Desafíos éticos del coche autónomo

Pablo Gómez-Abajo[a]

[a] *Grupo de Modelado e Ingeniería del SOftware (MISO)*
Universidad Autónoma de Madrid

Resumen

Investigadores y expertos tienden a centrarse en un desafío ético para los coches autónomos: el problema del tranvía. Desafortunadamente, el foco puesto en exclusiva en este problema ético parece cegar a estos profesionales y expertos que dejan al margen otros desafíos éticos aún sin resolver. Este artículo muestra que uno debe tomar la tecnología en sus propios términos. Una vez se tenga bien claro cómo funcionan estos vehículos, es decir, la lógica matemática detrás de su aprendizaje, se puede ver que existen otras opciones de diseño que son mucho más relevantes desde el punto de vista moral. También el artículo describe brevemente cómo aplicar la herramienta Wodel-Test de creación de entornos de pruebas de mutación independiente del dominio al coche autónomo. A continuación se hace una breve descripción de en qué sujetos ha de recaer la responsabilidad en caso de accidente de estos vehículos autónomos. Por último, como conclusión se incluye una breve reflexión sobre lo que supone este avance tecnológico así como otros avances de las tecnologías emergentes que se engloban dentro de la Inteligencia Artificial.

1. Introducción

Es habitual que cuando alguien se interesa por la ética de los coches autónomos, lo primero que surge en la conversación es "el problema del tranvía". El problema del tranvía es un experimento mental en el que a alguien se le presentan dos situaciones que presentan opciones similares y consecuencias potenciales [8, 19, 22, 23, 29, 31].

La situación A (conocida como *Interruptor*) es aquella en la que un tranvía fuera de control está recorriendo una vía y se estrellará y matará a cinco obreros a menos que un observador accione un interruptor y desvíe el tranvía por una pista lateral que sólo matará a un trabajador.

La situación B (conocida como *Puente*) tiene a un observador cruzando por un puente, desde donde observa que cinco personas serán matadas por el tranvía a menos que el observador arroje desde el puente a otro individuo corpulento a

Email address: Pablo.GomezA@uam.es (Pablo Gómez-Abajo)

las vías que se encuentran debajo, deteniendo así el tren y salvando a las cinco personas.

La mayoría de los filósofos están de acuerdo en que es moralmente permisible matar al individuo que está en el *Interruptor*, pero también piensan que no es correcto empujar al individuo corpulento en el *Puente* [18]. El caso tiene los mismos efectos: matar a uno para salvar a los cinco.

Esta discrepancia entre estos dos casos ha llevado a una gran cantidad de artículos sobre "el problema" y ha llevado a una investigación más amplia denominada "Tranviología".

El trabajo se estructura como sigue: primero, se describen los desafíos éticos que plantean los vehículos autónomos. A continuación, se hace una breve descripción de cómo utilizar la herramienta Wodel-Test para realizar pruebas de mutación sobre los algoritmos del coche autónomo. Por último, se realiza una breve reflexión acerca de quién es el responsable cuando uno de estos sistemas de conducción automática falla. Como cierre, se extraen varias conclusiones de este trabajo.

2. Todo va sobre el POMDP

Un proceso de decisión de Markov parcialmente observado (*Partially Observed Markov Decision Process*, POMDP) es una variante del proceso de decisión de Markov (*Markov Decision Process*, MDP). El modelo MDP es un modelo matemático para varios problemas de control y de planificación en ingeniería y computación. Por ejemplo, MDP es útil en un entorno que es *completamente observable*, tiene intervalos de tiempo *discretos*, y *pocas opciones de acción* en diversas condiciones [26]. Se puede pensar en un modelo MDP que es útil en algo así como el juego del ajedrez o el tres en raya. El algoritmo conoce completamente el entorno (el tablero, las piezas, las reglas) y espera a que su adversario haga un movimiento. Una vez que se ha efectuado el movimiento, el algoritmo puede calcular todos los movimientos potenciales que tiene en frente, tomando la decisión "mejor" u "óptima".

Desafortunadamente, muchos entornos del mundo real no son como el tres en raya o el ajedrez. Además, cuando tenemos sistemas robóticos como el coche autónomo con muchos sensores, el propio sistema *no puede tener completo conocimiento* de su entorno. Hay un conocimiento incompleto debido a las limitaciones en el rango y la fidelidad de los sensores y el retardo (en el tiempo que se tarda en leer los sensores, el entorno contínuo y dinámico puede haber cambiado). Además, un robot en esta situación toma una decisión utilizando las observaciones actuales así como un histórico de las acciones y observaciones previas. En términos más precisos, un sistema está midiendo todo lo que puede en un estado *s*, y el conjunto finito de estados $S = \{s1, \ldots, sn\}$ es el entorno. Cuando un sistema se observa a sí mismo en *s*, y toma una acción *a*, se traslada a un nuevo estado *s'*, y puede tomar la acción *a'*. El conjunto de acciones posibles es $A = \{a1, \ldots, ak\}$. Por tanto, en un punto dado, un sistema está decidiendo qué acción tomar basado en su estado actual, su estado previo (si lo hay), y el estado al que se espera trasladar. La diferencia crucial aquí es

que un POMDP está funcionando en un entorno en el que el sistema (o agente) tiene un conocimiento incompleto e incertidumbre, y está funcionando en base a probabilidades; en esencia, un coche autónomo no está funcionando en base a un MDP, su funcionamiento se parece más al de un POMDP.

