DOI: http://doi.org/10.12795/Argumentos/2022.i25.09

DILEMAS ÉTICOS DE LOS VEHÍCULOS AUTÓNOMOS: RESPONSABILIDAD ÉTICA, ANÁLISIS DE RIESGO Y TOMA DE DECISIONES

ETHICAL DILEMMAS OF AUTONOMOUS VEHICLES: ETHICAL RESPONSIBILITY, RISK ANALYSIS AND DECISION MAKING

JAVIER BUSTAMANTE DONAS
Universidad Complutense de Madrid
jbustama@ucm.es

RECIBIDO: 29/09/2022
ACEPTADO: 30/11/2022

Resumen: Este artículo plantea un análisis de los dilemas éticos creados en torno a los sistemas autónomos inteligentes, y más particularmente en torno a los vehículos autónomos. Se estudia el estado de la cuestión desde un punto de vista interdisciplinar. Se señalan los desafíos para la seguridad de los sistemas autónomos, así como las vulnerabilidades que afectan a la resistencia de los sistemas autónomos con respecto a ataques externos. Para tal fin se compara la habilitación para el manejo de vehículos con la certificación de sistemas autónomos. Se aplican dichos conceptos al estudio del caso del primer accidente mortal causado por un vehículo autónomo en relación con la asignación de responsabilidad ética y legal. Por último, se apunta a la tarea pendiente del análisis de la relación entre la ética de los vehículos autónomos y la teoría de juegos.
Palabras clave: Vehículos autónomos; Ética, Toma de decisiones; Ética de las máquinas; Ética de la ingeniería.

Abstract: This article presents an analysis of the ethical dilemmas created around intelligent autonomous systems, and more particularly around autonomous vehicles. The state of the art is studied from an interdisciplinary point of view. The challenges for the security of autonomous systems are pointed out, as well as the vulnerabilities affecting the resilience of autonomous systems

with respect to external attacks. For this purpose, vehicle operation licensing is compared to the certification of autonomous systems. These concepts are applied to the case study of the first fatal accident caused by an autonomous vehicle in relation to the assignment of ethical and legal responsibility. Finally, the pending task of analyzing the relationship between the ethics of autonomous vehicles and game theory is addressed.
Keywords: Autonomous vehicles; Ethics; Decision making; Machine Ethics; Engineering Ethics.

Definiciones y conceptos básicos sobre conducción autónoma

Se define como *vehículo autónomo* aquel que puede circular sin intervención humana por vías que no han sido diseñadas específicamente para este tipo de vehículos y por las que circulan colectivamente ciclistas, peatones, y otros usuarios de vehículos. Emplean sistemas de automatización que llevan a cabo total o parcialmente la llamada *tarea de conducción dinámica* (DDT)[1]. Estos sistemas dependen a su vez de los inputs que reciben a partir de otros subsistemas: sensores ultrasónicos, sensores infrarrojos, sistema inercial, sistemas de posicionamiento y navegación por satélite, sistemas de cámaras de visión artificial, sistemas de radar y sistemas de determinación de distancia a través de láser (LIDAR)[2]. Todos los vehículos se catalogan en una taxonomía de seis niveles según el grado de intervención de sistemas autónomos para asistir a la conducción humana. Las especificaciones técnicas de estos niveles de automatización han sido establecidas por un organismo global llamado *Sociedad de ingenieros de automoción* (SAE) en el documento J3016 publicado en 2014, cuya última revisión es del 30 de abril de 2021 (SAE, 2022). Los niveles de automatización hacen referencia al papel que desempeña cada uno de los tres

[1] DDT: Dynamic driving task.
[2] LIDAR: Laser imaging detection and ranging

agentes principales (el conductor, el sistema de automatización de la conducción, otros sistemas y componentes del vehículo) en la ejecución de la DDT habitual, o en la DDT de emergencia.

Estos niveles son los siguientes (Km77, 2022; SAE, 2022):

- Nivel 0: *ningún nivel de automatización*, conducción totalmente manual sin asistentes para el control en marcha del vehículo.
- Nivel 1: *conducción asistida* por dispositivos que pueden ayudar al control longitudinal o lateral del vehículo, pero no los dos a la vez. Por ejemplo, el programador de velocidad o el asistente de aparcamiento.
- Nivel 2: *automatización parcial de la conducción*. Asistencia a la conducción en ambos ejes al mismo tiempo, pero de forma limitada y dejando al conductor la responsabilidad de responder en situaciones de detección de peligros imprevistos. Por ejemplo, el sistema de mantenimiento dentro del carril y el asistente de conducción en atascos.
- Nivel 3: *automatización condicional de la conducción*. El conductor puede ceder el control al sistema autónomo, pero puede desactivar a voluntad dicho sistema frente a una situación de riesgo. Cuando el sistema detecta que se enfrenta a condiciones que le superan, avisa para que el conductor vuelva a tomar el control.
- Nivel 4: *alta automatización de la conducción*. El vehículo opera de forma automática de forma continuada, sin la expectativa de intervención por parte del conductor. El sistema está preparado para responder a situaciones de peligro imprevistas y escoger la acción más favorable, es decir, que propicie la denominada *situación de mínimo riesgo*. El conductor siempre puede solicitar recuperar el control del vehículo en cualquier momento, aunque el

vehículo puede no responder inmediatamente. Si es el vehículo el que solicita que el conductor retome el control y éste no lo hace, el sistema es capaz de detener el vehículo en una zona segura.

- Nivel 5: *automatización total de la conducción*. El sistema de conducción automatizada (ADS)[3] está diseñado para circular por cualquier vía que sea apta para un conductor humano en las mismas condiciones ambientales, incluidas las climatológicas. Por lo tanto, el vehículo puede carecer de pedales y/o de volante, o prescindir de un conductor de seguridad.

Los niveles se aplican a las funciones que están activadas durante la conducción. Por lo tanto, un vehículo de un nivel más alto puede tener desactivadas algunas funciones y actuar dentro de un nivel inferior.

Las ventajas de la generalización de los vehículos autónomos han sido recogidas por la *Organización Internacional de Constructores de Automóviles* (OICA), y son las siguientes: Se evitarían los errores humanos que causan más del 90 % de los accidentes de tráfico, por lo que aumentaría significativamente la seguridad vial. Se optimizaría la organización del tráfico mejorando el flujo de vehículos, con la consiguiente reducción de emisiones contaminantes. Se extendería el beneficio del uso del automóvil a conductores con necesidades especiales o inseguros, y a personas de edad avanzada. Por último, el ahorro de tiempo empleado en conducir supondría mayor productividad y disponibilidad de tiempo de ocio.

No obstante, también surge la necesidad de diseñar un marco ético y legal complejo con relación a la responsabilidad de los accidentes provocados por sistemas de conducción autónoma. También surgen dilemas éticos con relación a otros factores como

[3] ADS: Automated driving system.

la programación de la toma de decisiones en situaciones donde la mejor opción posible implica coste de vidas humanas, o la transparencia de los algoritmos empleados y la certificación de los sistemas de inteligencia artificial y las redes neuronales de aprendizaje profundo que se encuentran en el corazón del sistema autónomo. En los siguientes apartados iremos desentrañando estos dilemas éticos.

