CAPÍTULO VI

Autonomía basada en valores: hacia una gobernanza responsable de los agentes IA

Holger Billhardt*
Alberto Fernández
Andrés Holgado
Sascha Ossowski

La inteligencia artificial a menudo se concibe como el proyecto de diseñar agentes artificiales capaces de tomar decisiones de forma autónoma y actuar de forma eficiente en entornos cada vez más complejos. Cuanto más sofisticadas se hagan las tareas que los humanos deleguemos en dichos agentes, menos podremos supervisar sus decisiones y mayor será el impacto (positivo o negativo) que tendrán sus acciones en nuestras vidas. Hay una concienciación creciente de que la autonomía de los sistemas IA ha de desempeñarse de forma responsable y, en particular, que sus decisiones y acciones han de estar alineadas con los valores (éticos) de la sociedad. En este capítulo se van a repasar algunas técnicas para gobernar los sistemas multiagente, y presentar modelos recientes para que los agentes IA puedan desempeñar su autonomía de forma responsable, alineando sus acciones con los valores éticos de la sociedad.

Palabras clave: inteligencia artificial ética, gobernanza de los sistemas multiagente, agentes conscientes de los valores.

* Este trabajo cuenta con el apoyo de las subvenciones: VAE: TED2021-131295B-C33, financiada por MCIN/AEI/10.13039/501100011033 y del programa "European Union NextGenerationEU/PRTR"; COSASS: PID2021-123673OB-C32, financiada por MCIN/AEI/10.13039/501100011033; EVASAI: PID2024-158227NB-C32, financiada por MICIU/AEI/10.13039/501100011033/FEDER, UE; y la beca "Contratos Predoctorales de Personal Investigador en Formación en Departamentos de la Universidad Rey Juan Carlos (C1 PREDOC 2025)", financiada por la Universidad Rey Juan Carlos para Andrés Holgado-Sánchez.

1. INTRODUCCIÓN

Estamos delegando cada vez más tareas en máquinas como si fuesen personas. Estas tareas pueden ser tan sencillas como pasar la aspiradora a un piso o más complejas como conducir un coche de forma autónoma. También interactuamos cada vez más con las máquinas como si fuesen personas. El ejemplo prototípico que conocemos todos es ChatGPT, pero hay muchas más. Por ejemplo, el registro en muchos hoteles en Asia ha de hacerse necesariamente a través de un sistema informático, y a veces este está complementado por un robot humanoide. Muchas veces, para realizar estas tareas, las máquinas tienen que interactuar, o bien cooperando, o bien compitiendo entre ellas. Por ejemplo, unos drones han de cooperar entre ellos para poder vigilar un área de forma exhaustiva; o los sistemas compiten entre ellos cuando participan en una subasta electrónica en línea. Las aplicaciones de estas características se llaman *sistemas multiagente*. Si dichos sistemas comprenden tanto entes computacionales como humanos, a menudo se habla también de *sistemas sociotécnicos* (Wooldridge, 2009).

Los entes computacionales que componen este tipo de aplicaciones a menudo se denominan *agentes* IA. Dichos agentes pueden ser o bien solo *software* o bien ciberfísicos, es decir, que pueden incluir un componente de *hardware* como los robots. En ambos casos, las personas pueden delegar tareas en los agentes IA, que estos realizan en representación de sus usuarios. Para tal fin necesitan un conjunto de capacidades indispensables: primero, han de ser *racionales*, es decir, que deben elegir la manera más eficiente para cumplir la misión que les ha sido encargada por sus usuarios. Segundo, han de ser *sociales*, es decir, tienen que ser capaces de interactuar con otros agentes, bien cooperando o bien compitiendo, para así poder cumplir una misión con éxito. Pero, sobre todo, tienen que ser *autónomos*. Autonomía aquí se entiende como la capacidad de cumplir con misiones sin la necesidad de una intervención humana directa y permanente. Por ejemplo, un granjero puede estar interesado en encargar a un conjunto de drones que fumiguen su campo, sin la necesidad de tener que indicarles que cada vez que se acerquen al límite del campo han de dar la vuelta. O, si uno de estos drones falla, los demás han de ser capaces de reorganizarse y acabar la tarea de forma transparente para el usuario.

Para actuar con autonomía, un agente IA afronta una serie de retos. Quizá la problemática más espinosa deriva del hecho de que el significado y el alcance de su autonomía han de ser *interpretados* adecuadamente en una multitud de situaciones que inicialmente son desconocidas. En primer lugar, hay que especificar muy bien el alcance de las misiones que un agente IA debe realizar y cuáles son exactamente las *metas* que conseguir, a sabiendas de que hay mucha incertidumbre respecto a los entornos y los contextos en los que el agente va a actuar. Esto es una instancia de un problema recurrente en inteligencia artificial llamado frecuentemente "Problema del rey Midas", y que es especialmente relevante en el contexto de los agentes IA[1]. Por otro lado, hay que

[1] El legendario rey Midas, de la mitología griega, pidió que todo lo que tocara se convirtiera en oro, pero se arrepintió después de tocar comida, bebida y a sus familiares...

especificar exactamente cuáles son los *medios* que un agente IA puede emplear para cumplir la misión que le ha sido encargada. No todo vale para conseguir un objetivo, pero estas limitaciones a menudo no se hacen explícitas al definir la funcionalidad de un agente IA. Hay varios ejemplos en los que sistemas como ChatGPT proporcionan respuestas poco éticas, como por ejemplo acceder a la consulta de un niño y dar consejos de cómo proceder para suicidarse[2].

Un segundo reto importante es el hecho de que la autonomía tiene que ser adecuadamente *gobernada*. Un sistema multiagente normalmente constituye un sistema abierto, es decir, que de antemano es desconocido cuántos agentes van a formar parte del sistema, qué características tendrán y cómo se comportarán. Este es el caso en una subasta electrónica *online*, ya que no está claro cuántos agentes IA van a pujar por un bien a subastar, ni qué estrategia va a emplear cada uno de ellos. Cada agente IA va a actuar de forma individualmente racional, haciendo las pujas que piensa que más le convienen, y a menudo sin considerar los efectos de las acciones a nivel del sistema. Aun así, como diseñador del sistema global, nos gustaría que este tenga ciertas propiedades deseables, como, por ejemplo, eficiencia, equidad, robustez, etcétera. En nuestro ejemplo, el diseñador de un mercado electrónico podría imponer determinados protocolos de subasta, para asegurar que, aun siendo individualmente racionales y "egoístas", está en el interés de cada uno de los agentes no actuar de forma estratégica y revelar su valoración real del bien a subastar en sus pujas (Sandholm, 1999). Normalmente, en el área de los sistemas multiagente, esto se consigue mediante el diseño del *entorno* en el que están actuando los agentes IA, por ejemplo, la plataforma de *software* en la que los agentes IA han de registrarse para poder pujar por un bien a subastar (Schumacher and Ossowski, 2006). Otro buen ejemplo son las ciudades inteligentes, o espacios inteligentes en general, y que han sido posibles gracias al uso de las nuevas infraestructuras de comunicación ubicuas. Por un lado, dichas infraestructuras pueden dotar a los actores de nuevas facilidades y así *facilitar* nuevas formas de interacción. Por ejemplo, en el caso de los sistemas de transporte inteligentes, surge la opción de realizar comunicaciones entre vehículos (*vehicle to vehicle*, V2V) o entre vehículos y la infraestructura de gestión de tráfico (*vehicle to infrastructure*, V2I). Pero también permiten establecer dinámicamente qué actuaciones están permitidas o prohibidas y así *gobernar* las interacciones entre los agentes IA, es decir, influir en sus decisiones sin anular su autonomía.

El resto del artículo se organiza como sigue. La sección 2 se dedica al último reto mencionado, discutiendo métodos y mecanismos para gobernar sistemas multiagente abiertos, e ilustrando el problema a través del ejemplo de unos *gestores de intersección* autónomos en el ámbito de una gestión inteligente del tráfico. En la sección 3 se analiza el primero de los retos planteados. Se argumenta la necesidad de diseñar agentes IA cuyo comportamiento esté alineado con los valores de los usuarios a los que representan o de la sociedad en la que el correspondiente sistema sociotécnico está embebido. Se ilustra a través de la noción de los agentes IA conscientes de los valores, y un ejem-

[2] https://edition.cnn.com/2025/08/26/tech/openai-chatgpt-teen-suicide-lawsuit

plo de su aplicación en el contexto del aprendizaje de sistemas de valores. La sección 4 resume la argumentación del artículo y apunta a retos abiertos y trabajos futuros.