¿Cómo sabe un sistema qué acción debe tomar? Existen varios modos potenciales, pero me centraré en uno aquí: el aprendizaje reforzado. Para un POMDP que utiliza el aprendizaje reforzado, el vehículo autónomo aprende a través de un recibo de una señal de recompensa. Sistemas como éste que utilizan POMDPs tienen señales de recompensa (o a veces de 'coste') que indican las acciones que deben perseguir (o evitar). Pero estas señales están basadas en las distribuciones probables de qué acciones en el estado activo s llevarán a mayores recompensas en un estado futuro sn descontinuado para acciones futuras. Pongamos un ejemplo para explicarlo en términos más sencillos. Uno puede razonar que está cansado pero que necesita terminar este artículo (estado), por tanto uno puede decidir echarse una siesta ahora mismo (una acción del conjunto de acciones posibles), y a continuación despertarse tarde para terminar el artículo. Sin embargo, tampoco sabe si descansará bien en la siesta, dormirá de más, o se sentirá peor cuando se despierte, de este modo frustrando aún más sus planes para su artículo, aunque el pensamiento de una siesta inmediata podía darle una recompensa inmediata (dormir y descansar). La decisión óptima (o, en términos de POMDP, la política), podía ser dormirse ahora mismo. Sin embargo, esto no es así. Más aún, POMDP requiere que uno escoja la política más óptima bajo condiciones de incertidumbre, para tareas de decisión secuenciales, y descontadas para estados de recompensa futuros. Abreviando, la política óptima sería tomarse una taza de café, terminar el artículo, y a continuación irse pronto a la cama porque así verdaderamente maximizará la cantidad total de sueño al no haberse echado la siesta durante el día y haber terminado el trabajo.

Sin embargo, el desarrollo de un sistema de aprendizaje adecuado requiere contar con muchas experiencias para construir creencias sólidas. En el caso de agotamiento y escritura académica, si no se tiene una larga experiencia en la redacción de trabajos académicos y en el cansancio, uno podría pensar que cualquier decisión que tome sería, en el mejor de los casos, considerada aleatoria (es decir, es 50/50 como si uno se encuentra en su mejor política). De todas formas, si se tiene una larga historia de escritura académica y cansancio, así como las recompensas y los costes de las decisiones anteriores, se puede predecir con precisión qué acción maximizará la recompensa. Uno conoce su mejor política. Esto se debe a que las situaciones, en general, relacionan las creencias con las acciones [27, 28]. Sin embargo, matemáticamente, lo que realmente quiere decir es que una política π es un conjunto contínuo de distribuciones de probabilidad sobre el conjunto total de estados S. Y una política óptima es aquella que maximiza las recompensas.

Esa función se convierte en una *función de valor*, que es, una función de cómo la acción de un agente y su estado de creencia inicial ($b0$) puebla la recompensa obtenida esperada una vez recibe el feedback y *actualiza* sus creencias acerca de varios estados y observaciones, y por tanto mejora su política. Después continúa su patrón, una y otra vez, hasta que puede empezar a predecir mejor las acciones

que van a maximizar su recompensa. Aparentemente, esta estructura permite a un sistema aprender cómo actuar en un mundo que es incierto, ruidoso, confuso, y no observable completamente. El papel de un ingeniero es definir las tareas u objetivos de forma que el agente artificial que sigue a un modelo POMDP puede tomar una serie de acciones, aprender cuáles corresponden con las observaciones correctas sobre su estado del mundo y actuar en consecuencia, a pesar de la incertidumbre.

El mundo del coche autónomo hará sin duda un gran uso de los POMDPs en sus arquitecturas de software. Mientras hay una gran cantidad de técnicas de computación que se pueden elegir, las probabilísticas como los POMDPs se han demostrado por encima de las mejores candidatas para construir y desplegar robots autónomos. Tal como Thrun explica [30], "la aproximación probabilística [...] es la idea de representar la información a través de densidades de probabilidad" sobre las áreas de percepción y control, y "la robótica probabilística ha llevado a sistemas desplegados con niveles de autonomía y solidez sin precedentes". Por ejemplo, recientemente Cunningham et al. utilizaron los POMDPs para crear un proceso de creación-de-decisión-multi-política para la conducción autónoma que estimó cúando adelantar a un vehículo más lento así como el modo de integrarse en el tráfico teniendo en cuenta las preferencias de conducción, alcanzar los objetivos con agilidad y y la comodidad del usuario [5].

Es bien conocido, de todas formas, que los POMDPs son ineficientes computacionalmente y que a medida que la complejidad de un problema crece, algunos problemas pueden ser irresolubles. Para tener en cuenta esto, muchos investigadores realizan ciertas suposiciones sobre el mundo y las matemáticas para hacer que los problemas se puedan computar, o utilizan aproximaciones heurísticas para facilitar la búsqueda de políticas. Sin embargo, cuando uno realiza una aproximación, en cualquier sentido, no existe garantía de que un sistema actuará en un modo precisamente óptimo. De cualquier forma que manipulemos las matemáticas, pagamos el coste en un dominio o en otro: o bien computacionalmente, o bien en la búsqueda de los mejores resultados.

La cuestión importante a destacar en toda esta discusión es que cualquier coche autónomo que está utilizano un método probabilístico como un conjunto o arquitectura de POMDPs va a estar realizando dos cosas. En primer lugar, realizar "decisiones" no es lo que completa el conocimiento, como en el problema del tranvía. Más bien, lo que completa el conocimiento es elegir la probabilidad de que una acción cambie un estado de cosas. En esencia, las decisiones se toman en condiciones dinámicas de incertidumbre. En segundo lugar, esto quiere decir que para que un coche autónomo controlado por un sistema de aprendizaje opere de forma efectiva, se requieren cantidades sustanciales de "episodios" o entrenamiento en varios entornos bajo circunstancias similares y diferentes para que tome decisiones "buenas" cuando está desplegado en una carretera. Es decir, el coche autónomo necesita ser capaz de extraer las decisiones de una historia muy rica de interacciones para extrapolarlas de una manera prospectiva. De todas formas, al tratarse de sistemas de aprendizaje, no garantizan que algo completamente nuevo e imprevisto no vaya a crear confusión al coche autónomo o provocar que actúe de formas imprevistas. En el caso del problema del

tranvía, podemos empezar a ver de forma inmediata las diferencias entre cómo un observador puede razonar y como puede hacerlo el coche autónomo.

