Somos IA n.º 3: ¿Quién vive, quién muere, y quién decide?

Cubierta-alt.

Una enorme lata dorada: la tapa de la lata está abierta hacia arriba y sobre ella se proyecta la escena del clásico dilema del tranvía, con un tranvía acercándose a una bifurcación en la vía. En una de las rutas hay una persona atada a la vía, mientras que en la otra hay un grupo de 4 personas atadas a la vía. Hay un hombre parado en la intersección que controla con una palanca por qué vía se encaminará el tranvía. La lata está desbordada de gusanos rosados, algunos de los cuales han caído al suelo. Los gusanos tienen patrones abstractos en ellos y las palabras “desigualdad”, “valores”, “utilitarismo”, “incertidumbre” y “ética” escritas en ellos. Simboliza la idea de que el problema del tranvía abre una lata de gusanos de problemas éticos en la IA.

Citar como:
Julia Stoyanovich, Mona Sloane y Falaah Arif Khan.
¿Quién vive, quién muere, quién decide? We are AI Cómics, Vol. 3 (2021)
http://r-ai.co/comics

Página 1

Predecir es muy difícil, especialmente si es sobre el futuro.

Por más que sea difícil —debido a lo incierto y complejo que es el mundo— predecir el futuro es habitualmente el trabajo de la IA.

Un robot de IA con una bola de cristal y las manos ondeando alrededor de la bola; su expresión es seria y tiene las cejas levantadas.

Y debido a la dificultad de la tarea, que a veces es incluso imposible, los sistemas de IA cometen errores.

Por ejemplo, una IA de luz inteligente puede no acertar sobre si una luz debe estar encendida o apagada.

Encender la luz en medio de la noche nos despertará.

La mujer que estaba durmiendo se despierta abruptamente a causa de las luces. Tiene ojos de zombi, el cabello despeinado y los brazos levantados, y la expresión en su rostro refleja su exasperación.

Y dejar las luces encendidas sistemáticamente conlleva pagar una factura de energía más alta.
Una mujer tiene una factura de electricidad muy extensa en una mano y se asusta al ver la estimación. Tiene la otra mano en la cabeza en actitud de asombro y parece angustiada.

Otro ejemplo: una IA de servicio al cliente en tu zapatería favorita podría malinterpretar tu pedido, ...y tú recibirías un par de zapatos equivocados.

Vemos una conversación entre una IA de atención al cliente y un cliente. La IA de atención al cliente está usando auriculares y hablando con el micrófono, y una mujer sostiene un par de zapatos de colores y patrones muy vistosos, mientras frunce el ceño.

Estos errores pueden ser irritantes, pero plantean un escaso riesgo.

Las consecuencias de dichos errores no son graves y son reversibles.

Página 2

Sin embargo, hay casos en los que los errores pueden provocar daños irreversibles y catastróficos, incluso la pérdida de vidas humanas.

Consideremos un automóvil autónomo:
Una mujer se echa hacia atrás en un vehículo autónomo: está recostada en el asiento del conductor con las piernas en el tablero, mientras el volante se ilumina y el auto entra en modo de autoconducción.

Una IA está a punto de cruzar una intersección, y no reconoce a una persona en una bicicleta como uno de los tipos de objetos que esperaría ver en la carretera.

El vehículo no se detiene y atropella a la ciclista.

Vemos el punto de vista del conductor desde el auto. Hay dos autos cruzando la calle, árboles en la acera y una ciclista que cruza al frente. Un algoritmo de segmentación de imágenes identifica diferentes objetos en la escena: ha detectado el árbol y los automóviles, y los marca con cuadros de posición. No ha detectado a la ciclista como objeto en la carretera.

Otro ejemplo es cuando el vehículo autónomo no detecta la presencia de una persona en silla de ruedas cruzando la intersección.

Esto podría suceder si, por ejemplo, la persona cruza la intersección yendo hacia atrás,
de manera que la IA del coche autónomo no calcule bien la trayectoria del peatón.

Una mujer en silla de ruedas cruzando una intersección. El semáforo está en rojo, y la mujer gira la silla de ruedas y cruza yendo hacia atrás.

Pero los conductores humanos también causan accidentes.

Así que, ¿por qué dejar que lo perfecto sea el enemigo de lo bueno?

Un Elon Musk caricaturizado se encoge de hombros y apunta a la derecha de la pantalla, defendiendo los autos sin conductor.