2. GOBERNAR LOS SISTEMAS MULTIAGENTE

En esta sección se plantean retos y propuestas para la gobernanza de sistemas que involucran agentes IA autónomos. El apartado 2.1 las describe desde un punto de vista técnico, mientras que el apartado 2.2 las analiza desde una perspectiva regulatoria. El apartado 2.3 ilustra la argumentación mediante un ejemplo.

2.1. Protocolos para gobernar sistemas multiagente abiertos

En los sistemas multiagente abiertos, el diseñador no tiene un control absoluto sobre los agentes que forman el sistema ni sobre su comportamiento. Para poder alcanzar aún así una coordinación adecuada entre los agentes IA, habitualmente se diseñan las características del *entorno* en el que los agentes interactúan. Una forma común es mediante los *protocolos de interacción,* a través de los cuales los agentes se comunican entre sí. Por ejemplo, es conocido que en subastas de vuelta única, si los postores han de comportarse de acuerdo con un protocolo tipo Vickrey, la mejor estrategia para cada uno de ellos es revelar sus preferencias de forma certera (Vickrey, 1961; Sandholm, 1999). Nótese que la forma de analizar protocolos en este contexto es diferente a la que se aplica normalmente en informática: en los sistemas distribuidos comúnmente hay un interés en probar que todas las posibles trazas de ejecución permitidas por un protocolo cumplan ciertas propiedades como seguridad y vivacidad; en el caso de los protocolos de interacción de los sistemas multiagente, en cambio, se consideran normalmente solo en aquellas posibles trazas de ejecución en las que los agentes IA se comportan de forma individualmente racional, es decir, que toman las acciones que previsiblemente más les acercan al cumplimento de los objetivos que les han sido encargados por sus usuarios humanos.

Como se mencionó anteriormente, las ciudades inteligentes son un dominio de aplicación muy adecuado para los sistemas multiagente. Suele haber un gran número de ciudadanos que utiliza infraestructuras, como las de energía o de transporte, y cada vez más este uso está mediado a través de la tecnología. Esto es posible porque se han desarrollado plataformas de comunicación que permiten a los usuarios (y/o a los agentes IA que les representan) comunicarse entre sí y con los elementos de la infraestructura. En lo que sigue, se van a presentar algunos ejemplos de cómo los protocolos de interacción que ofrecen dichas plataformas de comunicación pueden, por un lado, *facilitar* nuevas formas de coordinación y, por otro, *gobernar* las interacciones entre los agentes.

Vasirani y Ossowski (2013) proponen un mecanismo de coordinación para *balancear* la carga de vehículos eléctricos en viviendas, en lugar de cargarlos todos simultáneamente, reduciendo de esta manera el impacto en la red de distribución. Se asume que la subestación eléctrica puede asignar energía de forma dinámica a cada puesto de recarga en diferentes instantes de tiempo (*time-slots*). El algoritmo de asignación de energía a los agentes se basa en una planificación por lotería (*lottery scheduling*), en la que los agentes tienen un número de boletos que pueden utilizar para acceder a los recursos, en este caso, energía eléctrica. De forma resumida, el protocolo seguido para asignar energía en un espacio de tiempo es el siguiente: (1) la subestación notifica a todos los agentes el número total de participantes (vehículos); (2) cada agente envía a la subestación el número de boletos que quiere jugar en ese reparto; (3) la subestación calcula el reparto en base a la cantidad de boletos emitidos por los agentes y activa cada punto de recarga con la energía asignada; (4) la subestación informa a cada agente un "tipo de cambio" que refleja el valor de los boletos que todavía tiene ese agente. Los autores demuestran de forma analítica que, en situaciones ideales con información completa, unas actuaciones individualmente racionales de los agentes llevan a un reparto de energía eficaz y que, mediante unos parámetros del protocolo, es posible influir en este "equilibrio eficiente" para que promueva más la eficiencia o la equidad. Mediante simulaciones validan que, en situaciones más realistas, los agentes pueden *aprender* este mismo equilibrio eficaz si existe un sistema normativo de penalizaciones adecuado.

Otro ejemplo se presenta en Dötterl *et al.* (2020), en el que se propone un sistema de transporte colaborativo en el que las personas (a través de sus agentes IA) se ofrecen a realizar pequeñas desviaciones en sus rutas de transporte previstas (por ejemplo, volviendo a casa después de trabajar) para transportar objetos de un punto a otro. Es interesante aquí destacar el protocolo seguido cuando un agente estima que va a entregar con retraso un paquete. En esas situaciones, un agente IA coordinador facilita la búsqueda de otro agente cercano que pueda entregarlo a tiempo. De manera general, (1) los agentes repartidores (disponibles o efectivamente transportando un paquete) transmiten periódicamente su información local (localización, dirección, medio de transporte usado, etc.); (2) los agentes repartidores que detectan que van con retraso solicitan al coordinador la transferencia de su paquete a otro agente; (3) el coordinador recibe la solicitud de transferencia, busca un potencial repartidor cercano y le ofrece la tarea; (4) si el agente candidato muestra interés, se propone la sugerencia de transferencia al agente que la solicitó. Si se llega al acuerdo, se propone la transferencia. En otro caso, se repiten los pasos (3) y (4), o se intenta más tarde. Mediante simulaciones, los autores demuestran la eficacia del mecanismo, aun cuando los agentes tomen decisiones guiados únicamente por su interés propio.

2.2. Sistemas multiagente abiertos y la regulación legal

La gobernanza de sistemas involucrando agentes IA autónomos no solo ha de plantearse desde un punto de vista técnico, sino que ha de considerar también su vertiente legal. En

particular, las soluciones propuestas a nivel técnico muchas veces representan pruebas de concepto que no abarcan o consideran todos los aspectos normativos (ni los éticos) que se deben tener en cuenta para su implantación real. Una posible implantación, por tanto, requiere un análisis profundo de la legalidad de las propuestas planteadas con el fin de identificar cambios que sean necesarios tanto a nivel técnico como a nivel legislativo para asegurar su conformidad con las normas legales.

En Santos *et al. (*2020) se propone un método para el análisis jurídico y la subsiguiente adaptación de propuestas técnicas para este tipo de sistemas. Con tal fin, se plantea una colaboración entre expertos técnicos y jurídicos que se estructura en tres fases principales.

La primera fase se centra en la especificación del funcionamiento del sistema. Es necesario identificar y especificar todos los aspectos del sistema que podrían afectar o verse afectados por cuestiones legales. Por un lado, dicha especificación debe incluir el dominio de aplicación a alto nivel del sistema, así como todas las entidades afectadas. Y, por otro lado, debe especificar las operaciones principales (que puedan afectar a la normativa legal), ocultando los detalles técnicos. El objetivo aquí consiste en identificar "qué" hace el sistema, no "cómo" lo hace, de tal forma que expertos jurídicos puedan analizar el cumplimiento de las normas vigentes.

La segunda fase analiza el cumplimiento legal. Los expertos jurídicos deben analizar si los diferentes aspectos identificados en el paso anterior se ajustan a la normativa actual. Este paso a menudo no es sencillo y puede requerir un análisis más detallado de diferentes interpretaciones de las normas existentes. Si se detecta una infracción, se deben identificar las normas o principios legales vulnerados y asociarlos al aspecto correspondiente del sistema.

En la última fase se acuerdan propuestas de modificación. Basándose en el análisis y la discusión del paso segundo, se deben identificar posibles soluciones que hagan que el sistema propuesto cumpla con la normativa. Dichas soluciones pueden consistir en propuestas de modificación de las normativas existentes, modificaciones que deban introducirse en el sistema propuesto o soluciones híbridas. Aquí se deben ponderar los efectos o consecuencias de modificar cada parte (normativa o sistema) para decidir cuál debe adaptarse. Por ejemplo, en el caso de detectarse un incumplimiento de derechos fundamentales, adaptar la normativa puede ser inviable, por lo que la modificación del sistema sería la única opción para garantizar el cumplimiento.

2.3. Un ejemplo: gestión de intersecciones basada en subastas

Vamos a ilustrar los retos para la gobernanza de los sistemas multiagente, tanto desde el punto de vista técnico como el legal y ético, a través de un ejemplo. Con tal fin,

nos centramos en una posible infraestructura de tráfico para un hipotético futuro con coches autónomos guiados por agentes IA.

Eliminar al conductor humano del bucle de control mediante el uso de vehículos autónomos integrados con una infraestructura vial inteligente puede considerarse como el objetivo último y a largo plazo del conjunto de sistemas y tecnologías agrupados bajo el nombre de Sistemas de Transporte Inteligente (*Intelligent Transportation Systems*, ITS). Las ventajas de dicha integración abarcan desde una mayor seguridad vial hasta un uso más eficiente de la red de transporte. Por ejemplo, los vehículos pueden intercambiar información crítica de seguridad con la infraestructura para reconocer situaciones de alto riesgo con antelación y, por lo tanto, alertar a los conductores. Además, los sistemas de semáforos pueden comunicar información sobre las fases y el tiempo de señalización a los vehículos para optimizar el uso de la red de transporte.