3. Los disparates de los tranvías

Patrick Lin [24] defiende la utilización del Problema del Tranvía como una "bomba de intuición" para hacernos pensar sobre qué tipos de principios debemos programar en los coches autónomos. Discute que la utilización de experimentos del pensamiento como éste "aisla un par de asunciones sobre cómo los coches sin conductor deben gestionar los choques inevitables, tan raros como éstos puedan ser. Desentraña las preguntas de (1) si los números importan y (2) si matar es peor que dejar morir". De forma adicional, indica que al ser los coches autónomos creaciones del ser humano a lo largo del tiempo, "los programadores y los diseñadores de coches autónomos [...] tienen el tiempo necesario para hacerlo bien, y por tanto, asumir más responsabilidad para los resultados desfavorables", por lo tanto, llevando a cabo alguna resolución de si había suficiente intencionalidad para que la acción sea juzgada como moralmente correcta o incorrecta.

Si bien uno puede estar de acuerdo con la evaluación de Lin de que muchos casos en la filosofía no están diseñados para escenarios del mundo real, sino para aislar y presionar sobre nuestras intuiciones, esto no significa que sean adecuados para todos los propósitos. Como dice Peter Singer [25], reducir la "filosofía... al nivel de resolver problemas de ajedrez" es de bastante poca ayuda, al "haber cosas que son más importantes". Hay que tomarlo con un cuidado especial para ver las asimetrías entre los casos como el del Problema del Tranvía y estos algoritmos, que no son agentes morales pero toman decisiones moralmente importantes. La primera y más sencilla forma de ver esto es reconocer que un coche autónomo que utiliza algo como un POMDP en un entorno dinámico no toma una decisión en un momento dado, sino que toma decisiones secuenciales. El coche autónomo toma una decisión basada en un conjunto de distribuciones de probabilidades sobre qué acción le dará la función de recompensa más alta (o minimizará el coste) basada en el conocimiento previo, las observaciones actuales y los estados futuros probables. Y esta forma en que funcionan los coches autónomos es diferente a la de de los casos del Tranvía, donde hay que tomar una decisión en un momento dado.

Segundo, y de una forma más radical, tendríamos que modelar problemas similares a los del tranvía en una variedad de situaciones, y entrenarlos en esas situaciones (o episodios) cientos, tal vez miles de veces para que el sistema aprenda qué hacer al respecto en ese ejemplo. No sólo no tomaría mágicamente la decisión "correcta" en ese caso porque las matemáticas y el conjunto de observaciones anteriores de hecho no lo permitirían. En realidad, tenemos que especificar previamente qué es "correcto" para que aprenda qué hacer. Esto se debe a que los tipos de algoritmos que utilizamos se están optimizando por naturaleza. Quieren encontrar la estrategia más óptima para maximizar su función de recompensa, y este aprendizaje, por cierto, implica que necesitan cometer

muchos errores. Por ejemplo, un grupo de investigadores de la Universidad Carnegie Mellon se negó a usar simulaciones para enseñar a un vehículo aéreo a volar y pilotar. Más bien, permitieron que este vehículo aéreo se bloqueara más de 11.500 veces para aprender políticas simples de auto supervisión de pilotaje [14]. De hecho, este aprendizaje mediante la práctica es exactamente a lo que se dirigen, en gran medida, las pruebas de los coches autónomos en condiciones reales de carretera: el aprendizaje en la vida real y no las meras simulaciones. Sin embargo, no estamos pidiendo a los coches que choquen contra las personas o que elijan si es mejor matar a cinco hombres o a cinco mujeres embarazadas.

Más aún, incluso si uno decidiese simular estos casos del tranvía una y otra vez, y diversificarlos hasta cierto tipo de grado suficiente, debemos considerar el punto sencillo pero estricto de que a menos que uno conozca de antemano la respuesta correcta, las matemáticas no van a ser de ayuda. Además, hay muchas dificultades para encontrar filósofos que estén de acuerdo con la forma de vivir y el código moral correcto a lo largo de más de 2.000 años, o incluso para encontrar un acuerdo sobre qué hacer en el problema del tranvía. Lo que es incluso peor es que si tomamos la opinión de que nuestras intuiciones deben guiarnos a encontrar datos para estos dilemas morales, no encontraremos de hecho datos fiables. Esto puede observarse fácilmente con dos ejemplos que muestran cómo de hecho las personas no actuamos de forma coherente: la Paradoja de Allais y la Paradoja de Ellsberg. Ambas paradojas desafían los axiomas básicos que Von Neumann y Morgenstern postularon para su teoría de la utilidad esperada [33]. La teoría de la utilidad esperada establece básicamente que las personas elegirán un resultado en función de si la utilidad esperada de ese resultado es mayor que la de todos los demás resultados potenciales. Abreviando, esta teoría quiere decir que las personas somos maximizadores de la utilidad. En la Paradoja de Allais, encontramos que en un experimento dado las personas no actúan de manera coherente para maximizar su utilidad (o lograr una satisfacción de preferencia) y, por lo tanto, violan el axioma de sustitución de la teoría [2]. En la Paradoja de Ellsberg, las personas terminan eligiendo cuando no pueden realmente inferir las probabilidades que maximizarán sus preferencias, violando así los axiomas de integridad y monotonicidad [6].