La deliberación ética acerca de los sistemas autónomos inteligentes. Estado de la cuestión

Los vehículos autónomos son una clase particular de sistemas autónomos inteligentes. Los sistemas autónomos inteligentes son aquellos que interaccionan con su entorno para adaptarse a las circunstancias dinámicas cambiantes. Ya sean barcos autónomos, vehículos de conducción automática o drones y otros tipos de aviones civiles o militares no tripulados, se engloban en uno de los campos de mayor éxito de las tecnologías de aprendizaje máquina e inteligencia artificial. La adopción de nuevos modelos de computación en nube, de robótica corporativa, de plataformas de arquitectura de red, etc., está generando toda una constelación de problemas éticos que tienen que ver no solamente con el uso de dichos artefactos, sino también con el diseño, el modelado, la verificación y la validación de los sistemas autónomos de complejidad funcional creciente tanto en escala como en prestaciones. Sin duda esta tendencia hacia una complejidad mayor tanto en estructura como funcional comienza a presentar un gran desafío de diseño y de validación para garantizar tanto la seguridad como la certificación de dichos sistemas con respecto a estándares técnicos y éticos aceptables en una sociedad democrática. (Yu et al., v-vi).

En los últimos años se han publicado un gran número de artículos académicos y *libros blancos* dedicados a orientar la investigación de las diferentes comunidades implicadas, como son las que trabajan en el control de sistemas, el procesamiento de señales, la visión artificial, el diseño de circuitos, la validación y verificación, la robótica corporativa y la inteligencia artificial. Sin embargo, este florecimiento de los enfoques especializados no debe olvidar que a menudo los dilemas éticos se generan en territorios de intersección de estas disciplinas, de forma que empezamos a percibir la necesidad de una visión holística y unificada del estado de la cuestión para poder responder a estos dilemas. Al ser un área de naturaleza fuertemente interdisciplinaria, los dilemas éticos generados tienen que ver también con avances en el estado de la cuestión producidos por la colaboración de investigadores industriales y académicos. En estos casos, no basta simplemente un abordaje deontológico, ya que se ponen en juego valores de eficacia y seguridad que no obedecen a conceptos ideales, sino a criterios y niveles de seguridad y fiabilidad aceptables por la sociedad. Esto implica que la problemática ética de los vehículos autónomos entendidos como un caso particular de sistemas autónomos inteligentes pone en acción una constelación de valores que no son solamente éticos, y que tienen que ver con criterios consecuencialistas o prudenciales, como veremos más adelante (Yu et al., v-vi).

Uno de los objetivos de este artículo será destacar la necesidad de establecer criterios éticos y protocolos de evaluación aplicables a los desafíos relacionados con la seguridad de la nueva generación de vehículos autónomos basados en el aprendizaje máquina, la aplicación de la computación en nube y la inteligencia artificial. Nuevas metodologías de análisis ético deben ser desarrolladas para orientar el trabajo colaborativo de las comunidades de tecnólogos y científicos que están trabajando en proyectos que no solo potencian

la capacidad humana para interactuar con su entorno, sino que redefinen las tareas hasta hace poco tiempo consideradas exclusivamente humanas.

Si bien es cierto que los vehículos de conducción autónoma llevan ya algunos años entre nosotros, el continuo avance en tecnología de sensores, aprendizaje de máquina visión artificial y modelos de toma de decisiones, está abriendo nuevas oportunidades para ampliar el uso de estos sistemas, pero también aumenta su complejidad y nivel de acoplamiento. En consecuencia, la discusión ética sobre sus condiciones de utilización se encuentra con requerimientos de seguridad cada vez más complejos para los cuales las metodologías convencionales de diseño y validación no están convenientemente preparadas para responder a los grandes desafíos de seguridad y consistencia que requieren estos sistemas. La opinión pública ha generado una particular expectación acerca de estos dilemas sobre la seguridad de las operaciones y decisiones de un vehículo autónomo a través de accidentes con graves consecuencias que lo ocurrido recientemente implicando a vehículos de Tesla y Uber.

Desafíos para la seguridad y resistencia de los sistemas autónomos

El número de desafíos que tienen relevancia ética para esta tecnología es muy amplio. El primero de ellos tiene que ver con las implicaciones para la seguridad de los sistemas autónomos sometidos a ciberataques desde la perspectiva del control de sistemas. Y el problema aquí se sitúa en conseguir un diseño de sistemas de control resistente a ciberataques que no sea excesivamente conservador. Cuando se evalúan los llamados *controladores híbridos* se verifican de forma analítica

exclusivamente para casos de horizonte de tiempo finito. Sin embargo, los resultados deben posteriormente extenderse al caso de horizonte de tiempo infinito para poder generalizar los resultados y garantizar así un satisfactorio nivel de protección frente a ciberataques (Kwon y Hwang, 2019).

El segundo problema tiene que ver con las metodologías de validación de sistemas autónomos de visión artificial en condiciones no ideales, es decir, circunstancias caracterizadas por la variación de las condiciones ambientales -- especialmente la temperatura y la visibilidad en situaciones de niebla --, y también por el envejecimiento de los circuitos. En estos casos el problema está en la imposibilidad de validar la robustez de estos sistemas en escenarios reales de casos más desfavorables, ya que resulta muy complicado diseñar experimentalmente *la peor situación posible*. Las investigaciones en este terreno se orientan a diseñar algoritmos, particularmente del tipo *subset sampling* (SUS), para obtener una estimación objetiva de la estadística de fallos raros utilizando a la menor cantidad posible de datos de test, con el objetivo de optimizar los costes de validación, mejorando los abordajes poco eficientes basados en el uso del método Montecarlo de fuerza bruta. (Handi Yu et al., 2019).

Un tercer problema tiene que ver con dos desafíos básicos en el diseño de sistemas autónomos seguros y confiables: la dinámica inadecuada del modelo del sistema y los entornos solo parcialmente observables. Tanto el aprendizaje máquina como la inteligencia artificial tienen un enorme potencial de transformación de los sistemas autónomos, dado que utilizan el diseño basado en datos (*data-driven design*). En este sentido, el análisis de las experiencias y de las consecuencias de los procesos de toma de decisiones en escenarios realistas es la principal herramienta de evolución. Sin embargo, dado el carácter crítico de la seguridad de los sistemas de vehículos autónomos, esta evolución requiere abordajes que sean al

mismo tiempo seguros, resistentes y robustos. Dada la cantidad de información que debe ser procesada por el sistema autónomo del vehículo en cada desplazamiento medio, tasas ínfimas de error dan lugar a cifras difícilmente soportables de accidentes y pérdida de vidas humanas (Bansal y Tomlin, 2019).

Resistencia de los sistemas autónomos con respecto a ataques externos

Un cuarto problema tiene que ver con la seguridad con respecto a ataques externos a los vehículos autónomos. Incluso cuando es posible garantizar el adecuado funcionamiento de los sensores en todo tipo de circunstancias climáticas y ambientales, aparte del desgaste y envejecimiento de los componentes de los subsistemas sensoriales, existe el problema de la vulnerabilidad de la integridad de los sistemas autónomos a las intervenciones externas, conocidas como *hacking*. El punto más vulnerable de dichos sistemas se localiza en su red neuronal profunda (*Deep Neural Network*, o DNN), que se ocupa fundamentalmente de interpretar y dotar de sentido a los estímulos o inputs captados por los sensores del vehículo. Varios artículos científicos recientes han demostrado la vulnerabilidad de las DNN y frente a pequeñas perturbaciones visuales en los inputs del sistema (Carlini et al., 2017; Carlini et al., 2020; Gandhi et al., 2020; Goodfellow et al., 2014; Kos et al., 2017; Li et al., 2014; Moosavi-Dezfooli et al., 2016; Neekhara et al., 2021; Nguyen et al. 2015; Papernot et al. 2016). Dado que los vehículos autónomos utilizan este tipo de redes neuronales para tomar decisiones en circunstancias críticas de seguridad, una interpretación errónea de información por parte de los sensores podría provocar decisiones erróneas y generar escenarios peligrosos. Estas técnicas pueden ser utilizadas para sabotear de

manera intencional los vehículos autónomos, y se convierten en una prueba de fuego para el desarrollo de algoritmos de aprendizaje máquina resistentes a ataques. Eykholt et al. (2018) proponen un algoritmo general de ataque llamado *Robust Physical Perturbations* (RP_2), que genera errores de lectura de los sensores bajo ciertas condiciones ambientales pruebas de laboratorio y de Campo. Utiliza como estudio de casos la perturbación creada por pequeñas pegatinas sobre señales de tráfico que provocan clasificaciones erróneas selectivas en el 100 % de las pruebas de laboratorio, y en el 84,8 % de las imágenes de vídeo capturadas por un vehículo en movimiento con respecto a una señal de tráfico en una prueba de campo.

