtual (servicios en la nube que incluyan AI). Por **agente AI** nos referiremos a un producto con un nivel de inteligencia superior a un producto AI, por ejemplo un agente AI sería un vehículo autónomo o un robot colaborativo.

¿Qué es ético y qué no?
En aras de la operatividad y para no enredarnos, diferenciaremos entre la ética en cuanto a reflexión sobre el bien y el mal y en cuanto a corpus de valores. Este corpus valorativo se sustancia en un sistema de normas que rigen el comportamiento humano y social. Una es a priori, el otro una consecuencia de la primera. Las normas, es cierto, se han ido destilando en base a una tradición configurada en el roce constante con la cuestión religiosa (o humanista) del bien y el mal, sin embargo, a efectos de lo que nos ocupa, no atenderemos a semejante cuestión, será nuestro objeto de estudio la ética en tanto sistema convencional o axiológico que se sustancia en un cuerpo de normas. Estas normas son las que rigen la valoración, calificación y gestión del comportamiento humano en sociedad; lo que esté dentro de ese corpus de normas es ético; el resto no lo sería.

¿Son los datos justos? ¿Tienen sesgos? ¿Discriminan?
Algoritmos y datos, eso es AI. Los datos, en sí mismos, sin un contexto de tratamiento, son neutros; sólo cuando se ponen en relación, convirtiéndose en información, es cuando pueden ser manipulados. La manipulación puede darse en el algoritmo de razonamiento pero es posible que los datos también sean manipulados, falseados desde su origen por un sesgo en la captura de los mismos, por ejemplo.

¿Cómo hacemos para que los resultados de la AI sean explicables y dejen de ser cajas negras?
¿Porqué pedimos transparencia, inmediatez, a la AI cuando la propia inteligencia natural sigue siendo en gran medida opaca y una gran desconocida?. La teoría de las inteligencias múltiples ha venido a hacer justicia atendiendo a aspectos de la inteligencia no reconocidos y al mismo tiempo a complicar aún más una definición global del concepto de inteligencia.
Pero, ¿acaso el término coloquial de “inteligencia” nos es inmediato? Por poner un ejemplo, la inteligencia que rige el debate político o la comunidad científica no nos es inmediata ni transparente. No se muestra como algo evidente a seres humanos que disponemos de forma natural de inteligencia, por ello, ¿no es acaso el conocimiento científico una caja negra para la sociedad?. El algoritmo no es una caja negra para el que lo diseña, codifica o entiende.

¿Cómo analizamos los parámetros/datos más influyentes? ¿Y su sensibilidad?
Parametrizar ya asume una toma de postura, determinar lo que resulta influyente y lo que no es influyente, otro tanto. Ajustar la sensibilidad es jugar en la frontera de lo

aceptable y lo no aceptable, esto es, tratar con valoraciones éticas. Me remito pues al punto siguiente.

¿Cuáles son los principios fundamentales que ninguna AI debería vulnerar?
En un plano general, yo referiría *ánimo de veracidad* y *universalidad*. Es decir, que la AI tiene que atenerse, de alguna forma, a lo que se conoce como "verdad" en ciencia, aunque éste sea un concepto difuso, tan cuestionado como cuestionable. Por otra parte, entiendo que la universalidad de los desarrollos y de la aplicabilidad de la AI es inevitable en tanto que es ineludible en un mundo globalizado.