Uno puede objetar este punto y afirmar que el utilitarismo no es la única teoría moral, y que, de hecho, no queremos que la utilidad maximice los coches autónomos. Prefiriríamos tener automóviles que respeten los derechos y las vidas, más como una ética de la virtud o un enfoque deontológico de la ética. Pero si eso es así, entonces hemos eliminado la necesidad de Problemas del Tranvía desde el principio. No está permitido matar a nadie si eso es cierto, a pesar de los números. O símplemente declaramos a priori que ganan las justificaciones menos malignas, y por lo tanto, *in extremis*, hemos evitado el problema [13, 20]. O, si garantizamos que los coches autónomos terminan realizando los cálculos como se haría para obtener la utilidad de una acción, entonces parece que no existe el problema - los números ganan. Espera, espera, uno puede responder, esto va demasiado rápido. Claramente, pensamos que los coches autónomos no deben matar a nadie, el problema del tranvía es válido y aún pueden encontrarse en situaciones en las que no tienen más remedio que matar a alguien, así que,

¿quién debería ser?

Cito aquí nuevamente que estamos atrapados en una posición incómoda con la necesidad de datos y entrenamiento frente a la necesidad de saber qué nos dicta la verdad moral: ¿queremos modelar dilemas morales o queremos resolverlos? Si es lo primero, podemos seguir haciéndolo de forma indefinida. Podemos modelar dilemas morales y pedir a las personas que participen en experimentos, pero eso sólo nos dirá la realidad empírica de lo que piensan esas personas. Y esa puede ser una respuesta significativamente diferente a la que la moralidad dicta que se debe hacer. Si es lo segundo, todavía me siento escéptico de que este sea el marco adecuado para discutir los problemas éticos que surgen con los coches que se conducen por sí mismos. Quizás el problema del tranvía no sea más que una distracción sin solución de la cuestión de los umbrales de seguridad y otros tipos de preguntas éticas con respecto a los efectos de segundo o tercer orden de la automatización de los automóviles en la sociedad.

De hecho, si uno está en lo cierto, entonces la configuración completa de un dilema moral para que un agente no moral "elija" la mejor acción es una elección falsa porque no hay ninguna opción que un ingeniero pueda planear de manera previsible. Lo que es más, incluso si el ingeniero exhibiera suficiente visión y construyera un sistema de aprendizaje que pudiera recoger pistas sutiles de las interacciones con el entorno y grandes cantidades de datos, esto supone que realmente se ha descubierto la acción correcta a tomar. Se han clasificado los datos como "buenos" o "malos" y los hemos enviado a un sistema. Sin embargo, los agentes morales humanos no hemos decidido esto, ya que hay un debate sobre lo que se debe hacer en cada situación, así como incertidumbre. Los problemas del tranvía se construyen de tal manera que el agente tiene una opción y sabe con certeza qué sucederá si toma esa decisión. Además, esa elección se construye como un dilema: parece que no importa la elección que haga el agente, ya que terminará cometiendo algún tipo de mal. Bajo condiciones de conducción reales, rara vez se dará este caso. Y si intentamos encontrar una solución a través de los medios tecnológicos disponibles, todo lo que hemos hecho es mostrar una gran cantidad de datos al sistema y optimizar su comportamiento para la tarea que se le asignó. Si se ve de esta manera, modelar dilemas morales como tareas y optimización parece moralmente rechazable.

Más importante aún para nuestros propósitos es que debemos tener bien claro que la Inteligencia Artificial no es humana. Incluso si la IA fuese un agente moral (y si estuviéramos de acuerdo en lo que eso parecería), el antropomorfismo de partida en el caso del Problema del Tranvía en realidad nos está cegando de algunos de los peligros reales. En los casos del Tranvía de la filosofía moral clásica, se asume desde el principio que: (i) hay un agente moral que afronta la elección; (ii) este agente moral es consciente de sí mismo con una historia de actuación en el mundo, comprende conceptos y posee suficiente inteligencia para identificar contextualmente cuándo las limitaciones morales se ven superadas por aquellas que son significativas; y (iii) la inteligencia puede, en algún sentido, equilibrar o medir aparentemente (o verdaderamente) bienes y obligaciones en conflicto. Además, como resume Barbara Fried sobre la estructura de los Problemas del Tranvía [12]:

Las hipótesis suelen compartir una serie de características que van más allá del dilema básico del daño/compensación de terceros. Estas incluyen las consecuencias de las opciones disponibles que se estipulan para ser conocidas con certeza antes del suceso; que los actores son todos individuos (en contraposición a instituciones); que las posibles víctimas (del daño que imponemos por nuestras acciones o que permitimos que ocurran por nuestra inacción) son individuos generalmente identificables que están muy cerca de los posibles actores; y que la cadena causal entre el acto y el daño es bastante directa y aparente. Además, los actores generalmente afrontan una decisión única sobre cómo actuar. En otras palabras, no se suele invitar a los lectores a considerar las consecuencias de ampliar el principio moral mediante el cual el dilema inmediato se resuelve en un gran número de casos.

Sin embargo, no sólo todos los atributos mencionados previamente van más allá de las capacidades actuales de cualquier sistema de inteligencia artificial: la situación en la que opera un coche autónomo no se comporta según ninguna de las suposiciones realizadas en casos similares a los del Problema del Tranvía [16]. Existe una disyuntiva entre decir que los humanos "programarán" el coche autónomo para hacer la elección moral "correcta" (por tanto, instanciando el caso del Problema del Tranvía en los coches autónomos) y entre afirmar que un coche autónomo es un autómata de aprendizaje suficientemente capaz de tomar decisiones importantes desde el punto de vista moral en el orden del Problema del Tranvía. Además, uno no puede quedarse símplemente preguntándose cuáles serán las mejores consecuencias, ya que en este caso no hay un "problema" real en el asunto: se salva a los cinco en lugar de a uno, sin hacer más preguntas.