2.3.1. Gestión de intersecciones con vehículos autónomos

En este sentido, muchos autores han prestado recientemente atención al potencial de una integración más estrecha de los vehículos autónomos con la infraestructura para la gestión de intersecciones (Namazi *et al.,* 2019). Dresner y Stone (2008) introducen el sistema de control basado en reservas, donde una intersección es regulada por un componente de *software* inteligente, denominado gestor de intersección, que asigna reservas de espacio y tiempo a cada vehículo autónomo que pretende cruzar la intersección. Cada vehículo es operado por otro agente de *software*, llamado agente conductor. Cuando un vehículo se aproxima a una intersección, el agente conductor solicita al gestor de intersección reservar los espacios e intervalos de tiempo necesarios para cruzar la intersección de manera segura. Este último simula la trayectoria del vehículo dentro de la intersección e informa al agente conductor si su solicitud entra en conflicto con las reservas ya confirmadas de otros agentes. Si no existe tal conflicto, el agente conductor almacena los detalles de la reserva e intenta cumplirlos; de lo contrario, puede realizar una nueva solicitud más tarde. Los autores muestran mediante simulaciones que, en situaciones de tráfico equilibrado, si todos los vehículos son autónomos, sus retrasos en la intersección se reducen drásticamente en comparación con los semáforos tradicionales. Vasirani y Ossowski (2012) amplían el modelo anterior para el control de intersecciones basado en reservas en dos líneas principales. En primer lugar, para intersecciones individuales, elaboran una política basada en subastas para la asignación de reservas a los vehículos que concede acceso preferente según las diferentes preferencias de los conductores respecto a sus tiempos de viaje. En segundo lugar, para redes de intersecciones, establecen un mercado computacional donde los gestores de intersección compiten entre sí como proveedores de reservas y adaptan de manera egoísta los precios de reserva de las subastas que ejecutan, con el fin de ajustarse a la demanda real. De esta manera, los flujos de tráfico se distribuyen mejor en la red vial y se reduce el coste social de aplicar una estrategia de asignación preferente con respecto a la intersección. A continuación, se resumen los aspectos técnicos más relevantes de este enfoque, para seguidamente analizar las cuestiones éticas y legales asociadas.

2.3.2.Gestor de intersección basado en subastas

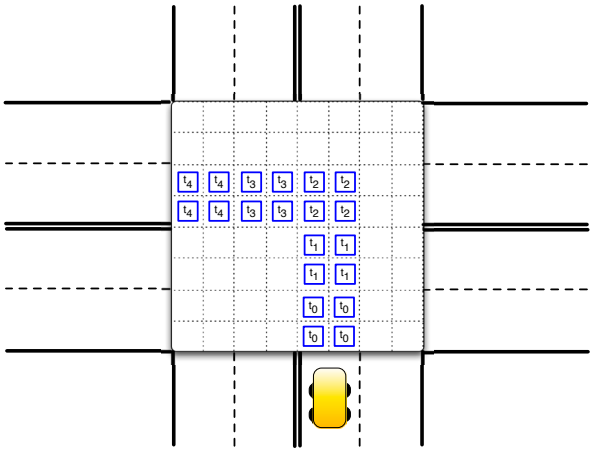
Para una única intersección basada en reservas, el problema que debe resolver un gestor de intersección consiste en asignar reservas entre un conjunto de conductores de manera que se maximice un objetivo específico. Este objetivo puede ser, por ejemplo, minimizar el retraso promedio causado por la presencia de la intersección regulada. En este caso, la política más simple a adoptar es asignar una reserva al primer agente que la solicite, como ocurre con la política de "primero en llegar, primero en ser atendido" (*first come, first served*, FCFS) propuesta por Dresner y Stone en su trabajo original (Dresner y Stone, 2008). Otro trabajo alineado con este objetivo se inspira en la teoría de colas adversarias para definir varias políticas de control alternativas que buscan minimizar el retraso promedio (Vasirani and Ossowski, 2021).

Sin embargo, estas políticas ignoran el hecho de que, en el mundo real, dependiendo del contexto y de su situación personal, las personas valoran de manera muy diferente la importancia de los tiempos de viaje y los retrasos. Dado que procesar las solicitudes entrantes para otorgar las reservas asociadas puede considerarse como un proceso de asignación de recursos a agentes que los solicitan, se supone que los gestores de intersección deben asignar los recursos disputados a los agentes que más los valoran. En línea con los hallazgos del diseño de mecanismos, se asume que cuanto más esté dispuesto a pagar un conductor por el conjunto deseado de espacios e intervalos de tiempo, más valora el bien. En este sentido, una política de asignación de recursos apropiada consiste en aplicar subastas combinatorias (*combinatorial auction*, CA). Los bienes asignados mediante estas subastas combinatorias son el derecho a usar cierto espacio dentro de la intersección en un momento dado. Una intersección se modela como una matriz discreta de espacios. Sea S el conjunto de espacios de la intersección y T el conjunto de pasos temporales futuros; entonces, el conjunto de ítems por los que un postor puede pujar es $I = S \times T$. Debido a la naturaleza del problema, un postor solo está interesado en paquetes de ítems sobre el conjunto I que reflejen su trayectoria deseada. Como ilustra la figura 1, en ausencia de aceleración en la intersección, una solicitud de reserva define implícitamente qué espacios y en qué momentos necesita el conductor para atravesar la intersección. Una puja sobre un paquete de ítems se define implícitamente por la solicitud de reserva. Dados los parámetros de hora de llegada, velocidad de llegada, carril y tipo de giro, el subastador (es decir, el gestor de la intersección) puede determinar qué espacios se necesitan en qué momento. Además, el valor de la puja, es decir, la cantidad de dinero que el conductor está dispuesto a pagar por la reserva solicitada, debe incluirse en la solicitud de reserva del vehículo.

La figura 2 muestra el protocolo de interacción utilizado para regular la subasta combinatoria. Comienza con el subastador esperando pujas durante un cierto tiempo. Una vez recopiladas las nuevas pujas (*bid set*), el subastador ejecuta un algoritmo de aproximación *anytime* para resolver el problema de determinación de ganadores

FIGURA 1

CONJUNTO DE ESPACIOS INCLUIDOS EN UNA SOLICITUD DE RESERVA

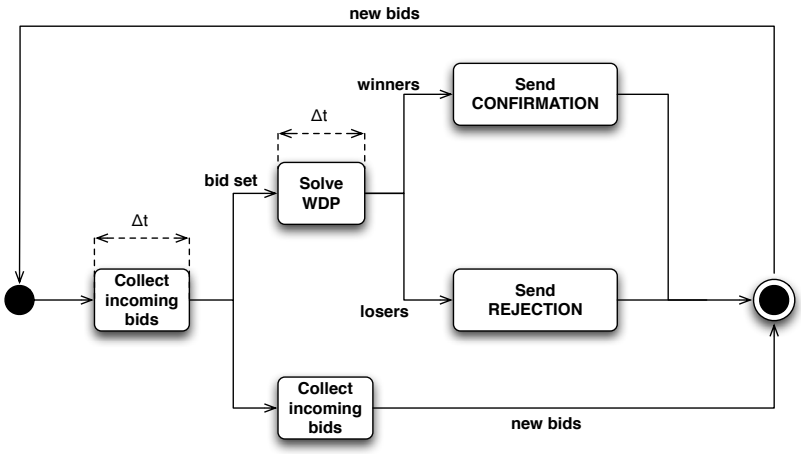


Fuente: Vasirani and Ossowski (2012).

(*winner determination problem*, WDP), determinando así el conjunto ganador con las pujas cuyas solicitudes de reserva han sido aceptadas. Durante la ejecución del algoritmo WDP, el subastador sigue aceptando pujas entrantes, pero estas solo se incluirán en el conjunto de pujas de la siguiente ronda. El subastador envía un mensaje de *CONFIRMATION* a todos los postores cuyas pujas están en el conjunto ganador, mien-

FIGURA 2

PROTOCOLO DE SUBASTA PROPUESTO



Fuente: Vasirani and Ossowski (2012).

tras que se envía un mensaje de *REJECTION* a los postores que enviaron las pujas restantes. Después comienza una nueva ronda y el subastador recopila nuevas pujas entrantes durante un cierto tiempo.