Por otra parte, en la respuesta a esta cuestión, veo necesario contar con alguna respuesta a una pregunta previa: *¿Es la AI una ciencia?*. De la respuesta a ella se derivan no pocas cuestiones asociadas, algunas de las cuales atañen a la cuestión de la ética que estamos tratando en esta tarea 50. Suponiendo que ambas, ciencia y AI, no caminan lejos una de otra y que lo harán juntas en los próximos tiempos, podemos decir algo más sobre AI y principios fundamentales. R. K. Merton adelantó, desde su estudio sobre la sociología de la ciencia, unos principios reguladores mínimos para la actividad científica que podrían ser traspuestos a los de la AI. Los principios enunciados por Merton son: **Universalismo**, **comunitarismo**, **desinterés** y **escepticismo organizado.** Así, universalismo sería el principio por el cual la AI no se constituye como desarrollos propietarios ni exclusivos de una sociedad o país concreto. Comunitarismo, se refiere a que el desarrollo de la AI se nutre de diversas fuentes y se dinamiza por la participación de la comunidad científica en un sentido de propiedad común al objeto de promover la colaboración mutua de los miembros de la comunidad científica. El secretismo es lo opuesto a este principio. Desinterés en cuanto a que la inteligencia está por encima del beneficio crematístico que pueda devenir de ella; que es puro ánimo de saber más allá de los beneficios que pueda entrañar el saber o la aplicación del saber. Finalmente el escepticismo organizado se refiere a la necesidad de una permanente actitud critica ante desarrollos, usos y derivas que la AI pueda tener en su evolución.

En definitiva, respecto a AI, podríamos afiliarnos a los mismos principios de racionalidad científica de Merton; más allá de eso, no creo que se pueda referir otra cosa que no sean los límites propios de las distintas ciencias particulares (sea médica, física,…). Es decir, podemos tratar la ética de la AI sólo en aspectos generales ya que en lo particular, la ética en la AI se debería circunscribir a cuestiones éticas propias de los códigos deontológicos respectivos de la disciplina concreta a la que sirve.

¿Deberían homologarse los desarrollos de AI?
En torno a AI solo se halla información en la web respecto a *homologación en estándares de la industria*; más concreta y casi exclusivamente, referidos a la homologación de la industria de la automoción y otras máquinas autónomas. No he encontrado nada referido a otros conceptos como *homologación científica* y *homologación ética* de los desarrollos de la AI, conceptos que pueden ser pertinentes en este debate.

Respecto al primero, es decir, al industrial, es necesario contar con estándares y homologación de parámetros y criterios de configuración y validación de componentes y subsistemas clave para diseñar productos y agentes AI seguros. Como es fácilmente comprensible, en un mercado global, es clave que el marco de seguridad de los desarrollos AI para vehículos autónomos o robots colaborativos sean los mismos independientemente del país en el que se diseñen. De forma similar, es necesario que el diseño de agentes AI (robots, vehículos…) cuente con especificaciones concretas y al mismo tiempo flexibles a las que puedan ser adaptadas en el futuro el desarrollo de nuevos componentes de automoción o elementos viales (carreteras y señalización inteligentes, onboard information, etc,…)

Respecto al segundo término se podrá hablar de que el conocimiento científico, la ciencia, es homologable y homologado en gran medida gracias a un marco de referencia que dentro de la comunidad científica se conoce con el nombre de *paradigma*. No es menos cierto que dentro de un paradigma conviven teorías diferentes que en algún momento pueden socavar dicho paradigma dominante, pero dichos marcos siempre incluyen algún tipo de homologación de metodologías, supuestos axiomáticos, etc. (ver T. S. Kuhn).

En tercer lugar diremos algo sobre homologación respecto a la ética. Las cuestiones éticas se materializan de forma diferente en diferentes culturas y tradiciones, por ello es fácil imaginar que cuestiones de ética en relación a una AI global no encuentren, en principio, un acomodo inmediato en culturas con principios éticos distintos.

Por otra parte, cualquier tipo de limitación al conocimiento que viniera de consideraciones éticas parece que estaría abocada al fracaso (las religiones lo intentaron de manera ingenua). El ánimo de saber, la curiosidad del ser humano, son difícilmente restringibles. Si, además, damos por sentado que la inteligencia es una capacidad adaptativa de ciertos animales con sistema nervioso superior, que es un sistema emergente, tenemos que, en principio, no resultará fácil ponerle límites al desarrollo de la inteligencia y al conocimiento. La AI así no aceptaría desarrollos parciales, ni límites impuestos, la AI sería así su propia potencialidad en desarrollo.