Es inútil continuar poniendo el foco e insistiendo en que el Problema del Tranvía agota el panorama moral de las cuestiones éticas con respecto a los coches autónomos y su despliegue. Todo lo que la Inteligencia Artificial puede hacer es poner de relieve las tensiones existentes en nuestra vida cotidiana que tendemos a asumir. Esto puede deberse a las limitaciones que tenemos como seres humanos para ver las estructuras y los sistemas sociales subyacentes porque no podemos procesar de golpe cantidades tan grandes de datos. Sin embargo, la Inteligencia Artificial es capaz de encontrar patrones novedosos en grandes cantidades datos y planificar en base a esos datos. Estos datos pueden reflejar nuestros sesgos, o pueden ser símplemente una agregación de cualquier situación que el coche autónomo haya encontrado. Lo único que los coches autónomos requieren de los seres humanos es hacer explícitas las tareas que requerimos y cuáles son las recompensas y los objetivos; no requerimos que los coches autónomos nos digan que estas son nuestras metas, recompensas y objetivos. Desafortunadamente, esta distinción no es algo que se haga explícito a menudo.

Más bien, el debate a menudo oscila entre si los *agentes humanos* deberían "programar" la respuesta correcta o si el *sistema de aprendizaje* puede de hecho dar la respuesta moralmente correcta. Si es lo primero, debemos admitir que aquí se ignora el hecho de que un sistema de aprendizaje no funciona en términos tan sencillos. Es un sistema de aprendizaje que se verá limitado por sus sensores, su experiencia y las diversas arquitecturas y subarquitecturas internas. Pero

actuará en tiempo real, lejos de sus desarrolladores, y en un entorno amplio y dinámico, por lo que los humanos responsables de su comportamiento tendrán, en el mejor de los casos, una responsabilidad sobre su manera de funcionar en cada situación.

Si es lo segundo, se ha intentado demostrar aquí que el aprendizaje del coche autónomo no posee cualidades relevantes para resolver el aberrante Problema del Tranvía. Incluso si se tuviera que entrenarlo en grandes cantidades y variantes de casos del Tranvía, siempre habrá situaciones diferentes que pueden surgir y que no producirían la reacción estimada o prevista por el ser humano. Esto son matemáticas sencillas. Sólo se pueden hacer generalizaciones acerca de los comportamientos, especialmente acerca de los comportamientos humanos desordenados que pueden dar lugar a casos similares a los del Problema del Tranvía - cuando hay un conjunto de datos significativamente grande (a esto se le denomina la ley de los grandes números). Desafortunadamente, esto significa que no hay forma de saber qué hará un controlador en cualquier circunstancia dada. Por tanto, si bien se puede identificar valor al pensar detenidamente en las cuestiones morales relacionadas con casos similares a los del Problema del Tranvía, también existen límites, particularmente con respecto a los pesos de decisión, las políticas y la incertidumbre moral.

4. Las funciones de *valor*

Si estamos de acuerdo en que el problema del tranvía ofrece poca orientación sobre los problemas sociales más amplios que existen, en particular el valor de un cambio masivo y la investigación científica, entonces se puede comenzar a reconocer los problemas de gran alcance que la sociedad tendrá que enfrentar con los coches autónomos. Como explican Kate Crawford y Ryan Calo [4], "los sistemas autónomos están cambiando los lugares de trabajo, las calles y las escuelas. Necesitamos asegurarnos de que esos cambios sean beneficiosos, antes de que se incorporen a la infraestructura de la vida cotidiana". En resumen, debemos identificar los valores que queremos actualizar a través de la ingeniería, el diseño y la implementación de tecnologías, como los coches autónomos. Por tanto, aquí hay un doble sentido: sabemos que el software que ejecuta estos sistemas intentará maximizar sus funciones de valor, pero también debemos asegurarnos de que estos sistemas maximizan los de la sociedad.

Entonces, ¿cuáles son los valores que queremos maximizar con los coches autónomos? Obviamente, queremos que los coches autónomos sean mejores conductores que las personas. Con más de 5,5 millones de accidentes por año y más de 30.000 muertes sólo en los EE.UU., la seguridad parece ser la principal motivación para automatizar la conducción. Más del 40 por ciento de los accidentes fatales involucran "una combinación de alcohol, distracción, participación de drogas y/o fatiga". Esto significa que si todos estuvieran utilizando vehículos de conducción automática, al menos en los EE.UU., podra haber al menos una reducción de 12.000 muertes por año. Aparentemente, salvar vidas es un valor primordial.

Son esfuerzos valiosos la forma en que esto ocurre, los efectos de las políticas, las elecciones en cuanto a infraestructura y el desarrollo tecnológico. No sólo tenemos una solución innovadora. No podemos "codificar" la ética y lavarnos las manos. La innovación, más bien, debe provenir de la intersección de las humanidades, las ciencias sociales y políticas, trabajando junto con la ingeniería. Esto se debe a que los valores que queremos mantener primero deben ser identificados, impugnados y comprendidos. Richard Feynman dijo popularmente: "No puedo crear lo que no entiendo". Es decir, no podemos crear, o mejor dicho, recrear aquellas cosas que ignoramos.

De hecho, la infraestructura normativa es lo más importante. Lo normativo aquí tiene dos significados que debemos tener en cuenta: (i) el "deber" filosófico o moral; y (ii) el enfoque foucauldiano de "normalización" que identifica las normas como aquellos conceptos o valores que buscan controlar y juzgar nuestro comportamiento [10]. Estas son dos nociones muy diferentes de lo que es la "normativa", pero ambas son de importancia crucial para la identificación de valor y la creación de funciones de valor para tecnologías autónomas.