El estudio citado se centra en la clasificación de las señales de tráfico por parte de los sistemas de aprendizaje profundo de vehículos dotados de sistemas de visión artificial. Realmente el reconocimiento de señales de tráfico se convierte en la *Drosophila Melanogaster* de los retos que enfrenta la visión artificial, por varias razones. En primer lugar, la relativa simplicidad visual de las señales de tráfico dificulta ocultar un sabotaje. En segundo lugar, las señales de tráfico forman parte de un entorno de condiciones físicas cambiantes, como pueden ser los cambios de distancia hay de ángulo de la cámara del vehículo según se aproxima a la señal. En tercer lugar, en este caso el saboteador no necesita tener acceso al propio vehículo, sino que se ocupa de alterar objetos en el mundo físico que el vehículo utiliza como base para tomar decisiones cruciales sobre seguridad. También encontramos un desafío importante, que consiste en que las alteraciones en el mundo digital pueden ser tan pequeñas que una cámara no pueda percibirlas debido a las imperfecciones del sensor. El estudio de Eykholt et al. concluye demostrando que un saboteador potencial puede modificar físicamente señales de tráfico utilizando técnicas de bajo coste que pueden provocar errores fatales de clasificación

en los sistemas de aprendizaje profundo y, que son cruciales en la operatividad de los vehículos autónomos.

En concreto, su ataque provocó que una señal de Stop modificada con cuatro pegatinas, dos blancas y dos negras, fuera interpretada por el sistema de visión artificial del vehículo como una señal de límite de velocidad de 45 millas por hora. Además, la colocación de las pegatinas simulaba ser un grafiti, forma habitual de vandalismo o arte urbano, según como se mire, tan frecuente hoy en día. Además de las fatales consecuencias que tiene un ataque de este tipo, ya que provoca que el vehículo no frene cuando deba hacerlo sino que continúe acelerando su marcha hasta alcanzar el límite de velocidad supuestamente permitido por la señal jaqueado, la peligrosidad consiste en que resulta muy difícil identificarlo como un sabotaje intencional y no despierta por lo tanto sospechas. Este estudio demuestra finalmente la vulnerabilidad de los sistemas de visión artificial cuando son los objetos físicos en sí mismos son saboteados, aunque el sistema mantenga su integridad de operación. El sabotaje de las redes neuronales profundas tiene que ver con el problema de la manipulación facial de imágenes y vídeos para difamar a una persona o para crear dinámicas de desinformación, fenómeno conocido como *DeepFake* (Dolhansky et al., 2020) que crea a su vez un nuevo abanico de cuestiones éticas y legales.

Habilitación para la conducción de vehículos vs. certificación de sistemas autónomos

Un quinto problema tiene que ver con la similitud existente con la certificación que tanto conductores de coches, autobuses y camiones como pilotos de aeronaves o capitanes de barco necesitan para demostrar su capacitación en el uso de dichos vehículos. Estos

procesos de certificación suelen incluir el examen médico general, examen detallado de agudeza visual, un examen teórico acerca de la legislación vigente aplicable y de casos teóricos, y por último un examen práctico y situaciones reales. A través de este proceso un examinador debe certificar si es candidato humano está capacitado para subir el nivel de responsabilidad que supone la conducción de un vehículo de transporte. Dada la importancia de las tecnologías de transporte autónomo y el necesario debate sobre si el razonamiento probabilístico que caracteriza a estas tecnologías debe ser certificado, debemos preguntarnos si dicho sistema de licencias de habilitación para conductores humanos de vehículos de transporte debe extenderse también a los vehículos autónomos.

Cummings (2019) responde a esta pregunta y explora este problema comparando los procesos de habilitación para vehículos de superficie y aeronaves comerciales, extrayendo patrones que nos guíen a la hora de establecer parámetros y criterios para la homologación de sistemas autónomos, y este apartado continúa y amplía su línea de argumentación. A pesar de que el grado de dificultad de los exámenes de habilitación está en función de la complejidad de operación de un vehículo así como del número de pasajeros y el riesgo total en función de los dos parámetros, la obtención de licencias tiene una estructura similar en todos los casos: condiciones físicas y de salud, examen escrito sobre conocimientos operacionales y legislativos, y un examen práctico en el que un evaluador acompaña al candidato en situaciones prácticas del mundo real. Al estudiar la ética de los vehículos autónomos, nos debemos plantear si este proceso de obtención de licencias y habilitación profesional puede arrojar luz sobre el proceso de certificación de los sistemas autónomos. Entendemos por sistemas autónomos aquellos vehículos con nivel de automatización 4 o 5 según está definido por el estándar SAE J3016, citado al comienzo de este artículo. La certificación de estos

vehículos sería un proceso análogo al de la habilitación para los operadores humanos. Sin embargo, existe una diferencia fundamental. Mientras que el examen de habilitación se aplica a un individuo, la certificación del hardware y el software que componen un sistema autónomo se aplicaría al grupo homogéneo de vehículos que compartiera dichos dispositivos. Mediante tal certificación se obtendrían garantías de que dicho vehículo autónomo puede ser operado de manera segura y efectiva por los subsistemas computacionales de abordo sin necesidad de supervisión o intervención humana.

En los Estados Unidos, la obtención de una licencia de conducción de automóviles. comienza por un examen de la vista, ya que está demostrado que es el sentido responsable del 95 % de los estímulos recibidos en la conducción del automóvil (Shinar y Schieber, 1991). Otras condiciones de salud que pueden ser limitantes se consideran Y resuelven caso al caso. El examen teórico y se concentra en la legislación de tráfico, las normas de circulación y las prácticas de conducción segura. La tercera parte es un examen práctico acompañado por un examinador en circunstancias de tráfico real. Este test evalúa la capacidad del examinado para conducir un automóvil de manera segura, y evalúa su capacidad perceptiva, de atención y motora. En la Unión Europea y en otros países de nuestro entorno, la obtención del carnet de conducir obedece a parámetros similares.

La obtención de la licencia de aviación comercial tiene una estructura similar. El examen físico debe demostrar la capacidad para utilizar y entender adecuadamente los sensores a través de los que el piloto obtiene información del mundo exterior, y la capacidad de tomar decisiones adecuadas a tiempo. Deben pasar también un examen que demuestre que conoce y sabe interpretar los elementos de gestión de situaciones de riesgo definidos en el *Airman Certification Standard* (ACS). Son en total no menos de 60

tareas en las que debe demostrar su capacidad y suficiencia. Una de ellas es, por ejemplo, desviar la aeronave a otro aeropuerto cuando el destino previsto no está disponible por causas meteorológicas o de cualquier otro tipo. Mediante un examen oral los evaluadores plantean cuestiones basadas en escenarios concretos, ya que no existen procedimientos que puedan agotar todas las contingencias o emergencias que se pueden presentar en la vida real. Una vez que el candidato ha aprobado tanto el examen escrito como el oral, el examen práctico de vuelo evalúa su habilidad y para llevar a cabo tareas y maniobras estándar, además de la capacidad para actuar mitigando el riesgo en situaciones de emergencia simulada. Luego veremos que este tipo de evaluación interactiva del manejo del riesgo en escenarios prácticos será un factor clave en la certificación de sistemas autónomos.