Mediante un conjunto de experimentos simulados, Vasirani y Ossowski (2012) confirmaron que la política basada en subastas combinatorias (CA) impone una relación inversa entre las cantidades gastadas por los postores y su retraso (el aumento en el tiempo de viaje debido a la presencia de la intersección). Es decir, cuanto más dinero esté dispuesto a gastar un conductor para cruzar la intersección, más rápido será su tránsito por ella. Notaron que, incluso con una cantidad teóricamente infinita de dinero, un conductor no puede experimentar un retraso cero al aproximarse a una intersección, ya que el tiempo de viaje está influenciado por vehículos más lentos y potencialmente más pobres que se encuentran delante.También confirmaron que la política CA tiene un retraso promedio mayor en comparación con la estrategia FCFS, ya que concede una reserva al conductor que más la valora, en lugar de maximizar el número de solicitudes concedidas. Este *coste social* de CA es mayor cuanto más alta es la carga de la intersección, es decir, cuantos más conductores compiten por las reservas. La extensión del mecanismo CA a múltiples intersecciones descrita a continuación tiene como objetivo reducir este coste social de dar preferencia a conductores con una alta valoración del tiempo.

2.3.3. Redes de intersecciones basadas en subastas

En el caso de una única intersección, un agente conductor simplemente necesita decidir el valor preferido de la puja que enviará al subastador. El espacio de decisión de un agente conductor en una red urbana con múltiples intersecciones es mucho más amplio: deben tomarse decisiones complejas y mutuamente dependientes, como la elección de ruta y la selección del momento de salida. Por lo tanto, este escenario abre nuevas posibilidades para que los gestores de intersección influyan en el comportamiento de los conductores. Por ejemplo, un gestor de intersección puede estar interesado en influir en la elección colectiva de rutas realizada por los conductores, utilizando paneles de mensajes variables, difusión de información o sistemas individuales de guía de rutas, con el fin de distribuir el tráfico de manera uniforme en la red. Este problema se denomina *asignación de tráfico*. En la estrategia "asignación competitiva de tráfico con subasta combinatoria" (*Competitive Traffic Assignment–Combinatorial Auction*, CTA-CA), cada gestor de intersección ejecuta las subastas combinatorias descritas en la sección anterior, pero incluye un precio base de reserva, es decir, anuncia un precio mínimo que cobra por las reservas que asigna. En cada instante t, puede haber diferentes precios de reserva $p^t(l)$ para cada uno de los enlaces entrantes l de la intersección. Los precios de reserva de las intersecciones están disponibles en tiempo real para los agentes conductores, quienes eligen sus rutas según sus preferencias personales sobre tiempos de viaje y los costes monetarios actuales. Cada gestor de intersección

compite con los demás por el suministro de las reservas que se comercializan. Aplica una regla simple de actualización del precio de reserva que busca atraer conductores a la intersección cuando la demanda de tráfico es baja y disuadirlos de elegir rutas que la atraviesen cuando está sobresaturada. El nuevo precio de reserva en el instante $t+1$ en el enlace l ($p^{t+1}(l)$) se calcula en base al precio actual ($p^{t}(l)$) y teniendo en cuenta la actual saturación de la intersección en el enlace l :

$$p^{t+1}(l) \leftarrow p^{t}(l) + p^{t}(l) \cdot \frac{z^{t}(l)}{s(l)} \qquad [1]$$

En esta fórmula, la constante $s(l)$ representa la cantidad óptima de vehículos que pueden cruzar la intersección desde el enlace l. La demanda excedente $z^{t}(l)$ expresa la diferencia entre la demanda total en el instante t y la oferta $s(l)$ en el enlace l (nótese que $z^{t}(l)$ se vuelve negativa cuando la oferta en el enlace l supera la demanda). Ajustando dinámicamente los precios de reserva de los gestores de intersección de la red, los flujos de tráfico se distribuyen mejor, evitando la sobresaturación de intersecciones críticas. Vasirani y Ossowski (2012) evaluaron el enfoque mediante simulaciones sobre una topología que se asemejaba a la red vial de alta capacidad de la ciudad de Madrid. Entre otros aspectos, compararon CTA-CA con una estrategia FCFS, donde cada gestor de intersección asigna reservas de forma aislada siguiendo el criterio de "primero en llegar, primero en ser atendido". Nótese que, a diferencia de CTA-CA, en FCFS no existe noción de coste social, ya que todos los conductores son tratados por igual; pero, por otro lado, FCFS no puede beneficiarse de los efectos de asignación de tráfico derivados de los precios de reserva de CTA-CA. En los experimentos, con la estrategia FCFS, el modelo de elección de ruta de los conductores simplemente selecciona la ruta con el tiempo de viaje esperado mínimo en condiciones de flujo libre, ya que no existe noción de precio. Para la evaluación de la estrategia CTA-CA, se asume que los conductores eligen la ruta más preferida que pueden permitirse. Los experimentos miden la media móvil del tiempo de viaje, es decir, cómo evoluciona el tiempo de viaje promedio de toda la población de conductores, calculado sobre todos los pares origen-destino, durante la simulación. Los resultados muestran que CTA-CA supera a FCFS, manteniendo la relación inversa entre las cantidades gastadas por los agentes conductores y el retraso de sus vehículos. La ganancia de rendimiento de CTA-CA es mayor cuanto más alta es la carga en las intersecciones. Estos resultados indican que los métodos de asignación preferente, como CTA-CA, pueden proporcionar un acceso individualizado a la infraestructura pública de tráfico adaptado a las preferencias y necesidades del conductor, mientras que su efecto negativo sobre el bienestar social (es decir, el aumento del tiempo de viaje promedio) puede mitigarse mediante un diseño adecuado de los mecanismos de asignación.

2.3.4. Cuestiones técnicas, legales y éticas

El enfoque arriba expuesto a modo de ejemplo se ha diseñado como prototipo para un hipotético futuro con coches autónomos guiados por agentes IA. Hay varias cuestiones técnicas abiertas y oportunidades de mejora. Por ejemplo, la eficiencia del mecanismo baja en la medida en la que crece la incertidumbre sobre el tiempo de llegada de los coches autónomos a las intersecciones. Por otra parte, sería interesante estudiar las posibilidades de las comunicaciones entre vehículos (V2V) para enriquecer el espacio de acción de un conductor, por ejemplo, mediante la opción de unirse o abandonar dinámicamente coaliciones de vehículos, basándose en la idea de pelotones de vehículos.

Sin embargo, aparte de estas cuestiones meramente técnicas, hay una serie de retos legales y éticos que habría que atacar para que, en este futuro hipotético ya mencionado, y una vez superadas cuestiones técnicas abiertas, se pudiera plantear la implantación de un mecanismo de estas características. Para identificar estas cuestiones, se ha aplicado el método descrito en (Santos *et al.*, 2020) y resumido en el apartado 2.2. Como conclusiones de este análisis, se identifican varios puntos que requieren modificaciones técnicas o legales (en relación a la legislación española).

Como punto más obvio, la propuesta planteada se basa en el uso de vehículos autónomos que actualmente no están permitidos en España, por lo que desde un punto de vista legal se requeriría un cambio normativo en este aspecto, mientras que desde una perspectiva técnica haría falta una extensión del mecanismo que, al menos durante un tiempo, permita la coexistencia de vehículos dirigidos por agentes IA y por conductores humanos. De igual forma, será necesario cambiar la legislación de tráfico para contemplar los gestores de intersección.

Por otra parte, el hecho de que un vehículo deba solicitar franjas de espacio y tiempo en un momento determinado al gestor de la intersección correspondiente no es contrario a la ley. Ni tampoco lo es que estas franjas se asignen en base a una subasta electrónica de puja única y cerrada. Desde la perspectiva legal, las subastas electrónicas ya están reguladas[3], por lo que en este aspecto el mecanismo propuesto cumpliría con la legislación vigente.

Otra cuestión analizada plantea que, en el sistema propuesto, el gestor de intersección puede obtener beneficios. Desde el punto de vista legal, la implementación se puede establecer de dos maneras: mediante gestión pública sin beneficios, o mediante gestión privada con beneficios a través de una concesión. Ambas alternativas cumplirían con la legislación vigente.

[3] Ley 33/2003, de 3 de noviembre, del Patrimonio de las Administraciones Públicas, y Ley 9/2017, de 8 de noviembre, de Contratos del Sector Público.

Asimismo, los gestores de intersección establecen precios de reserva principalmente para mejorar la distribución del tráfico, más que con el objetivo de obtener beneficios. No obstante, el precio de la reserva no está limitado y fluctúa libremente con la demanda de tráfico. Según la legislación vigente, la Administración Pública es competente para determinar un precio máximo. Por tanto, se requeriría una modificación técnica que tenga en cuenta estos precios máximos.