Desde la perspectiva moral, uno debe ser capaz de identificar todos aquellos valores morales que deben operacionalizarse no sólo en el sistema del vehículo autónomo, sino en los métodos de adjudicación que se usarán cuando estos valores entren en conflicto. Esto no es, se podría pensar, un retorno al Problema del Tranvía. Más bien, es una opción de valor tecnológico sobre cómo uno decide diseñar un sistema para seleccionar un curso de acción. En los sistemas de aprendizaje de objetivos múltiples, a menudo ocurren situaciones en que los objetivos (es decir, las tareas o los comportamientos que se deben cumplir) entran en conflicto entre sí, están relacionados o incluso son endógenos. El ingeniero debe diseñar la forma de encontrar cómo priorizar objetivos particulares o crear un sistema para intercambio, como si se quiere conservar la energía o mantener la comodidad [32]. La forma en que lo hacen es una cuestión de matemáticas, pero también es una opción sobre si están privilegiando tipos particulares de matemáticas que a su vez privilegian determinados tipos de comportamientos (como la satisfacción).

Además, el hecho de alejar el enfoque de eventos trágicos y raros como los del Problema del Tranvía permite abrir más problemas sistémicos que se deben tener en cuenta para reducir los daños y garantizar la seguridad. Como sostiene Allen Wood [35], la mayoría de las personas nunca tendrían que enfrentar un caso del Tranvía si hubiera tranvías más seguros, se imposibilitara a los transeúntes el acceso a los interruptores y se dispusiera de una buena señalización para "evitar que alguien se sitúe en lugares donde podría morir o ser atropellado por un tranvía fuera de control". En resumen, hay que pensar en el uso, el diseño y la interacción de la experiencia diaria de los consumidores, usuarios o transeúntes con la tecnología. Se trata de entender cómo los coches autónomos pueden cambiar el diseño y la composición de las ciudades y los pueblos, y qué efectos pueden tener en todo, desde el acceso a los recursos básicos hasta la creación de nuevas formas de desigualdad.

Desde la perspectiva foucaldiana, las cosas se vuelven un poco más interesantes, y aquí es donde creo que muchas de las preocupaciones éticas comienzan

a aparecer. Las normas que rigen la forma en que actuamos, las suposiciones que hacemos acerca de la idoneidad de las acciones o comportamientos de los demás y el valor que asignamos a esos juicios, son aquí una cuestión de evaluación empírica [9, 10]. Por ejemplo, las encuestas de opinión pública son instrumentos que nos dicen lo que las personas "piensan" sobre algo. Sin embargo, menos obvias son las formas en que ajustamos sutilmente nuestro comportamiento sin hablar o, en algunos casos, incluso a partir de señales culturales y sociales. Estos son los tipos de normas que preocupan a Foucault. Este tipo de normas son las que aparecen en grandes conjuntos de datos, en sesgos, en "patrones de vida". Y son estas clases de normas las que son más difíciles de identificar, las más difíciles de cambiar.

La importancia de todo esto para los vehículos autónomos se basa en las suposiciones que los ingenieros hacemos sobre el comportamiento humano, los valores humanos o incluso cómo se ve de "apropiada" una acción. Desde el punto de vista del diseño sensible al valor (DSV), uno puede considerar no sólo la cuestión del daño letal a los pasajeros o transeúntes, sino una gran cantidad de valores como la privacidad, la seguridad, la confianza, los derechos civiles y políticos, el bienestar emocional, la sostenibilidad ambiental, la belleza, el capital social, la equidad y el valor democrático. Para el DSV, se busca encapsular no solo los aspectos conceptuales de los valores que una tecnología en particular traerá (o afectará), sino también cómo "las propiedades tecnológicas y los mecanismos subyacentes apoyan o dificultan los valores humanos".

Pero uno se da cuenta de que en todas estas opciones, los Problemas del Tranvía no tienen lugar. Por ejemplo, muchas de las implicaciones sociales y éticas de los coches autónomos pueden ser extremadamente sutiles o símplemente obvias. Hay que tener en cuenta la idea recientemente distribuida por la firma Deloitte: "los coches autónomos afectarán a los servicios minoristas y de entrega de bienes". Como argumenta, los minoristas intentarán utilizar los coches autónomos para aumentar las áreas de captación, brindar mayores niveles de servicio al cliente enviándoles coches, reduciendo el tiempo de entrega o actuando como "centros de vecindarios", similar a una tienda de esquina móvil que entrega productos a tu propia casa. En esencia, los minoristas pueden atender mejor a sus clientes y "los no conductores no están" obligados "a tomar el autobús, el metro, el tren o la bicicleta [...] y por tanto, esto tendrá un impacto en las tiendas".

Sin embargo, este beneficio previsto de los coches autónomos puede que sólo se aplique a personas acomodadas que viven en una proximidad razonable a los puntos de venta de los minoristas. Sin duda, será difícil encontrar incentivos económicos en los "desiertos alimentarios" donde las personas de bajos ingresos que no tienen acceso al transporte viven a distancias cada vez más alejadas de los supermercados o tiendas de alimentación. El hecho de que estas personas en la actualidad no poseen medio de transporte y padecen la falta de acceso a productos de alimentación frescos y verduras, es un indicio de que seguramente tampoco tendrán capacidad para pagar los precios de la automatización y entrega, debido tal vez al aumento de los precios por el lujo de ser transportados de un lado a otro. Todo esto puede, en efecto, tener consecuencias más perju-

diciales sobre la pobreza y aumentar la brecha entre ricos y pobres, y extender en lugar de reducir las áreas que actualmente se consideran como "desiertos alimentarios".

Para estar seguros, existe una gran especulación sobre cómo los coches autónomos proporcionarán realmente beneficios netos para la sociedad. Muchos informes, desde diversos puntos de vista, estiman que los coches autónomos asegurarán que todas las estaciones de aparcamiento se convertirán en hermosos parques y jardines [1], y que las personas mayores, discapacitados y personas en riesgo de exclusión social tendrán acceso a un transporte seguro y confiable [7, 17, 21, 34]. Pero parece que se presta menos atención a cómo las limitaciones actuales de la tecnología requerirán una reformulación sustancial de la planificación urbana, la infraestructura y las vidas y comunidades de aquellas personas tanto cercanas como alejadas de los coches autónomos. Los circuitos Hyper-Loop para los coches autónomos, por ejemplo, pueden requerir pasos elevados para peatones o, como sugiere un investigador, "cercas eléctricas", para evitar que los peatones crucen al nivel de la calle. Otros sugieren que la mayor adopción de los coches autónomos será costosa pero beneficiosa para el medio ambiente, por lo que deberá ser comunitaria y deberá operar en trayectos largos [3]. Si esto es así, entonces surgirán preguntas sobre la presencia de vigilancia y la intervención en delitos potenciales, acoso o cualquier otro comportamiento ofensivo.