La *taxonomía SRKE* (skills, rules, knowledge, and expertise)[4] (Cummings, 2014) muestra por qué la certificación de los vehículos autónomos se parece mucho más a las licencias de piloto que las licencias de conducción de automóviles. En situaciones de incertidumbre, la fortaleza de un sistema computacional tiende a mostrarse en las habilidades y el seguimiento de las reglas, mientras que los humanos son más capaces cuando entran en juego el conocimiento y, sobre todo, la pericia. El comportamiento basado en capacidades y en reglas suelen ser acciones motoras y sensoriales y entremetido automáticas, que los humanos adquieren

[4] Habilidades, reglas, conocimiento y pericia. Hace referencia a los tipos de comportamiento que son necesarios con relación al grado de incertidumbre y en una situación de riesgo. Por ejemplo, para ir del punto A al punto B un vehículo autónomo puede seguir las indicaciones de Google Maps aplicando sus habilidades con un alto grado de automatismo. Pero cuando las condiciones meteorológicas son adversas, hay un accidente en la ruta o el tráfico está complicado, la simple habilidad mecánica o el seguimiento de las reglas estándar deja paso a la aplicación sabia de un conocimiento basado en la interpretación de la situación y en la pericia humana.

a través del entrenamiento. Por ejemplo, controlar una aeronave comercial del vuelo que es una tarea altamente automatizada, que requiere habitualmente de 3 a 7 minutos de intervención manual por parte del piloto, independientemente de la duración total del vuelo. Dado que el piloto está sometido a factores humanos como la fatiga, la distracción, la falta de atención y otros problemas neuromusculares, resulta más fiable dejar el gobierno de la nave en manos del piloto automático en situaciones donde las habilidades estandarizadas tienen el papel preponderante. Las situaciones que exigen un comportamiento basado en reglas suponen tareas cognitivas más complejas. Aquí es necesario interpretar a una situación concreta, sobre todo cuando existe un fallo múltiple del sistema, para decidir qué procedimiento debe seguirse. Dada la estructura *si P, entonces Q* de las reglas, los sistemas autónomos tienen aparente ventaja. Sin embargo, cuanto mayor es el riesgo de incertidumbre más se necesita la capacidad humana de pericia y discernimiento. En situaciones computables o determinísticas los algoritmos muestran grandes ventajas frente a la variabilidad del comportamiento humano, pero solamente pueden tomar en cuenta aquellos factores que estén definidos como argumentos críticos de una ecuación. En un sistema complejo con un alto grado de incertidumbre inherente, como es el tráfico rodado, no es posible prevenir todas las condiciones variables que pueden intervenir en la toma de decisiones por parte del vehículo autónomo. Por ejemplo, en hora punta o a la salida del trabajo los conductores suelen demostrar un alto grado de ansiedad y realizan maniobras más impredecibles que de costumbre, y en situaciones meteorológicas adversas la lluvia o la niebla pueden crear situaciones peligrosas en cuestión de segundos.

Al igual que ocurre con los humanos, los vehículos autónomos también dependen de la percepción de los estímulos para saber qué procedimiento debe ser aplicado. Por lo tanto, un sistema autónomo

que esté basado en procedimientos muy sofisticados dependerá siempre de la precisión y fiabilidad de sus sensores, ya que a partir de ellos obtienen la información del mundo exterior que está en la base de sus procesos de toma de decisiones. Ya que en ética el problema de la responsabilidad está siempre a la base, algunos ejemplos nos demuestran la relación entre percepción del mundo y toma de decisiones. El conductor de un Tesla falleció porque confió en el sistema automático de detección frente a obstáculos cuando un tráiler se cruzó en su trayectoria. El sistema estaba programado para detenerse frente a un obstáculo de ese porte, el problema es que el sensor no funcionó adecuadamente provocando el accidente mortal. (NTBS 2017).

Cuanto mayor es el grado de incertidumbre, en mayor medida tiene que ser sustituido el conocimiento deductivo por otro inductivo, ya que no contamos con todos los elementos de juicio necesarios. Aquí más que en cualquier otro escenario se cumple la máxima de Samuel Butler: *la vida consiste en sacar conclusiones de premisas insuficientes*. En casos como el de las señales de tráfico hackeadas, los humanos actuamos en función de intuiciones y juicios situacionales, basados habitualmente de experiencia. Esta es la característica del nivel de experticia o maestría, en el que no es suficiente una base de conocimiento si no está apoyada en la experiencia previa en situaciones de incertidumbre similares. Este nivel de conocimiento experiencia tiene mucho que ver con lo que tradicionalmente se entiende como sabiduría, y es un terreno muy alejado aún del alcance de la inteligencia artificial, a pesar de los progresos realizados hasta el momento. Serán necesarios avances cualitativos para poder obtener niveles satisfactorios de imitación del comportamiento humano en este tipo de situaciones límite.

Viendo estos antecedentes, Cummings aboga por la conveniencia de aplicar un sistema de certificación de vehículos autónomos que aprenda de la experiencia obtenida con los

procedimientos de obtención de licencias para conductores de vehículos terrestres o pilotos de una vez. De la misma forma que el chequeo de la vista es fundamental para aún operador humano, la certificación de los sistemas de visión artificial son un elemento clave en los vehículos autónomos. No es solamente un problema de sabotaje externo malintencionado, como hemos visto en el caso estudiado por Eykholt, sino también de otras causas medioambientales, como es una situación de niebla, de lluvia o de nieve, de baja visibilidad, de deslumbramiento cuando está el sol a baja altura en el amanecer o en el ocaso, de viento que arrastre hojas que queden pegadas a las señales de tráfico impidiendo su correcta identificación, etc. En los viajes nocturnos por carretera es habitual ver al llegar a destino que los faros y el parabrisas quedan manchados por una multitud de mosquitos y otros insectos que son atraídos por las luces del vehículo y quedan incrustados. Tampoco son infrecuentes las defecaciones de los pájaros sobre la carrocería del vehículo, sobre todo cuando aparcamos debajo de un árbol. Factores tan triviales pueden bloquear un sensor óptico o provocar fallos de identificación de elementos esenciales como las señales de tráfico, peatones y otros vehículos, para el desarrollo de una conducción segura. Por último, fallos eléctricos en el vehículo o el envejecimiento de componentes ópticos o electrónicos puede ser importantes factores de riesgo. Como ya hemos indicado, la certificación debería establecerse a dos niveles. Por un lado, de forma genérica para un solo componente del vehículo autónomo, y por otro individualmente para cada vehículo en forma de inspección periódica.

En cuanto al equivalente de los exámenes teóricos, la certificación para los vehículos autónomos debería contemplar simuladores que replicaran situaciones aleatorias de riesgo tanto para la seguridad de los pasajeros y el conductor del vehículo como para los peatones y conductores y pasajeros de otros vehículos. Se

debería prestar atención especial a los llamados *elementos de mitigación de riesgo*. Aquí de nuevo el problema principal es de índole ética. Tenemos que definir el umbral de riesgo que consideramos aceptable, y también la respuesta *suficientemente buena* por parte del sistema autónomo. Es decir, una respuesta que minimice los daños dentro del abanico de posibilidades reales, aceptando que ni humanos ni máquinas pueden tener un conocimiento exhaustivo de todas las condiciones implicadas en la situación, por lo que si hace necesario definir el umbral de riesgo que estamos dispuestos a aceptar. En teoría de la decisión se utiliza un concepto de *satisfactorio* para calificar este tipo de decisiones.