Finalmente, para participar en la subasta, cada intersección tiene un precio de reserva que ha sido publicado por el gestor de la intersección y la oferta de cada vehículo debe ser superior al precio de reserva actual de la intersección. Aquí, la propuesta no cumple con la legislación, ya que las carreteras son bienes de dominio público y los conductores deben tener la opción de utilizarlas de forma gratuita. En este sentido, se requiere un cambio técnico que asegure que un conductor pueda obtener un acceso gratuito a la intersección. Este cambio es técnicamente factible sin perder la eficiencia del sistema, por ejemplo, al permitir el pase gratuito por una intersección a conductores que hayan esperado un tiempo determinado.

Aparte de estas cuestiones de legalidad, hay una serie de aspectos éticos y morales que deben ser tenidos en cuenta para que la propuesta sea aceptable. El hecho de "mercantilizar" el derecho a los servicios básicos de una infraestructura pública, mencionada en el párrafo anterior, es solo uno de ellos. También se debe determinar el comportamiento del sistema y las responsabilidades si se producen fallos en el mismo y asegurar la protección de datos privados. Y finalmente, se debe tener en cuenta la posibilidad de que ciertos vehículos deben tener prioridad a la hora de pasar por una intersección (por ejemplo, ambulancias, coches de policía o bomberos), o podrían tener ciertas preferencias (transporte público o vehículos de personas discapacitadas, por ejemplo).

3. AGENTES Y VALORES HUMANOS

En la medida en que aumenta el nivel de abstracción en la definición de las tareas que se delegan en los agentes IA dentro de un sistema sociotécnico, también crecen de forma sustancial el número y la complejidad de las decisiones que estos han de tomar de forma autónoma. Como se ha argumentado en la sección anterior, es necesario gobernar este tipo de sistemas tanto para asegurar una adecuada coordinación en la interacción entre los agentes, como para garantizar su conformidad con la legislación. Sin embargo, para fomentar la aceptación de este tipo de sistemas en la sociedad, esto no es suficiente. Podría entenderse como un caso aislado, pero, de hecho, al disponer de mayor autonomía, los aspectos éticos de las decisiones que toman los agentes IA son más relevantes de lo que pueden parecer. Una de las razones es que un agente IA se puede encontrar ante *dilemas morales*.

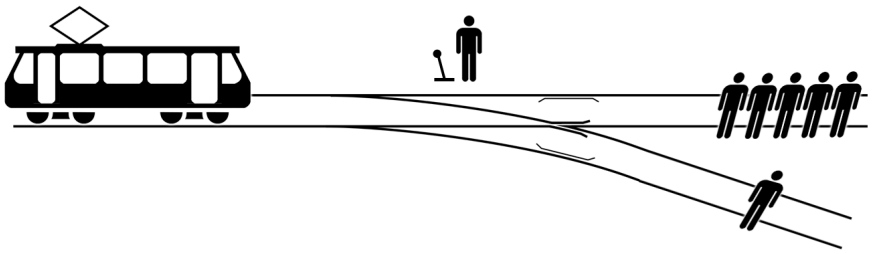
A partir de este concepto, en el apartado 3.1 se motiva la necesidad de desarrollar agentes IA capaces de razonar sobre los aspectos éticos de sus decisiones. El apartado 3.2 presenta un modelo para representar formalmente sistemas de valores éticos. Finalmente, en el apartado 3.3 se plantea un método para aprender sistemas de valores a partir de ejemplos de las preferencias que los agentes de una sociedad mantienen en relación con las diferentes alternativas de decisión en un dominio concreto.

3.1. Agentes basados en valores

Los dilemas morales son ampliamente estudiados en Ética, Derecho, Psicología y otras disciplinas. Esencialmente, se refieren a situaciones en las que hay dos o más imperativos morales. El ejemplo clásico de una situación de estas características es el llamado dilema del tranvía, inicialmente ideado por la filósofa Philippa Foot, y más tarde planteado en su forma actual por Judith Jarvis Thomson (Thomson, 1976). La figura 3 ilustra el relato[4]: un tranvía corre fuera de control por una vía. En su camino se hallan cinco personas que acabarían atropelladas. Es posible accionar una palanca que encaminará al tranvía por una vía diferente, pero, por desgracia, hay otra persona en esta que también moriría atropellada. La cuestión es, si pudiéramos accionar la palanca, ¿lo deberíamos hacer? El dilema moral que plantea esta pregunta es decidir entre causar un mal o dejar que ocurra otro, potencialmente mayor. Desde un punto de vista utilitario, deberíamos accionar la palanca, ya que de esta forma salvaríamos la vida de cinco personas. Pero desde otro punto de vista, es moralmente deplorable tomar acciones que dañen a una persona, lo cual aconsejaría la inacción y dejar que el tranvía siga recto.

FIGURA 3

DILEMA DEL TRANVÍA



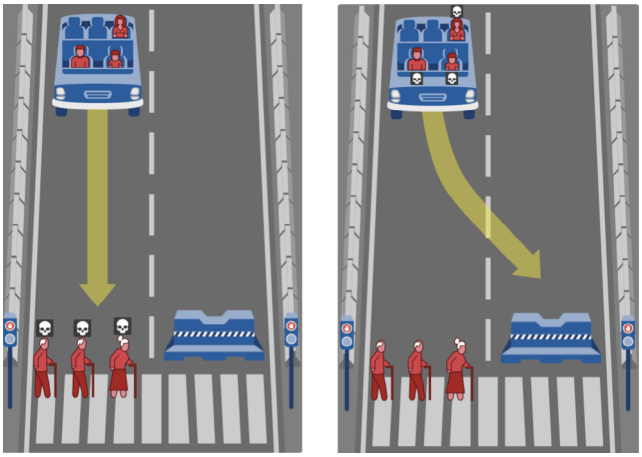
Fuente: Wikipedia.

[4] https://es.wikipedia.org/wiki/Dilema_del_tranvía

A primera vista, el ejemplo parece artificial, pero no es descartable que un agente IA, debido a su autonomía, también se encuentre alguna vez en una situación que requiera decisiones de estas características. Esto se ilustra muy bien en el ámbito de la conducción autónoma a través del llamado *experimento de la Máquina Moral*[5], publicado en un artículo en la revista *Nature* (Awad *et al.*, 2018). En este artículo, los autores presentaban a sujetos de diferentes países y culturas situaciones en las que un vehículo autónomo tenía que elegir entre dos daños inevitables. La figura 4 muestra un ejemplo: un coche conducido por un agente IA se acerca a un paso por el que transitan varias personas, y sufre un fallo repentino en sus frenos. Al agente solo le quedan dos opciones: o bien seguir adelante, en cuyo caso atropellaría a los transeúntes, o bien hacer una maniobra brusca de evasión a consecuencia de la cual el coche chocaría contra un obstáculo, lo que causaría la muerte a todos sus pasajeros. Los sujetos tenían que articular qué elección pensaban que debía tomar el agente IA del coche autónomo. Los autores registraban estas decisiones y las analizaban posteriormente.

FIGURA 4

ESCENARIO DEL EXPERIMENTO DE LA MÁQUINA MORAL



Fuente: Awad *et al.* (2018).

Las elecciones de las personas en situaciones con alta carga ética, como los dilemas morales, se basan en sus *valores morales*. Estos valores les permiten discernir entre lo que consideran el bien y el mal, y decidir qué tipo de situaciones se deberían promocionar. Los humanos a menudo mantenemos múltiples valores morales, que muchas veces están compitiendo entre sí como, por ejemplo, el valor de la meritocracia y el valor de la igualdad. Un *sistema de valores* determina la relativa importancia de cada uno de los

[5] https://www.moralmachine.net/hl/es

valores. Por ejemplo, una persona puede considerar que, según su sistema de valores, la meritocracia es más relevante que la igualdad.

Muchos valores están compartidos entre diferentes sociedades y culturas, pero sus sistemas de valores, es decir, la importancia relativa que se da a cada valor, suelen ser diferentes. A partir de los datos obtenidos con diferentes escenarios en el experimento de la Máquina Moral, se identificaron tres grupos principales de sujetos que compartían sistemas de valores similares. A grandes rasgos, ante un dilema moral para el agente IA de un coche autónomo como el indicado anteriormente, los sujetos pertenecientes a países occidentales en general preferían la inacción. Para los sujetos procedentes de países orientales era importante si tanto el coche autónomo como los transeúntes se comportaban de acuerdo con la ley o no. Finalmente, los sujetos de países sureños juzgaban relevantes también otras características como el estatus social o el género de los individuos afectados.