Todos estos efectos aparentemente pequeños o indirectos de los coches autónomos normalizarán su utilización, generarán pautas de comportamiento y sistemas de poder, y darán mayor importancia a algunos valores en concreto sobre otros. En el sentido foucauldiano, la adopción y el despliegue de los coches autónomos comenzará a cambiar la organización y la construcción de la "infraestructura colectiva" y esto requerirá alguna forma de racionalidad gubernamental (una imposición de estructuras de poder) en la sociedad [11]. Este tipo de planificación urbana, símplemente para permitir una mayor adopción de los coches autónomos, es una opción política; permitirá una "cierta asignación de personas en el espacio, una canalización de su circulación, así como el establecimiento de sus relaciones recíprocas". Por tanto, hacer que este tipo de decisiones sean transparentes y evidentes para los diseñadores e ingenieros de los coches autónomos será de gran utilidad para ver las suposiciones que hacen sobre el mundo y lo que ellos y otros valoran en él.

5. Pruebas de mutación en el coche autónomo

Como se indicó en el trabajo anterior de este curso [15], la técnica de las pruebas de mutación se utiliza para evaluar la calidad de los conjuntos de pruebas de un software o artefacto como sigue: Se introducen errores artificiales en el sistema generando una serie de sistemas mutados a los que posteriormente se les aplican los conjuntos de pruebas, y la medida de la calidad del conjunto de pruebas se obtiene a partir de la cantidad de estos errores artificiales que detecta. Esta técnica permite identificar defectos en el conjunto de pruebas y proporciona una receta al desarrollador para mejorarlo.

En el caso del coche autónomo, se puede pensar que cada uno de los sensores utilizados para dirigir su comportamiento está diseñado conforme a unos casos previstos. Por ejemplo, si el coche autónomo detecta que se aproxima a un objeto estático que está justo delante, el sensor correspondiente indicará al sistema que debe frenar. Supongo que el conjunto de sensores del coche autónomo seguirá una metodología de pruebas para su puesta a punto. Es en este punto donde se pueden utilizar las pruebas de mutación. Partiendo de un conjunto de situaciones posibles, se pueden utilizar estas situaciones como parámetros para la introducción de mutaciones en los casos previstos por el sistema, es decir, para cambiar una situación que ya se ha comprobado por otra diferente, ver la reacción que tiene el sistema ante esta variación, y de esta forma descubrir si el conjunto de pruebas es completo.

El conjunto de pruebas ha de mejorarse en caso de que no detecte la variación introducida artificialmente en la reacción prevista para cada caso. Por ejemplo: Mi conjunto de pruebas comprueba que el vehículo ha de reducir la velocidad en caso de atasco en la autopista. La mutación introducida incluye que hay un carril vacío a la derecha del vehículo autónomo. Si el conjunto de pruebas que utilizamos no es capaz de detectar que la reacción correcta en este caso sería incorporarse al carril vacío de nuestra derecha, entonces se ha detectado una carencia en el conjunto de pruebas, y este conjunto de pruebas debe mejorarse e incluir el nuevo caso. Se me ocurre que para este ejemplo bastaría con incluir una cláusula en el conjunto de pruebas que establezca que el coche autónomo debe "reducir la velocidad en caso de atasco e incorporarse a un carril vacío accesible si lo hay".

6. Responsabilidades en caso de accidentes

Teniendo en cuenta que el entorno en el que deben circular los coches autónomos deberá ser de uso exclusivo por parte de estos vehículos (de forma similar a como circulan los trenes por las líneas de ferrocarril), puedo definir las siguientes responsabilidades en caso de accidente:

- En caso de que el conductor humano haya desactivado el sistema de conducción autónomo y sea él el que haya tomado el control: Si se produce un accidente, la responsabilidad será de este conductor humano.
- En caso de que el accidente se produzca por un desperfecto en la infraestructura por la que circulan los vehículos autónomos: La responsabilidad si se produce un accidente en este caso es de la empresa de mantenimiento de la infraestructura y según lo estipulado en el protocolo de prevención de riesgos que se haya acordado con la institución pertinente.
- En cualquier otro caso que no sea provocado por un agente externo (como puede ser una catástrofe natural o un atentado): Desde mi punto de vista, si se produce un accidente en cualquier otro caso, y teniendo en cuenta que los vehículos autónomos además de conducirse por sí mismos mantienen

un sistema de geolocalización y son capaces de dibujar un mapa con la localización de todos los demás vehículos autónomos circulando en sus proximidades en ese instante, la responsabilidad ha de recaer sobre el equipo desarrollador del sistema.

7. Conclusiones

Creo que en estos aspectos de aplicación de sistemas de Inteligencia Artificial que tienen tanto impacto sobre la actividad cotidiana se requieren profesionales con formación multidisciplinar, capaces de interactuar entre sí manejando un lenguaje común, entendible por todos. De esta forma se pueden afrontar estos problemas tan amplios con un rigor mínimo.

Bajo mi punto de vista, el coche autónomo puede ser un gran avance si se hacen bien las cosas. En cierto sentido, creo que no será tan necesario para el transporte privado e individual, ya que entiendo que con las infraestructuras telemáticas cada vez será menos necesario trasladarse diariamente a los centros donde actualmente desarrollamos las actividades de estudio o trabajo. Sin embargo, sí pienso que tendrán mucha utilidad en el transporte de mercancías y en el transporte colectivo, ya sea público o privado.