Pero por otra parte, aun asumiendo la pertinencia de lo dicho anteriormente, es cierto que no podemos pasar por alto los interrogantes que surgen a nivel global ante el desarrollo de la AI hoy en día. Prueba de ese interés son las acciones que ya están tomando las instituciones a nivel mundial en pos de regular el desarrollo de una tecnología disruptiva como la AI. Así, el ENAI, marco regulador para el estado español, tiene como objetivo *"establecer un marco ético y normativo que refuerce la protección de los derechos individuales y colectivos,"* y que logre *"generar un entorno de confianza respecto al desarrollo de una Inteligencia Artificial"* Dicha confianza es necesaria para la industria, pero también para la ciudadanía, para los usuarios.

Efectivamente, gran parte de la cuestión, si no toda, se dirime en la línea fronteriza entre derechos individuales y colectivos. No es algo nuevo, se ha dado en otras áreas, pero adquiere radical actualidad en el contexto de pandemia y las llamadas a priorizar lo colectivo (la inmunidad de grupo por medio de la vacunación, asepsia y

distancia social) frente a las libertades del individuo (derecho a no ser vacunado por ejemplo). Una cuestión aún no resuelta en otras áreas de la ciencias sociales, aflora, en el caso de las máquinas y a los algoritmos en la AI.

Las cuestiones éticas despuntan en forma de dilemas refractarios a cualquier intento de resolución. Por ejemplo el dilema del tranvía[1], el cual podemos poner en relación a la AI, concretamente un agente AI del tipo *coche autónomo*[2] que se está desarrollando en la actualidad. Una inteligencia artificial, no humana, puede decidir "sacrificar" una vida para salvar otras sin tener que acarrear con el peso de la conciencia de haber sacrificado una vida, algo que con toda certeza sí afectaría a un ser humano. El coche autónomo contará en breve con algún tipo de algoritmo que le instruya a través de una premisa del tipo *"la vida de un ser humano no prevalece sobre la de dos o más humanos"* y actuar en consecuencia. Dicho agente AI autónomo, valoraría, en su caso, un escenario concreto y, aplicando tal premisa, movería el vehículo para esquivar al grupo de dos o más personas y arrollar a una sola persona. Pero ¿y si la situación de riesgo no es de un *sí* o un *no* evidentes?. Si el agente AI (vehículo autónomo) tiene en su base de datos de premisas una del tipo *"la vida de un ser humano no prevalece sobre la posibilidad de salvar a dos o más humanos"*, entonces el agente autónomo AI habría de valorar qué probabilidades hay de salvar a más de uno y actuar en consecuencia de esa probabilidad. No sería algo materialmente imposible, pero sí habría de hacer unos cuantos cálculos en un lapso de milésimas de segundo; el tiempo justo preciso en las reacciones en casos de accidente de automóvil. Ese rango de probabilidad acaptable habría de venir dado previamente por los creadores del agente autónomo de AI.

Bien, hasta aquí hemos dibujado un escenario de situaciones en las que la AI ya ha de vérselas con la ética. La situación se complica aún más cuando, de forma previsible, la AI tendrá que responder a situaciones no previstas y decidir sobre situaciones nuevas en las que se pueda dar algún dilema ético. Para estos casos es por lo que se requiere una AI con capacidad de aprendizaje. Este es un ámbito de la AI que inicia sus pasos de forma paralela al de la formalización de los denominados contrafácticos[3] (cuestiones de la forma de *"Que habría pasado si hubiéramos actuado diferentemente en el caso…"* y las respuestas que se den a esas preguntas). El tema de los contrafácticos se da en el ámbito de la causalidad, largamente debatido y muy controvertido en filosofía y ciencia. La causalidad como estructura del mundo y de los hechos del mundo ha sido cuestión recurrente que después de siglos de ostracismo puede tener una segunda vida con el despegue de la AI.