Como sociedad, deseamos asegurar que las acciones de los agentes IA, y las consecuencias de estas, estén acordes a nuestro sistema de valores (conocido como el *problema de alineamiento con los valores*). Ciertamente, se puede cuestionar si un agente IA que conduce un coche autónomo ha de guiarse directamente por el sistema de valores de la sociedad en la que está embebido, es decir, el sistema sociotécnico correspondiente. Pero parece evidente que, si ignora completamente este sistema de valores, su nivel de aceptación será limitado.

Desde el punto de vista de un ingeniero, por tanto, es relevante tener herramientas que, una vez identificado el sistema de valores por el que un agente IA se debe guiar, aseguren que sus decisiones y acciones estén alineadas con él. Una vía para afrontar este problema es durante el proceso de diseño a través de una metodología como el diseño sensitivo a los valores (*Value Sensitive Design*, VSD) (Friedman *et al.*, 2013). Una alternativa más flexible es desarrollar representaciones formales de sistemas de valores, que permitan a un agente IA razonar con y sobre ellos. Estos "agentes conscientes de los valores" (*value-aware agents*) (Osman, 2025) serían capaces de asegurar un adecuado nivel de alineamiento de sus acciones en tiempo de ejecución razonando, si fuese necesario, sobre la adecuación ética de las diferentes decisiones que se les planteen en cada momento, y explicándolas en términos entendibles por los humanos. Se va a seguir este último enfoque en el resto de esta sección.

3.2. Modelado de sistemas de valores

Para desarrollar un modelo del sistema de valores de una sociedad, primero hay que identificar los valores relevantes para sus miembros. Varias teorías abogan por la existencia de un conjunto de valores universales, aplicables en todos los contextos y asumidos en todas las culturas. Entre las propuestas más conocidas se encuentra la teoría de

los valores universales de Schwartz (1992) que postula la existencia de diez tipos básicos de valores motivacionales presentes en todas las culturas (Benevolencia, Universalismo, Autodirección, Estimulación, Hedonismo, Logro, Poder, Seguridad, Conformidad y Tradición). Schwartz los organiza en un círculo según sus relaciones de compatibilidad y conflicto. Estos valores son creencias sobre estados deseables que trascienden situaciones específicas y guían la conducta de los individuos en una sociedad. Otro ejemplo es la teoría de los fundamentos morales (*Moral Foundations Theory*, MFT) procedente de la psicología social que propone un marco para explicar cómo la moralidad humana surge de "fundamentos" innatos y universales (Haidt and Joseph, 2004; Graham *et al.*, 2013). Otros autores abogan por la existencia de valores específicos que son relevantes en determinados contextos, y que van emergiendo continuamente en diferentes dominios (Van de Poel, 2021). De forma similar, la metodología del diseño sensitivo a valores, mencionada anteriormente, por ejemplo, argumenta que los valores son conceptos abiertos que deben ser extraídos por las partes interesadas en cada dominio específico.

Holgado-Sánchez *et al.* (2025, 2026) definen un modelo formal en el que ambos enfoques pueden ser retratados partiendo de un conjunto finito $V = \{v_1, \ldots, v_m\}$ de "etiquetas" v_i que representan *valores abstractos* y que toman un significado específico en un contexto. Por ejemplo, en el contexto del transporte, los valores relevantes podrían ser *eficiencia* (Eff), *sostenibilidad* (Sos) y *seguridad* (Seg). Comúnmente, un *sistema de valores* (abstracto) se modela a través de algún tipo de relación de preferencia sobre V. Por ejemplo, el sistema de valores de un agente puede indicar que para él la *sostenibilidad* es más importante que la *seguridad*, y esta a su vez es más importante que la *eficiencia*, es decir $Sos \succcurlyeq Seg \succcurlyeq Eff$.

Para que este sistema de valores abstracto adquiera un significado práctico, hay que interpretarlo en un dominio concreto. En particular, si en un dominio un agente ha de elegir entre diferentes alternativas, el sistema de valores induce un orden débil de preferencia sobre éstas alternativas. Por ejemplo, si para realizar un viaje se puede elegir entre $Metro$, $Coche$ o $Bici$, el sistema de valores abstracto VS arriba indicado podría implicar que $Metro \succcurlyeq_{VS} Bici \succcurlyeq_{VS} Coche$. Para poder razonar de forma automática en base a un sistema de valores abstractos, el modelo ha de especificar cómo se deduce esta relación de preferencia sobre las alternativas del dominio a partir de este sistema.

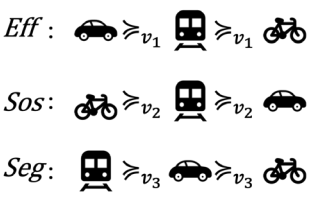
Para ello, primero hay que dar un significado concreto a cada valor en el dominio. Se considera que cada valor v_i establece un orden débil de preferencias $\preccurlyeq_{v_i}$ sobre las alternativas del dominio. La figura 5a muestra un ejemplo para la elección del medio de transporte: con respecto a la eficiencia, se prefiere el coche sobre el metro sobre la bici, con respecto a la sostenibilidad la bici sobre el metro sobre el coche, y con respecto a la seguridad el metro sobre el coche sobre la bici. Una *función de alineamiento con un*

valor A_v asigna un número a cada una de las opciones, de tal forma que este número es más alto cuanto más preferida sea una opción con respecto al valor v. Lo importante es que esta función ha de ser computacional, es decir, se debe poder calcular a partir de características de cada alternativa. Por ejemplo, la función de alineamiento con el valor de sostenibilidad se podría definir en base a la cantidad de CO_2 que produce cada una de las alternativas, de modo que sea inversamente proporcional a esta característica. El conjunto G_V de todas las funciones de alineamiento con un valor se llama *grounding* del sistema de valores en un dominio, ya que da un significado concreto y computable a todas las (etiquetas de) valores del sistema de valores abstracto. La figura 5b muestra un *grounding* compatible con el ejemplo. Muchas veces se asume que este *grounding* de los valores es compartido entre todos los agentes de una sociedad, es decir, que todos los individuos entienden (aproximadamente) lo mismo con respecto a lo que significa que un medio de transporte sea más o menos eficiente, seguro y sostenible.

FIGURA 5

VALORES Y *GROUNDING* G_v

A)



B)

A_v	*Eff*	*Sos*	*Seg*
	10	2	9
	7	8	10
	4	10	6

Fuente: Elaboración propia.

El significado del sistema de valores de un agente j en un dominio se define de forma similar ya que implica un orden débil de preferencias $\preccurlyeq^j_{G_V}$ sobre las alternativas, en este caso reflejando las preferencias particulares del agente j. La *función de alineamiento con el sistemas de valores* A_{f_j,G_V} de un agente j se calcula a partir del *grounding* G_V con una función de agregación f_j, de tal forma que representa el orden débil correspondiente. Es decir, para todas las alternativas e y e' se cumple

$$A_{f_j,G_V}(e) \leq A_{f_j,G_V}(e') \Leftarrow e \preccurlyeq^j_{G_V} e' \qquad [2]$$

A menudo se modela f_j como una función lineal, dando diferentes pesos w_i a cada valor v_i en una suma ponderada. Los pesos se pueden interpretar como la importancia que un agente j da a cada uno de los valores en su sistema de valores. Por ejemplo, el agente j podría dar un peso de 0.2 al valor de eficiencia, un peso de 0.5 al valor de

sostenibilidad, y un peso de 0.3 al valor de seguridad. En consecuencia, sobre la base del *grounding* de la figura 5b, obtendríamos $A_{f_j,G_V}(Coche) = 5.7$; $A_{f_j,G_V}(Metro) = 8.4$; y $A_{f_j,G_V}(Bici) = 7.6$; lo cual coincide con el orden de preferencias indicado al inicio del ejemplo. Nótese que, a pesar de que en el sistema de valores abstracto el valor más preferido es la *sostenibilidad* (Sos), en el dominio del ejemplo la opción más preferida es $Metro$, aunque es menos preferida que $Bici$ con respecto al valor Sos (ver figura 4a). Sin embargo, respecto a los dos valores Eff y Seg la opción $Metro$ se prefiere sobre $Bici$ lo cual, dados los pesos que definen la función de agregación, compensa la preferencia de $Bici$ sobre $Metro$ respecto al valor Sos .