¿No deberíamos estar preparados para sufrir algunos errores cometidos por automóviles autónomos en aras de una mayor seguridad general de nuestro sistema de transporte y la conveniencia de los conductores?

Página 3

De hecho, ¿no podríamos codificar nuestro criterio sobre los errores más importantes a evitar, y dejar que una IA resuelva los casos dudosos?

¿No podemos equipar nuestra IA con esos valores?

Un ejemplo famoso que nos hace pensar en nuestros valores, y las compensaciones que introducen, es
El problema del tranvía.

Es un experimento sobre el pensamiento que plantea un dilema ético:
¿Deberíamos sacrificar la vida de una sola persona para salvar la vida de un grupo de personas?

Curiosamente, los experimentos en ética y psicología han demostrado que no hay una respuesta clara.

Lo que decidimos depende de nuestros valores: de lo que consideramos correcto o incorrecto,
de los diversos elementos de nuestra identidad, de nuestro contexto cultural, y también de la configuración específica del problema (el contexto en el que se toma la decisión).

Imagen de una extensa red de tranvías y vías de intersección. De arriba hacia abajo: un tranvía se acerca a una bifurcación. En la ruta superior hay una

persona atada a la vía, mientras que en la ruta inferior hay un grupo de 4 personas atadas a la vía. Hay un hombre parado en la intersección que controla con una palanca por qué vía se encaminará el tranvía. La ruta inferior luego conduce a otra ruta que vuelve a bifurcarse, con las rutas superior e inferior similares con una sola persona o un grupo de personas atadas. Cada ruta sigue bifurcándose en más rutas, mostrando la naturaleza interconectada de estos "dilemas del tranvía" y el impacto combinado de las decisiones individuales.

Por interesante que suene, el dilema del tranvía sigue siendo un experimento sobre el pensamiento y ha sido criticado por ser demasiado extravagante y poco realista.

Página 4

Pero los vehículos autónomos ahora nos presentan una versión real de ese dilema.
Si decidimos incorporar vehículos autónomos de manera amplia, entonces ¿cómo lidiamos con los errores que están destinados a ocurrir, aunque se den relativamente pocos errores de ese tipo?

y ¿qué sucede con un sistema de transporte completo compuesto por vehículos autónomos, con personas, climas y diferentes condiciones de la carretera?

¿Cómo gestionamos simultáneamente cientos de dilemas del tranvía que dependen unos de otros?

Una dificultad adicional importante es que, a diferencia del clásico dilema del tranvía, donde se sabe cuántas personas hay en cada lado de la vía, un automóvil autónomo, y otros tipos de tecnología, operan bajo un alto grado de incertidumbre.

Puede que ni siquiera sepamos si hay personas en las vías, y mucho menos el número concreto que hay y a qué grupos pueden representar.

¿Cómo hacemos juicios de valor frente a semejante incertidumbre?

Imagen de una variación del problema clásico del tranvía con información incompleta: un tranvía se acerca a una bifurcación. Hay un hombre parado en la intersección que controla con una palanca por qué vía se encaminará el tranvía. El hombre sospecha que hay personas atadas a las vías y, por lo tanto, se perderán vidas humanas si el tranvía continúa, pero no sabe cuántas personas hay, quiénes son, etc.

Página 5

El dilema del tranvía ilustra una teoría específica dentro de la filosofía moral:
El utilitarismo
¿Puede esta teoría ofrecernos algunas pautas?

El utilitarismo se refiere a un principio moral que sostiene que la mejor acción, en cualquier situación, es la que produce el mayor equilibrio entre beneficios y daños para todas las personas implicadas.

El utilitarismo proviene de los filósofos y economistas ingleses de finales del siglo XVIII y principios del XIX, Jeremy Bentham y John Stuart Mill.

Un caricaturizado Jeremy Bentham se para y posa ante la cámara o el espectador.
Una cita famosa de Bentham es: “La mayor felicidad del mayor número es la medida del bien y del mal”.
¡Suena genial!

Un robot o figura personificada de la IA “ducha” a las personas con felicidad: un grupo de personas bailan y saltan de emoción mientras un robot masivo se para sobre ellas y rocía partículas del emoji de “cara sonriente” y emoji de “amor” al estilo del famoso Salt Bae.

Desafortunadamente, la aplicación de estas ideas a los vehículos autónomos, y al diseño e implementación de la tecnología de manera más general, abre una lata de gusanos.