En términos prácticos, se suele establecer que un sistema autónomo funciona adecuadamente cuando genera *niveles de seguridad equivalentes o mejores* (EBLS)[5] a los que se exigen a los operadores humanos, especialmente en situaciones poco probables pero de consecuencias potencialmente graves. El problema que nos enfrentamos para establecer estos estándares de EBLS es la opacidad del funcionamiento del software del vehículo autónomo. Los algoritmos de aprendizaje profundo y las redes neuronales son muy complicados de entender, y a ello se suma en muchos casos el carácter propietario del software, que impide la llamada *ingeniería inversa*. Es decir, que estemos autorizados a decodificar el programa inspeccionando lo que realmente hace su código fuente. El problema se resolvería en gran parte exigiendo que toda la programación de los subsistemas del vehículo obedezca a estándares de software libre. Otro camino sería el desarrollo de algoritmos transparentes y modalidades de visualización e interacción que permitan a los especialistas conocer y entender cómo el vehículo genera planes de acción y de toma de decisiones que realmente cumplan los niveles de EBLS. En caso contrario resulta muy difícil establecer controles para garantizar el

[5] ELBS: Equivalent or better levels of safety.

cumplimiento de las prescripciones éticas que debe cumplir el vehículo autónomo para obtener su certificación. (Cummings, 2019, p. 147).

De la misma forma en que los conductores noveles deben llevar en el vehículo una señal que indique a los demás conductores que se encuentran en periodo de formación (la famosa "L" blanca sobre fondo verde) sería ético que los vehículos autónomos utilizaran algún indicativo para prevenir a los demás conductores de su presencia en la carretera. Al menos en esta etapa de desarrollo en la que aún no han alcanzado un nivel satisfactorio de seguridad. Prueba de ello es que en el estado de California están autorizados a circular sin que haya un conductor dentro del vehículo, siempre que haya un operador que controle remotamente el vehículo y que pueda intervenir telemáticamente cuando así lo requiera una situación de peligro. Sin embargo, este tipo de operaciones sufren de un defecto congénito. Se trata del desfase en la reacción del teleoperador, ya que evitar un accidente requiere frecuentemente tiempos de reacción de una fracción de segundo. Al actuar a distancia el teleoperador necesita un tiempo extra para tomar control del vehículo y tomar las decisiones oportunas, para después actuar de forma motora sobre los dispositivos del vehículo autónomo. Además de este tiempo extra necesario para tomar el control del vehículo hay que tener en cuenta la limitación física llamada *desfase neuromuscular*, que consiste en un retardo de 500 milisegundos entre la aparición del estímulo y la capacidad de reaccionar. Una prueba de ello y es la gran cantidad de accidentes de los drones controlados a distancia. En el caso del avión no tripulado *Predator*, la fuerza aérea norteamericana ha perdido un tercio de su flota por fallos de operación humana, principalmente en maniobras de aterrizaje. Estas maniobras se realizan a velocidades similares a las de un automóvil que circula por una autopista. Por lo tanto, es muy probable que los vehículos

autónomos, dado el estado actual de la tecnología, sufren también de las mismas dificultades para que el teleoperador consiga evitar un accidente. El camino consistiría en el desarrollo de interfaces más sofisticadas de realidad Virtual para que los teleoperadores tuvieran una percepción más inmersiva y realista, diseñando sensores táctiles y olfativos que complementen los estímulos visuales y auditivos. Haciendo intervenir más sentidos se consigue una mejor percepción e interpretación de la situación de peligro.

Una segunda fase consistiría en la implantación subcutánea de una interfase que permitiera que dichos estímulos se dirigieran directamente al sistema nervioso del teleoperador. De esta manera se reduciría la distancia que tendrían que recorrer los impulsos eléctricos de su respuesta motora. No obstante, este tipo de intervenciones presenta nuevos dilemas éticos que deben ser tenidos en cuenta, que tienen que ver con la integridad del cuerpo de la persona. En vehículos autónomos de nivel 5 se elimina la esta fuente de error, ya que no existe ningún tipo de intervención humana, sino que la conducción se realiza con un altísimo nivel de conectividad con el entorno, a través del reconocimiento de señales, la información ambiental contenida en la nube, los sistemas de geoposicionamiento de alta precisión, y todo tipo de tecnologías que requieren un nivel de acoplamiento muy elevado. No olvidemos que en cualquier sistema de alto acoplamiento, cualquier fallo de un componente se contagia rápidamente al resto de componentes, provocando un fallo múltiple de difícil resolución.

Sin embargo, aparece otra fuente de error que puede ser más grave. En el sector del automóvil cada vez tiene mayor importancia los fallos de software. En vehículos convencionales que no tienen sistemas de inteligencia artificial, estos fallos han sido responsables de más de un 19 % de las llamadas a fábrica para reparación (*recalls*) en 2019. En vehículos más sofisticados el problema es mucho más grave. Entre 2016 y 2022 la compañía BMW llamar a

revisión a más de 60.000 vehículos de los modelos X3, X4 y serie 5 debido a un fallo de software que provocaba la pérdida de torque y la parada del motor. A su vez, Tesla ha tenido que llamar a fábrica más de 130.000 vehículos en Estados Unidos debido a un fallo de software que provocaba que la pantalla de infoentretenimiento no mostrara imágenes de la cámara trasera, la selección de marcha el estado de los limpiaparabrisas y de las luces de aviso, incrementando así el riesgo de sufrir un accidente. En 2019, solo en Estados Unidos más de 10 millones de vehículos han sufrido estos fallos, y el 51,6% no han pasado por fábrica y circular sin haber resuelto el fallo. (Sibros, 2020). A medida que los automóviles cada vez tienen más elementos de conectividad, aumenta también la cantidad de líneas de código de programación necesarias para soportar este aumento de capacidades inteligentes. Esto conlleva un aumento exponencial del riesgo de fallo de programación. En este panorama se plantean nuevos dilemas como el que afecta a la política de la compañía Tesla de actualizar automáticamente el software de sus vehículos sin consentimiento ni conocimiento previo del dueño del vehículo. A medida que se introduzcan a niveles superiores de conducción autónoma, parece una cuestión moralmente delicada el hecho de que el fabricante actualice sin consentimiento un software que permite funcionales desconocidas para el usuario, o asociando la descarga de la actualización a la aceptación tras la lectura de las condiciones de uso, cuando todos sabemos que casi nadie lee dichas condiciones antes de instalar una aplicación en su teléfono móvil o en su computadora. Hoy en día no existen test homologados y estandarizados que puedan evaluar adecuadamente el software basado en una racionalidad estocástica y no determinista, como es el caso del aprendizaje máquina, principalmente en entornos críticos de seguridad con alto grado de incertidumbre. Ese hecho permite prever un aumento muy significativo de este problema en un futuro próximo. Y este

problema se suman otros que quedan claramente puestos en el caso del primer accidente con consecuencias mortales en el que se visto involucrado un vehículo autónomo. Lo veremos a continuación.

El primer accidente mortal de un vehículo autónomo: ¿quién (o qué) fue responsable?