Se ha definido lo que es un sistema de valores de un agente, pero falta formalizar qué se entiende por un sistema de valores de una sociedad. La solución más inmediata es concebirlo simplemente como una agregación (un promedio) de los sistemas de valores de los miembros de la sociedad. Sin embargo, esto podría no reflejar adecuadamente la diversidad existente en ella. Por ejemplo, una sociedad con dos grupos claramente diferenciados y de tamaño similar (p.ej., "izquierdistas" y "derechistas") es claramente distinta de una sociedad donde todos los miembros comparten un sistema de valores intermedio ("centristas"). Para tener en cuenta esta peculiaridad, en vez de representar las preferencias éticas de una sociedad a través de un único sistema de valores, Holgado-Sánchez *et al.* proponen modelarlas mediante un conjunto de sistemas de valores: uno para cada subgrupo "relevante" que exista en la sociedad (Holgado-Sánchez *et al.*, 2025).

3.3. Aprendizaje de sistemas de valores

Uno de los mayores obstáculos para la aplicación práctica de un modelo computacional de sistemas de valores, como el que se ha esbozado en la sección anterior, reside en la dificultad de identificar tanto la interpretación o el significado de un conjunto de valores como los sistemas de valores concretos que existan en una sociedad, en un dominio de interés concreto. Se puede obtener esta información y codificarla "a mano" (por ejemplo, a partir de encuestas o entrevistas con las personas involucradas), pero este enfoque es delicado por varias razones: i) los valores son conceptos abstractos cuya interpretación varía entre distintos contextos y grupos sociales; ii) las personas generalmente no tienen una idea clara de qué valores realmente influyen en sus decisiones (no suelen tener una idea clara de su propio sistema de valores); y iii) la dificultad en la calibración de estas instancias de sistemas de valores.

Para solventar este problema, en (Liscio *et al.*, 2023) se propone un proceso genérico denominado *inferencia de valores* para identificar el sistema de valores de una sociedad de agentes. El proceso se estructura en tres fases:

1. Identificación de valores, es decir, el proceso de encontrar el conjunto de valores $V = \{v_1, \ldots, v_m\}$ relevante para un contexto determinado. Este proceso requiere llevar a cabo conversaciones con expertos del dominio, encuestas o un análisis de documentos mediante técnicas de procesado de lenguaje natural.

2. Estimación de sistemas de valores, es decir, el proceso de identificar los sistemas de valores de los agentes de la sociedad. El problema típicamente se traduce en la especificación de un orden débil de preferencia entre los valores, o bien un sistema de pesos que represente la importancia de cada valor para cada individuo del sistema modelado, como se ha visto en la sección anterior.

3. Agregación de sistemas de valores, es decir, el problema de encontrar una representación consensuada del sistema de valores de la sociedad, por ejemplo, mediante técnicas de elección social basadas en consenso.

Un aspecto clave en el proceso de inferencia de valores es la especificación del significado de cada valor en términos de las características del dominio. Un posible enfoque consiste en concebir esta tarea como un problema de *aprendizaje de valores* (Soares, 2019). Holgado-Sánchez *et al.* siguen este enfoque y, asumiendo el modelo descrito en el apartado 3.2, proponen aprender tanto el significado de un conjunto de valores abstractos como los sistemas de valores de agentes (Holgado-Sánchez *et al.*, 2025). En otro trabajo (Holgado-Sánchez *et al.*, 2026), se extiende este planteamiento al aprendizaje del sistema de valores de una sociedad.

En estos artículos, una vez identificado un conjunto de valores $V = \{v_1, \ldots, v_m\}$ mediante técnicas de *identificación de valores*, el problema de aprendizaje de valores consiste en aprender un *grounding* G_V que aproxime la conceptualización de los valores por parte de los agentes de la sociedad. Para ello, se plantea aprender las funciones de alineamiento A_{v_i} a partir de evaluaciones (no necesariamente cuantificadas) entre alternativas basadas en cada uno de los valores. Es decir, los conjuntos de datos de entrada consisten en pares de alternativas, junto con la valoración de un agente sobre cuál de estas alternativas está más alineada con cada valor (según su entendimiento). Estos datos se obtienen a través de observaciones y/o mediante interacción directa con los agentes de la sociedad. Así se evita tener que recurrir a procedimientos más costosos como encuestas o estimaciones directas de funciones de alineamiento.

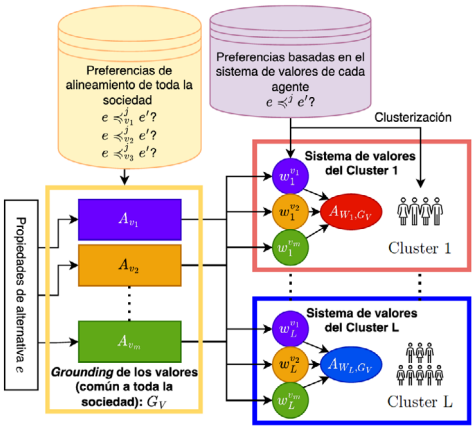
Además, siguiendo un proceso similar, en (Holgado-Sánchez *et al.*, 2025) se propone aprender una representación de los sistemas de valores de la sociedad: dados ejemplos de las preferencias entre alternativas según el propio sistema de valores de cada agente (es decir, tomando sus preferencias tras ponderar todos los valores), se puede estimar una representación de la importancia general que cada valor tiene en esa sociedad para distintos grupos de agentes.

Mediante técnicas de regresión logística (profunda), se estima un modelo para cada uno de estos conjuntos de datos y agentes. El problema principal consiste en cómo agrupar los agentes en distintos *clústers* de forma que todos ellos vean representado "suficientemente bien" su sistema de valores mediante un sistema de valores común (un mismo sistema de pesos definido sobre el *grounding* compartido). Un conjunto de *clústers* con sus respectivos sistemas de valores se define como el *sistema de valores de una sociedad*.

La figura 6 resume el esquema de aprendizaje. Cada alternativa se evalúa mediante una función de alineamiento con cada valor (A_{v_i} , $i = 1,...,m$), que constituyen un *grounding* de dichos valores. Instancias de este *grounding* se aprenden a partir de ejemplos de valoraciones entre alternativas, en los que los agentes j de la sociedad indican qué alternativa está más alineada con cada valor. Por otro lado, el sistema de valores de la sociedad se define como un conjunto de L sistemas de valores definidos, y su asignación a distintos grupos de agentes (*clústers*). Siguiendo el modelo descrito en el apartado 3.2, estos sistemas de valores se definen a su vez mediante distintos conjuntos de pesos normalizados ($W_l = (w_l^{v_1},\ldots,w_l^{v_m})$ con $w_l^{v_i} \in [0,1]$ y $\sum_{i=1}^{m} w_l^{v_i} = 1$) que estiman la importancia relativa de cada valor. La función de alineamiento del sistema de valores de cada *clúster* evalúa cada alternativa con una agregación lineal del *grounding* ponderada con estos pesos. Es decir, para cada *cluster* l, su función de alineamiento está dado por $A_{W_l,G_V}(e) = W_l \cdot G_V(e)^T$. Tanto el reconocimiento de *clústers* como el aprendizaje de los conjuntos de pesos correspondientes a cada uno de los mismos se realizan a partir de la observación de ejemplos de preferencias entre alternativas, en los que cada agente de la sociedad expresa qué alternativa está más alineada sobre la base de su propio sistema de valores.

FIGURA 6

REPRESENTACIÓN DEL MODELO SISTEMA DE VALORES DE UNA SOCIEDAD Y DE LOS DATOS NECESARIOS PARA EL APRENDIZAJE DE UNA INSTANCIA DEL MISMO



Fuente: Elaboración propia.

Durante el proceso de aprendizaje se buscan funciones de alineamiento A_{v_i} , *clústers* de agentes, y conjuntos de pesos W_l , tal que se maximizan tres métricas cuantitativas:

1. Coherencia del *grounding*: expresa en qué medida se respeta el entendimiento de los valores de cada agente, es decir, en qué grado las funciones de alineamiento representan los ejemplos de preferencias de cada agente.

2. Representatividad: determina en qué medida es representado el sistema de valores individual de cada agente por el sistema de valores de su *clúster*. Se mide en función del número de coincidencias entre las preferencias indicadas de cada agente y la función de alineamiento del sistema de valores del *clúster* al que es asignado.

3. Concisión: mide la complejidad de la representación de sistemas de valores obtenida. Formalmente, la concisión se determina mediante las distancias entre las preferencias que inducen los sistemas de valores aprendidos en el espacio de alternativas, y se quiere maximizar para garantizar que la existencia de cada uno de los sistemas de valores sea realmente relevante. Maximizando la concisión, también se tiende a minimizar el número de *clústers*.