Por último diré algo respecto a problemas éticos que pudieran colegirse del posible hecho de que interactuemos con agentes AI sin nosotros saberlo. En la presentación

---

[1] Dilema del Tranvía': imagina que ves un tren sin frenos acercándose peligrosamente a un grupo de cinco personas. Tienes a tu lado una palanca para desviarlo y que tome otra vía, en la que es seguro que atropellaría mortalmente a una sola persona. La mayoría de las personas accionaría la palanca. Pero, ¿y si el único modo de evitar la tragedia fuera que empujaras a un hombre desde un puente, que éste cayera sobre la vía y frenar así al tren? Las decisiones que implican vidas humanas nunca son fáciles de tomar.

[2] A pesar de haberse enfriado las altas expectativas iniciales, el futuro del coche autónomo de Waymo-Google, Tesla, etc continuará siendo un factor dinamizador de la AI en el futuro.

[3] Ver Pearl y Mackenzie en "The Book of Why"

que Google realizó de su motor de AI, se muestra cómo un agente AI realiza por teléfono una reserva de cita en una peluquería sin levantar sospechas de su identidad en la encargada de peluquería. Habrá que estudiar las posibles derivadas que un hecho de este tipo pudiera tener pero seamos conscientes de que los nuevos medios de comunicación ya plantean cuestiones problemáticas en este aspecto, caso paradigmático las *fake news* y cómo consiguen modelar la realidad social en red y el impacto que tienen en la realidad social *real*. Sin duda debemos considerar esa posibilidad pero, ¿acaso los principios éticos no son volteados o retorcidos por la manipulación mediática, el debate político, la ingeniería del consenso o la realidad virtual?. Tal vez no sea descabellado decir que nos será más fácil, en su momento, defendernos de la posible manipulación de los agente AI que de la manipulación realizada por mentes humanas.

Para finalizar, hagamos un poco de ciencia ficción. Si la AI llegara a hacerse autónoma (aprendizaje desatendido, auto-programación del aprendizaje y cosas así) nos pondría ante las puertas de un conocimiento de corte no humano, *conocimiento desantropologizado o* inteligencia no humana, con las consecuencias que ello tendría a todos los niveles. ¿podríamos hablar de la socialización de los robots?, ¿de cognición social de agentes AI autónomos?, En este aspecto son particularmente sugestivas las obras del escritor Stanislav Lem.

¿Deberían pasar por una revisión para validar su certificación?

Estaría relacionado con la cuestión anterior. Sería deseable y tal vez ineludible en cuestiones de homologación técnica y científica y, si acaso, deseables, aunque no exenta de dificultades, en la cuestión de la homologación ética. La posible validación yo lo situaría en paralelo a los requerimientos éticos en las ciencias: validación inter pares dentro de la comunidad científica que certifique la atención a la veracidad, exactitud y no tergiversación de los algoritmos y de captura y disposición de los datos, etc.

¿Qué papel juegan la sociedad y los gobiernos en la definición de estos principios?

La cuestión de lo público y lo privado se extiende en paralelo a la cuestión de la tradicional distinción entre lo natural y lo artificial y se vuelve crecientemente conflictiva. El peso que el binomio público-privado tiene hoy en día en la sociedad y su evolución es central en la definición de nuestro futuro. Es clave como hemos visto en la cuestión ética y también en la cuestión del propio desarrollo sostenible de la AI y lo viene siendo desde décadas ya en la propia cuestión de la ciencia y su modelo de desarrollo vigente, la tecnociencia. La privatización progresiva de la ciencia augura una privatización de los desarrollo tecnocientífico asociados a la AI. Lo público, a través de los pocos instrumentos que les son propios (gobiernos y la sociedad civil) habrán de plantar batalla en un contexto en el que no es difícil anticipar la visión de una nueva e inquietante distopía.