El 18 de marzo de 2018 un SUV Volvo de conducción autónoma atropelló a un peatón a una velocidad de 39 millas por hora (63 km/h), a pesar de la presencia de un conductor de seguridad o *back-up driver*. La misión de este conductor es tomar el control del vehículo en el caso de que éste tome una decisión errónea que ponga en peligro la vida del propio conductor, la de otros conductores o peatones, o la propia integridad del vehículo. El accidente tuvo lugar en la localidad de Tempe, Arizona, cuando Elaine Herzberg, de 49 años, se cruzó en la carretera con su bicicleta y fue arrollada por el vehículo de la compañía Uber. La investigación alegó que la conductora de seguridad, Rafaela Vasquez (nacido Rafael), estaba en el momento del accidente distraída mientras asistía a un programa del concurso de televisión *La voz*. Los datos de streaming de la plataforma Hulu demostraron que su teléfono móvil estaba visualizando dicho programa. Se trata del primer atropello con resultado de muerte en el que está implicado un automóvil autónomo, y es un caso paradigmático para poder analizar los dilemas éticos que se plantean acerca de la responsabilidad de los accidentes en los vehículos autónomos. Fue un día fatídico para la industria del automóvil y la percepción pública de la seguridad de la nueva tecnología de vehículos autónomos

En su defensa Rafaela Vasquez declaro que la mujer saltó a la carretera justo delante de ella montando en una bicicleta y llevando

varias bolsas de compra, y que no tuvo ninguna posibilidad de frenar. Su velocidad en ese momento era de 40 millas por hora, por lo que no superaba el límite de 45 millas en ese punto de la carretera. Sylvia Moir, jefe de policía de Tempe, declaró que "Está muy claro que habría sido difícil evitar esta colisión en cualquier tipo de modo (autónomo o conducido por humanos) basándose en cómo salió de las sombras directamente a la calzada". Fuentes de la policía local también señalaron que los vídeos grabados por las cámaras del vehículo de Uber apoyaban la versión de Vasquez (Styles, 2018).

Actualmente se encuentra a la espera de juicio, quizá por la complejidad del caso y también por las repercusiones que una sentencia en un sentido u otro puede tener para la industria del automóvil. No hay que olvidar que a partir de 2015 Uber competía con Google en la carrera por el vehículo autónomo, planteándose el desarrollo de los *robotaxis* como una batalla existencial. Efectivamente, contar con una flota de taxis sin conductor reduciría enormemente los costes de operación, lo que daría una ventaja estratégica sustancial a la compañía que fuera la primera en conseguir dicha gesta. Rafaela Vasquez se enfrenta a una acusación de homicidio por negligencia. El fiscal solicita una pena de 2,5 años de prisión, similar a la de un homicidio involuntario. A la hora de cerrar este artículo todavía no hay fecha prevista para el juicio.

Este caso muestra claramente dónde se encuentra el eslabón más débil del sistema legal, más allá de cualquier consideración ética. Lo que está en juego es la asignación de responsabilidad civil o penal a los agentes involucrados. Las grandes compañías cuentan con equipos de abogados especializados en su defensa legal, y siempre existe la coartada del error humano. La defensa sostiene que Vasquez no estaba en el momento del accidente viendo un programa concurso, sino chequeando en su teléfono móvil información enviada por la compañía. Además de intentar saber

quién fue culpable, este accidente debería obligar a las compañías implicadas en el desarrollo de los automóviles autónomos a reevaluar sus políticas de seguridad. Sin embargo, parecen más preocupadas por el hecho de que con sucesos con éste la opinión pública ha reducido su fe en la promesa de la tecnología que se presenta con innegables beneficios sociales. Un fiscal de Arizona liberó a la compañía Uber de cualquier responsabilidad criminal, pero muchos observadores consideran que hay indicios de que la compañía tiene alguna responsabilidad, y que Rafaela Vasquez ha sido señalada como chivo expiatorio (Stern, 2021).

Lauren Smiley (2022) ha escrito para la revista Wired un revelador artículo que presenta la primera entrevista concedida por la conductora después del accidente, y también expone una exhaustiva recopilación de datos acerca del accidente que pone en tela de juicio la explicación oficial y nos lleva a concluir que existe un profundo interés por presentar el primer accidente mortal de un vehículo autónomo como un error humano, sacando de los focos cualquier deliberación ética acerca de la planificación de los test en vías públicas y de las garantías de funcionamiento adecuado de estos vehículos de convivencia con ciclistas, peatones y otros conductores.

Cito textualmente un fragmento de la reconstrucción del momento del accidente:

> El sistema de conducción de Uber - que en ese momento llevaba 19 minutos controlando el coche - registró un vehículo que circulaba por delante a 5,6 segundos de distancia, pero no alertó a Vásquez. A continuación, el ordenador anuló su evaluación inicial: no sabía qué objeto era. A continuación, volvió a cambiar la clasificación a un vehículo, y luego osciló entre *vehículo* y *otra cosa*. A 2,6 segundos del objeto, el sistema lo identificó como "bicicleta". A 1,5 segundos, volvió a considerarlo *otra cosa*. Luego volvió a *bicicleta*. El sistema generó un plan para intentar esquivar lo que fuera, pero decidió que no podía. Entonces, a 0,2 segundos del impacto, el coche emitió un sonido para

> alertar a Vásquez de que el vehículo iba a reducir la velocidad. Dos centésimas de segundos antes del impacto, a 39 Mp/h, Vásquez agarró el volante, lo que sacó al coche de la autonomía y lo puso en modo manual. Era demasiado tarde. La bicicleta destrozada dejó una estela de 7 metros en la acera. Una persona yacía gravemente herida en la calzada (Smiley, 2022).

Días después, representantes de la compañía explicaron a la policía que habían anulado el sistema de frenado automático de Volvo para que no interfiriera con los propios sistemas de seguridad instalados por Uber. Una vez que esta compañía retiró sus vehículos autónomos de pruebas, realizó una autoevaluación sobre el accidente y sus prácticas de seguridad. En su análisis de los datos reconoció que su tecnología de visión artificial no había reconocido en ningún momento a la víctima del accidente como una persona:

> Casi cada vez que el sistema cambiaba lo que creía que era Herzberg - un coche, una moto, otra cosa - empezaba de cero para calcular hacia dónde podría dirigirse el objeto, es decir, cruzando la carretera hacia el carril del Volvo. Uber había programado el coche para retrasar la frenada brusca durante un segundo a fin de permitir que el sistema verificara la emergencia - y evitar falsas alarmas- y que el ser humano tomara el control. El sistema sólo frenaría bruscamente si podía evitar por completo el accidente; de lo contrario, reduciría la velocidad gradualmente y avisaría al conductor. En otras palabras, en el momento en que consideró que no podía evitar por completo a Herzberg aquella noche, el coche no pisó a fondo los frenos, lo que podría haber hecho que el impacto fuera menos grave (Smiley, 2022).

Más adelante Volvo realizó pruebas de simulación del accidente con sus propios recursos, y emitió un informe para la *National Transportation Safety Board* en el que demostraba que su sistema patentado de frenada de emergencia, el que Uber anuló para sustituirlo por su propio sistema, habría detenido el vehículo antes de provocar el atropello en 17 de 20 escenarios posibles, y habría

minimizado los daños del atropello al conseguir reducir significativamente la velocidad del vehículo en los otros tres escenarios.