Y tiene nombre:
moralidad algorítmica
Una lata abierta se encuentra de lado, con su contenido cayendo. La tapa de la lata está abierta a la derecha y el reflejo en ella muestra el dilema del tranvía. El contenido de la lata son pequeños gusanos que salen de la lata.

Página 6

La moralidad algorítmica es el acto de atribuir juicios morales a los algoritmos.

Dos mujeres paradas en cada hombro de una figura de robot de IA. La de la izquierda está vestida de diablo, con un tridente en la mano, cuernos en la cabeza y túnica roja. La de la derecha está vestida de ángel, con un halo y alas, y túnica blanca. Ambas mujeres hacen gestos y hablan a la IA. La IA parece perpleja, con las cejas levantadas y los dientes apretados.

Y hacerlo es problemático. Esta es la razón:

Para empezar, ¿cómo medimos la felicidad y la infelicidad?

Una mujer muestra una gran sonrisa a la cámara o al espectador. Un par de manos usan una cinta métrica para determinar el ancho de su sonrisa, como indicador de la cantidad de felicidad.

¿Y cómo codificamos esas medidas en un conjunto de objetivos que sean comprensibles para un algoritmo?

Rara vez existe una fórmula matemática o una declaración lógica que pueda establecer el equilibrio entre los beneficios y los daños.

Una científica escribe frenéticamente fórmulas en una pizarra. Lleva anteojos y sostiene un rotulador, con el que garabatea fórmulas de felicidad, tristeza, cantidad de beneficio y grado de daño.

En otras palabras: simplemente no hay una fórmula para lo “correcto” o lo “incorrecto”.
Y no existe una fórmula para definir los valores, y cómo surgen y cambian esos valores en situaciones sociales complejas.

Otra razón por la cual la moralidad algorítmica es problemática es que, cuando se comete un error de juicio sobre lo que está bien o mal—y, como ya sabemos, se cometerán errores porque el mundo es complejo, incierto y, quizás, incluso impredecible—,la moralidad algorítmica requeriría de un algoritmo para asumir la responsabilidad por el error.

Se están tomando fotografías policiales de un robot o una IA personificada. Lo vemos en dos poses diferentes y sosteniendo una pizarra: una frente al espectador o la cámara, y la otra mirando de perfil, con expresión de culpabilidad.

Página 7

Pero responsabilizar a un algoritmo por un error no tiene sentido:
Una IA personifica o robot está manipulando a un humano: sostiene una marioneta humana en la mano izquierda y mueve los hilos para que el humano salude con la mano y mire hacia arriba. El robot tiene el brazo derecho apoyado en la cintura.

un algoritmo no posee conciencia, ni libre albedrío, no toma una decisión intencional que conduce a un error, y por lo tanto no puede ser considerado responsable.

¿Dónde nos deja esto?
El abrelatas que es el problema del automóvil nos mostró que no podemos delegar la ética en las máquinas.

Que todavía depende de nosotros, los humanos, tomar decisiones y medidas (o elegir no actuar), de acuerdo con nuestros valores, y con las leyes vigentes.

Y luego depende de nosotros asumir la responsabilidad de las consecuencias de cualquier error.

Un ser humano sonríe mientras sostiene una marioneta robot: mueve los hilos para hacer que el robot baile un poco.

No podemos externalizar hacia una máquina el trabajo de ser humanos.

Página 8

En resumen, para incorporar la ética en sistemas sociotécnicos como la IA, debemos pensar qué valores están alrededor de esos sistemas, quién se beneficia cuando los sistemas funcionan bien, y quién se resulta perjudicado por sus errores.

Y debemos asumir colectivamente la responsabilidad de decidir sobre el equilibrio entre los beneficios y los daños, para que “la mayor felicidad” que Jeremy Bentham promete al mayor número de personas también sea disfrutada por la mayor diversidad de partes interesadas.

Este trabajo de comprensión y negociación colectiva de las compensaciones es lo que hace que el diseño de las tecnologías se sustente en las personas.

Imagen de un gran “árbol de IA”: las hojas están compuestas por una gran red o gráfico, con nodos y bordes coloridos y densamente interconectados. Las raíces del árbol abarcan una gran área subterránea, donde un grupo de humanos se aferran a las raíces y mantienen el contacto del árbol con la tierra. Los humanos están dibujados con líneas abstractas y solo son identificables por una silueta, y aparecen en una variedad de colores.