Para mí lo más importante antes de poner en marcha un sistema de transporte autónomo es adaptar las infraestructuras urbanas y de transporte para (i) evitar accidentes, y (ii) que realmente los vehículos autónomos sean un beneficio para la sociedad.

Por último, queda señalar que en este momento del siglo XXI nos encontramos en los albores de la revolución digital de la Inteligencia Artificial que está transformando la sociedad en una dimensión similar a la transformación que supuso la revolución industrial durante los pasados siglos XIX y XX. Considero por tanto que es imprescindible estudiar la historia reciente para no cometer una y otra vez los mismos errores del pasado. De esto depende que las transformaciones sociales provocadas por estos avances tecnológicos tengan un impacto positivo en los próximos siglos.

Fuentes

Este trabajo es una adaptación parcial del artículo "The folly of trolleys: Ethical challenges and autonomous vehicles"[1] de Heather M. Roff (2018).

[1] `https://www.brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles/`

Referencias

[1] Aarian Marshall. How to Design Streets for Humansand Self-Driving Cars, 2010. https://www.wired.com/story/nacto-streets-self-driving-cars/.

[2] M. Allais. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de lécole américaine. *Econometrica: Journal of the Econometric Society*, 21:503–546, 1953.

[3] Andrew Small. The Self-Driving Dilemma, 2017. https://www.citylab.com/transportation/2017/05/the-self-driving-dilemma/525171/.

[4] K. Crawford and R. Calo. There is a blind spot in ai research. *Nature*, 538:311–313, 10 2016.

[5] A. G. Cunningham, E. Galceran, R. M. Eustice, and E. Olson. Mpdm: Multipolicy decision-making in dynamic, uncertain environments for autonomous driving. *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1670–1677, 2015.

[6] D. Ellsberg. *Risk, Ambiguity and Decision*. Routledge, 2015.

[7] Eric Madia. How Autonomous Cars and Buses Will Change Urban Planning (Industry Perspective), 2017. https://www.govtech.com/fs/perspectives/How-Autonomous-Cars-Buses-Will-Change-Urban-Planning-Industry-Perspective.html.

[8] P. Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15, 1967.

[9] M. Foucault. *The Archeology of Knowledge and the Discourse of Language*. New York, NY: Pantheon Books, 1971.

[10] M. Foucault. *Discipline and Punish: The Birth of the Prison*. New York, NY: Vintage Books, 1975.

[11] M. Foucault. *Interview with Michel Foucault on Space, Knowledge, and Power from Skyline in (Ed.) Paul Rabinow*. New York, NY: Pantheon Books, 1982.

[12] B. H. Fried. What does matter? the case for killing the trolley problem. *Philosophical Quarterly*, 62(248):505–529, 2012.

[13] H. Frowe. Ii—claim rights, duties, and lesser-evil justifications. *Aristotelian Society Supplementary Volume*, 89(1):267–285, 2015.

[14] D. Gandhi, L. Pinto, and A. Gupta. Learning to fly by crashing. *CoRR*, abs/1704.05588, 2017.

[15] P. Gómez-Abajo. Un entorno de pruebas de mutación para sistemas de control del tráfico. *Metodologías y problemas contemporáneos de la investigación científica*, 2019.

[16] Heather Roff. How understanding animals can help us make the most of artificial intelligence, 2017. http://theconversation.com/how-understanding-animals-can-help-us-make-the-most-of-artificial-intelligence-74742.

[17] James M. Anderson. Self-Driving Vehicles Offer Potential Benefits, Policy Challenges for Lawmakers, 2014. https://www.rand.org/news/press/2014/01/06.html.

[18] S. Kagan. *The Limits of Morality*. Oxford University Press, 1989.

[19] F. M. Kamm. Harming some to save others. *Philosophical Studies*, 57(3):227–260, 1989.

[20] F. M. Kamm. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford University Press USA, 2006.

[21] Michele Bertoncello and Dominik Wee. Ten ways autonomous driving could redefine the automotive world, 2015. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/ten-ways-autonomous-driving-could-redefine-the-automotive-world.

[22] M. Otsuka. Double effect, triple effect and the trolley problem: Squaring the circle in looping cases. *Utilitas*, 20(1):92110, 2008.

[23] D. Parfit. *On What Matters: Two-Volume Set*. Oxford University Press, 2011.

[24] Patrick Lin. Robot Cars And Fake Ethical Dilemmas, 2017. https://www.forbes.com/sites/patricklin/2017/04/03/robot-cars-and-fake-ethical-dilemmas/.

[25] Peter Singer. Interview in Philosophy Bites, 2010. https://philosophybites.com/2010/08/peter-singer-on-the-life-you-can-save-1.html.

[26] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[27] M. T. J. Spaan. Partially observable Markov decision processes. In M. Wiering and M. van Otterlo, editors, *Reinforcement Learning: State of the Art*, pages 387–414. Springer Verlag, 2012.

[28] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[29] J. J. Thomson. Killing, letting die, and the trolley problem. *The Monist*, 59(2):204–217, 1976.

[30] S. Thrun, W. Burgard, and D. Fox. *Probabilistic robotics*. MIT Press, Cambridge, Mass., 2005.

[31] P. Unger. Living high and letting die. *Philosophy and Phenomenological Research*, 59(1):177–181, 1999.

[32] K. Van Moffaert and A. Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *J. Mach. Learn. Res.*, 15(1):3483–3512, Jan. 2014.

[33] J. von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1947.

[34] D. West. *Securing the future of driverless cars*, pages 201–208. Brookings Big Ideas for America, 01 2017.

[35] A. Wood. Humanity as end in itself. *Proceedings of the Eighth International Kant Congress*, 1:301–319, 1995.