El hecho de que Rafaela Vasquez sea una persona transgénero con antecedentes delictivos ha sido también un factor que ha desviado la atención mediática identificándola como la responsable más probable, la culpable perfecta. Una simple prueba de ello es el título del artículo del *Daily Mail* que está citado en la bibliografía: "Convicted Felon Behind the Wheel of Uber Self-Driving Car Was Streaming the Voice on Her Phone and Laughing Before Crash Which Killed a Pedestrian in Arizona." (Style y Smith, 2018).[6]

Solo en Estados Unidos los accidentes de tráfico se llevan la vida de 38.000 personas al año, y entre el 90 y el 95 % de dichos accidentes son debidos a causas humanas. El fallecimiento de esta ciclista es precisamente el tipo de accidentes que los vehículos autónomos pretenden evitar. Los conductores humanos estamos sometidos al cansancio, al consumo de alcohol o drogas, al despiste o al enfado, condiciones de riesgo que son ajenas a una inteligencia artificial. Si embargo, para llegar a este objetivo hace falta salvar dos obstáculos. El primero de ellos es que las tecnologías que intervienen en la operación de los vehículos autónomos no están certificadas, y por lo tanto se encuentra en una fase equivalente al de un conductor novato que acaba de aprobar el examen de conducir. Ello supone una constante experimentación basada en ensayo y error con vehículos experimentales que están aprendiendo a interactuar con las formas de comportamiento temerarias e improvisadas propias de los seres humanos.

Después de este accidente la compañía Uber dejó de realizar test de vehículos autónomos en las vías públicas, a pesar de que la

[6] La traducción textual del titular del Daily Mail es la siguiente: "Una delincuente condenada al volante de un coche autónomo de Uber veía *La Voz* en su teléfono y se reía antes del accidente en el que murió un peatón en Arizona".

investigación exonera a la compañía de cualquier responsabilidad. Otras compañías como Toyota han decidido detener temporalmente sus test en vías públicas, alegando el efecto emocional que dicho accidente ha tenido en los conductores de pruebas de la compañía. Quizá uno de los problemas éticos está en el hecho de permitir los test de vehículos autónomos en convivencia con peatones, ciclistas y conductores humanos sin las debidas garantías. De cualquier forma, la razón siempre está de lado de quien construye el relato.

Conclusiones provisionales: Reflexiones personales acerca de la sustitución del ser humano por la máquina

Cuando intento desentrañar los dilemas éticos que generará el uso extensivo de vehículos autónomos, siempre me viene a la cabeza la maravillosa película de dibujos animados *Wall-E* de la factoría Pixar. En esta cinta de 2008 se narran las aventuras de un robot diseñado para recoger la basura que cubre todo el planeta después de que la humanidad haya decidido a dejar su hogar hecho un asco y marcharse en un gran crucero espacial lleno de comodidades, destinado a proporcionar a los humanos sobrevivientes una sedentaria vida de placeres vacíos. La humanidad se ha hecho acomodaticia y no tiene que luchar por nada, ya que en el útero materno que es la nave espacial en forma de crucero de vacaciones pueden disfrutar de toda la diversión que les pide el cuerpo. Es una nueva sociedad digital, en tanto que el dedo es el único órgano creador que utilizan. Todo queda a golpe de botón. Para pedir comida, para desplazarse sin levantarse del asiento, o para disfrutar de cualquier tipo de ocio. En esta sátira el más humano de todos es el robot Wall-E, que al final acaba sacando a la humanidad de su deshonrosa ociosidad y recuperando el espíritu de lucha que caracteriza al ser humano. Me pregunto si algo parecido no puede

ocurrir con la implantación de los vehículos autónomos como una *última frontera*, como fue en sus tiempos la *conquista del Oeste*, y más recientemente la conquista del espacio. Parece ser que se abrirán nuevas posibilidades para el ser humano. En lugar de tener que estar atentos al volante durante todo el trayecto, podremos dedicarnos tranquilamente a ver suculentas series de televisión o maravillosos concursos para ver quién cocina el plato más extravagante. Me pregunto si alguien ha realizado algún estudio sobre el ocio dentro de los vehículos autónomos. Quizá los pasajeros hablen más entre ellos ahora que no hace falta estar callados para no desconcentrar del conductor. Quizá las familias se unan más al poder jugar a las cartas dentro del vehículo, o quizá el hábito de lectura de los clásicos vuelve a florecer gracias a la serenidad de espíritu de millones de usuarios que se sentirán liberados con un maravilloso tiempo libre extra con el que no contaban hasta ahora. Sinceramente, lo dudo.

Lo primero que la película Wall-E nos muestra es la pérdida de especialización de los seres humanos cuando automatizamos una tarea que tradicionalmente requería aprendizaje, habilidad, prudencia y destreza. El dueño del vehículo ya no necesita ningún conocimiento técnico para poder utilizarlo. Por supuesto, asumo que los sistemas de aparcamiento automático ya estarán mucho más avanzados, y que habrá otros sistemas automáticos que controlen la presión de los neumáticos y los inflen en caso de necesidad. Como beneficio, tendremos un extra de tiempo libre que nos permitirá una mayor concentración en nuestros propios objetivos, al no tener que preocuparnos por la conducción del vehículo que dejamos en manos más fiables que las nuestras. Pero es evidente que también existe un peligro como contrapartida. Igual que en la computación en nube, donde la carga de procesamiento se desplaza desde el dispositivo personal a los recursos centralizados del sistema con total transparencia para el usuario, también en el caso de los

vehículos autónomos aumenta la dependencia con respecto a la máquina. Existe el riesgo de *shutdown*, es decir, de fallo global del sistema en el momento en que se produzca alguna contingencia en alguno de los múltiples elementos críticos. El ser humano, el que pueda comprar uno de estos vehículos, gana en comodidad, pero pierden *know-how*, esa herramienta de valor añadido que empodera a través de la *ley de rendimientos crecientes de adopción*: cuantos más usuarios adoptan una tecnología y la hacen suya, cuanto más la transforman, más valor añadido hay en la misma. Y además, se crea la capacidad de nuevos empoderamientos sociales aplicando los conocimientos adquiridos a nuevas formas de vida, transformando una tecnología institucional en tecnología social. Claramente se reduce la autonomía técnica del ser humano. Acostumbrados a la transparencia de los procesos de comunicación y de información, el mundo se va convirtiendo en una gran caja negra que decidimos no abrir. No sabemos lo que ocurre dentro de ella, y como un día firmó Arthur C. Clarke, *la técnica más avanzada tiende a confundirse con la magia.* Siempre nos quedará el consuelo de poder promover un plan estatal para la instalación en centros cívicos y asilos de ancianos consolas de videojuegos con programas como *Gran Turismo* o *Forza Horizon*, u otros simuladores de conducción para evitar el deterioro de las capacidades motoras.

La tarea pendiente: la relación entre la ética de los vehículos autónomos y la teoría de juegos

Por otra parte, también existe un fenómeno curioso que estudiaremos en la segunda parte de este análisis, que serás publicada el próximo año, dedicada a la relación entre la toma de decisiones en vehículos autónomos y la teoría de juegos. Tiene que

ver con la aplicación de una ética utilitarista a la formación de los sistemas de toma de decisiones de los vehículos autónomos, de forma que en una situación de peligro escojas siempre el escenario que minimice el daño en términos cuantitativos. Es decir, el menor número posible de heridos o fallecidos. Algunos estudios realizados en 2015 demostraron que la mayoría de las personas aprobarían vehículos autónomos utilitaristas, es decir, que sacrifiquen a sus pasajeros por un bien mayor, y que también querrían que los demás compraran estos vehículos, pero ellos preferirían conducir vehículos que protegían a sus pasajeros por encima de todo. En consecuencia, la adopción de algoritmos utilitarios podría incrementar paradójicamente las muertes en carretera ya que la gente no compraría una tecnología que admiten que es más segura. (Guevara 2021; Bonnefon et al. 2017).

En definitiva, a todos nos gustaría tener un vehículo que protegiera a los ocupantes en primer lugar, pero nos gustaría también que los demás escogieran vehículos que protegieran el mayor bien en caso de accidente. Se parece bastante al argumento de Rousseau acerca de *la voluntad general* frente a *la voluntad de todos*. Podríamos resumir el argumento de base en cuatro puntos:

1. Los usuarios prefieren que los vehículos autónomos empleen algoritmos utilitaristas.
2. Sin embargo, cada uno de ellos no compraría tales vehículos, sino otros que protegieran a sus ocupantes a cualquier precio.
3. Al retrasarse la implantación de vehículos autónomos, el tráfico es más peligroso y mueren muchas más personas.
4. Corolario: la búsqueda de la maximización de la utilidad de las decisiones personales (máximo beneficio), acaba derivando justo en todo lo contrario (*tragedia de los comunes*).

El *dilema del prisionero, el dilema de los padres de* D. Parfit, *el problema del tranvía sin frenos, la tragedia de los comunes* y otros modelos de la teoría de juegos nos permitirán abordar otro conjunto de problemas éticos de carácter fundamentalmente prudencial. Este será el abordaje de un nuevo estudio de los dilemas éticos relacionados con los vehículos autónomos.

En definitiva, se necesita un trabajo conjunto intensivo de industria y academia para responder a las cuestiones éticas que plantean los vehículos autónomos, especialmente en el problema de la certificación de los algoritmos de control que toman decisiones a partir de sensores sometidos a fallos, sabotaje y envejecimiento, y en metodologías de razonamiento que hacen uso intensivo de datos para tomar decisiones por parte de sistemas de seguridad crítica. Para poder enfrentarse a estos problemas será necesario avanzar en el campo emergente de una inteligencia artificial explicable e inteligible. Los vehículos autónomos traen la promesa de una drástica reducción de los accidentes de tráfico y de las muertes provocadas por la carretera, y también permitirán un acceso universal al uso del automóvil por parte de personas portadoras de necesidades especiales. Sin embargo, plantea problemas éticos basados en la ausencia de estándares de seguridad. Necesitaremos protocolos que nos permitan evaluar la adecuación de los sistemas automáticos a principios éticos y de funcionamiento que sean satisfactorios, y para ello podemos aprender lecciones de cómo los seres humanos obtienen sus licencias de habilitación para conducir automóviles o pilotar aeronaves, como argumenta Cummings (2019). Sin duda debemos responder a estos interrogantes sin la presión mediática que la industria impone sobre nosotros, sin identificar necesariamente la automatización con el avance de la humanidad.

Bibliografía

Bansal, S., & Tomlin, C. J. (2019). Control and Safety of Autonomous Vehicles with Learning-Enabled Components. In H. Yu, X. Li, R. Murray, S. Ramesh, & C. J. Tomlin (Eds.), *Safe, Autonomous and Intelligent Vehicles*. Springer Nature. https://doi.org/10.1007/978- 3- 319- 97301- 2.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The Social Dilemma of Autonomous Vehicles. *Science*, *352*(6293): 1573-1576. https://doi.org/10.1126/science.aaf2654

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M. & Canton Ferrer, C. (2020). The deepfake detection challenge (DFDC) dataset. *arXiv preprint* arXiv:2006.07397.

Carlini N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *Security and Privacy* (SP), *2017 IEEE Symposium*, 39–57.

Carlini, N., & Farid, H. (2020). Evading deepfake-image detectors with white- and black-box attacks. *arXiv preprint* arXiv:2004.00622.

Cummings, M.L. (2017). The Brave New World of Driverless Cars: The Need for Interdisciplinary Research and Workforce Development. *TR News*, 34–37.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A, Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). Robust Physical-World Attacks on Deep Learning Visual Classification, *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1625-1634, doi: 10.1109/CVPR.2018.00175.

Gandhi, A., & Jain, S. (2020). Adversarial Perturbations Fool Deepfake Detectors. *arXiv preprint*.
https://doi.org/10.48550/arXiv.2003.10596.

Goodfellow I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint* arXiv:1412.6572.

Guevara, J. F. (2021). Moral machine: descubriendo el desafío ético de las inteligencias artificiales. *Arbor*, *197*(800), a607. https://arbor.revistas.csic.es/index.php/arbor/article/view/2421.

Handi Yu et al. (2019). Efficient Statistical Validation of Autonomous Driving Systems. In H. Yu, X. Li, R. Murray, S. Ramesh, & C. J. Tomlin (Eds.), *Safe, Autonomous and Intelligent Vehicles*. Springer Nature. https://doi.org/10.1007/978- 3- 319- 97301- 2.

Km77.com (2022) Conducción autónoma, niveles y tecnología. https://www.km77.com/reportajes/varios/conduccion-autonoma-niveles.

Kos, J., Fischer, I., & Song, D. Adversarial examples for generative models. *arXiv preprint* arXiv:1702.06832, 2017.

Kwon, C., & Hwang, I. (2019). Cyberattack-Resilient Hybrid Controller Design with Application to UAS. In H. Yu, X. Li, R. Murray, S. Ramesh, & C. J. Tomlin (Eds.), *Safe, Autonomous and Intelligent Vehicles*. Springer Nature. https://doi.org/10.1007/978- 3- 319- 97301- 2.

Li, B., & Vorobeychik, Y. (2014). Feature cross-substitution in adversarial classification. In *ANIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2087–2095.

Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., & Frossard, P. (2016). Universal adversarial perturbations. *CoRR*, abs/1610.08401.

Neekhara, P., Dolhansky, B., Bitton, J., & Ferrer, C. C. (2021). Adversarial Threats to DeepFake Detection: A Practical Perspective. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA. 923-932. https://openaccess.thecvf.com/content/CVPR2021W/WMF/html/N

eekhara_Adversarial_Threats_to_DeepFake_Detection_A_Practical_Perspective_CVPRW_2021_paper.html
Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 427–436.
NTSB (2017). Highway Accident Report: Collision Between a Car Operating with Automated Vehicle Control Systems and a Tractor-Semitrailer Truck near Williston, Florida May 7, 2016. *NTSB/HAR-17/02 PB2017–102600*.
Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, 372–387.
Quach, K. (14 de marzo de 2022). Driver in Uber's self-driving car death goes on trial, says she feels 'betrayed'. *The Register*. https://www.theregister.com/2022/03/14/in_brief_ai/
SAE (2022). Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_202104. *SAE.org*. https://www.sae.org/standards/content/j3016_202104/
Sandel, M. (2011) *Justicia. ¿Hacemos lo que debemos?* Debate.
Shinar, D., & Schieber, F. (1991). Visual requirements for safety and mobility of older drivers. *Hum. Factors*, 33(5), 507–519.
Sibros medium (2022). *The Current State of Automotive Software Related Recalls*. https://sibros.medium.com/the-current-state-of-automotive-software-related-recalls-ef5ca95a88e2
Smiley, L. (2022). 'I'm the Operator': The Aftermath of a Self-Driving Tragedy. *Wired*, 8 de marzo. https://www.wired.com/story/uber-self-driving-car-fatal-crash/
Stern, R. (12 de mayo de 2021). Trial Delayed for Backup Driver in Fatal Crash of Uber Autonomous Vehicle. *Phoenix New Times*.

https://www.phoenixnewtimes.com/news/uber-crash-arizona-vasquez-herzberg-trial-negligent-homicide-charge-11553424.
Styles, R. y Smith, J. (20 de marzo de 2018). Convicted armed robber who was behind the wheel of self-driving Uber when it killed pedestrian as she wheeled her bike across a road. *Daily Mail*. https://www.dailymail.co.uk/news/article-5524031/PICTURED-Felon-wheel-killer-self-driving-Uber-car.html
Yu, H., Li, X., Murray, R., Ramesh, S., & Tomlin, C. J. (Eds.). (2019). *Safe, Autonomous and Intelligent Vehicles*. Springer Nature. https://doi.org/10.1007/978- 3- 319- 97301- 2.