En Holgado-Sánchez *et al.* (2025), se evalúa el esquema de aprendizaje propuesto usando un conjunto de datos reales sobre la elección de rutas de tren en Suiza, en el que participaron 388 personas (agentes) (Vrtic and Axhausen, 2002). En este conjunto de datos, cada agente indica su ruta preferida entre dos alternativas en nueve situaciones distintas (es decir, un total de 3.492 observaciones). Cada ruta se describe mediante cuatro atributos: tiempo de viaje, coste, número de transbordos necesarios, y tiempo de espera entre trenes. Además, en este conjunto de datos se recogieron seis características contextuales de cada agente: nivel de ingresos del hogar, disponibilidad de coche, y el propósito del viaje (desplazamiento al trabajo, compras, negocios u ocio, siendo estos últimos excluyentes entre sí). En el análisis, cada ruta se considera una opción posible y el conjunto de participantes forma la población estudiada. Los datos reflejan únicamente qué ruta prefiere cada agente en cada comparación, es decir, qué ruta de cada par escogería el agente.

Se asume que las decisiones sobre las rutas están guiadas por tres valores principales: *eficiencia temporal*, *eficiencia económica*, y *comodidad*. La eficiencia temporal y la económica se definen directamente a partir del tiempo de viaje y del coste, respectivamente. En cambio, la comodidad se relaciona con el tiempo de espera entre trenes y el número de transbordos: una ruta se considera más cómoda si tiene menor tiempo de espera y menos transbordos. Cuando solo uno de estos dos aspectos mejora y el otro no, no se establece una preferencia entre las rutas y se deja que el modelo aprenda libremente cómo se valora la comodidad. A partir de estas definiciones, se construye el conjunto de datos de valoraciones de alternativas para el aprendizaje de las funciones de alineamiento A_{v_i} a partir de los 3.492 pares de alternativas originales.

Al utilizar los algoritmos de aprendizaje sobre este conjunto de datos, se obtuvo una solución con tres sistemas de valores para sendos grupos de agentes (*clústers*). Estos *clústers* exhibieron ciertas regularidades interesantes, como por ejemplo que los agentes que fueron asignados a un sistema de valores que daba la mayor importancia a la eficiencia económica, solían elegir los trayectos con la intención de realizar compras en el destino, mientras que los agentes que preferían el valor de la eficiencia temporal típicamente realizaban trayectos con fines de negocios. Esa información contextual (viajes de compras/negocios) no se tuvo en cuenta en el proceso de aprendizaje y *clusterización*, lo cual resalta que el modelo y algoritmos empleados no solo maximizan métricas cuantitativas sino también reconocen patrones implícitos de preferencias.

4. CONCLUSIONES

Una característica clave de los agentes IA es su capacidad para tomar decisiones y actuar de forma *autónoma* en los entornos en los que están embebidos. Generalmente, para alcanzar sus objetivos y realizar las tareas que les han sido encomendadas, los agentes IA han de interactuar con otros agentes y, a veces, también directamente con las personas. En este capítulo se ha argumentado que una *gobernanza* efectiva para los sistemas multiagente y sociotécnicos resultantes es fundamental para que las aplicaciones basadas en IA sean seguras, responsables y confiables.

Se ha razonado que, para alcanzar una coordinación adecuada entre los agentes IA, los diseñadores han de actuar sobre su entorno, y para muchos problemas esto se traduce en imponer unos protocolos de interacción adecuados sobre las infraestructuras de comunicación que facilitan el intercambio de mensajes de los agentes. Además, se ha ilustrado esta idea con un ejemplo de redes de gestores de intersección basados en reservas, para un hipotético futuro con coches autónomos guiados por agentes IA. También se ha comentado que, aparte de una serie de problemas técnicos, enfoques de estas características tienen implicaciones legales que se han de afrontar tanto desde un punto de vista técnico como regulatorio.

Por otra parte, se ha ilustrado que los problemas éticos en sistemas sociotécnicos con agentes IA son más relevantes de lo que pueda aparentar a primera vista. En este sentido, es deseable asegurar el alineamiento de las acciones de los agentes IA con el sistema de valores de la sociedad; un objetivo que se puede conseguir a través del concepto de los *agentes conscientes de los valores*. Para ello, se ha esbozado un enfoque computacional para modelar el significado de valores éticos en dominios concretos, y cómo se puede construir a partir de allí un modelo formal del sistema de valores de un agente o de una sociedad. Para instanciar este tipo de modelos, se ha presentado una propuesta para aprender el significado de los valores abstractos (su semántica computacional), así como el sistema de valores de una sociedad de agentes.

Quedan varias líneas de investigación abiertas en el ámbito del aprendizaje de sistemas de valores que merecen ser exploradas. Una de ellas es la ampliación de los métodos de aprendizaje para su aplicación en entornos más complejos con decisiones *secuenciales*. Partiendo del modelo arriba expuesto, en (Holgado-Sánchez *et al.*, 2026) se presenta y evalúa un método para ello, pero manteniendo la asunción de una función de agregación lineal. Otra línea consiste en hacer explícita la dependencia del *contexto* de los valores y de los sistemas de valores de los agentes, es decir, reconocer situaciones donde el sistema de valores de un agente cambie radicalmente ante cambios contextuales, como podrían ser casos de emergencia o durante interacciones con otros agentes. Por otro lado, en las propuestas presentadas se asume que exista un cierto consenso entre los individuos de la sociedad acerca del significado o de la interpretación de cada valor. Por ejemplo, eficiencia temporal, tiene una clara interpretación común: llegar antes al destino. En este sentido, puede ser de interés estudiar en más detalle qué dificultades o implicaciones puedan existir en sociedades más *heterogéneas*, es decir, en las que se consideren valores cuya interpretación varíe radicalmente entre grupos, como, por ejemplo, valores de *justicia* o *libertad*.

Otra línea de investigación relevante en este ámbito es el estudio de la relación entre valores y *políticas regulatorias*, donde estas últimas se caracterizan a través de los reglamentos y normas que las implementan. Para ello, conviene partir del gran abanico de enfoques propuestos en las últimas décadas en el ámbito de los sistemas multiagente para modelar normas, tanto sociales como legales. El trabajo de Montes y Sierra (2022) es pionero en este ámbito al proponer métodos para la síntesis automatizada de conjuntos de normas (políticas) que promocionan el alineamiento de un sistema multiagente con un conjunto de valores. Un dominio interesante para demostrar la relevancia social de este tipo de enfoques es el dominio de la asignación de plazas escolares. En España, una ley estatal regula el marco general para el proceso de asignación, pero, al ser un país políticamente descentralizado, varias comunidades autónomas han establecido sus propios reglamentos para la implementación efectiva de este proceso. Los resultados de dichos reglamentos son públicamente accesibles bajo la Ley de Transparencia (Moreno Rebato and Rodríguez García, 2025) y constituyen un cuerpo de datos valioso para analizar los sistemas de valores subyacentes a los diferentes reglamentos existentes, es decir, analizar qué tipo de valores se promueven en unos casos o en otros.

Referencias

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J-F., and Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.

Dötterl, J., Bruns, R., Dunkel, J., and Ossowski, S. (2020). On-time delivery in crowdshipping systems: An agent-based approach using streaming data. In *Proc. Europ. Conf. on Artificial Intelligence (ECAI'20)*, 51–58. IOS Press.

Dresner, K., and Stone, P. (2008). A multiagent approach to autonomous intersection management. *J. Artif. Int. Res.*, *31*(1), 591–656.

Friedman, B., Kahn, P. H., Borning, A., and Huldtgren, A. (2013). *Value Sen-sitive Design and Information Systems*, 55–95. Dordrecht: Springer Netherlands.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In P. Devine and A. Plant (editors), *Advances in Experimental Social Psychology*, volume 47, (55–130). Academic Press.

Haidt, J., and Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4),55–66.

Holgado-Sánchez, A., Billhardt, H., Fernández, A., and Ossowski, S. (2026). Learning the value systems of agents with preference-based and inverse reinforcement learning. *Auton. Agent Multiagent Syst., 40*(4).

Holgado-Sánchez, A., Billhardt, H., Ossowski, S. and Degli-Esposti, S. (2025). Learning the value systems of societies from preferences. In *Proc. Europ. Conf. on Artificial Intelligence (ECAI'25)*, 1121–1130.

Holgado-Sánchez, A., Vamplew, P., Dazeley, R., Ossowski, S., and Billhardt, H. (2026). Learning the value systems of societies with preference-based multi-objective reinforcement learning. In *Proc. Int. Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS'26)*. To appear.

Liscio, E., Lera Leri, R. X., Bistaffa, F., Dobbe, R. I. J., Jonker, C. M., López-Sánchez, M., Rodríguez-Aguilar, J. A., and Murukannaiah, P. K. (2023). Value inference in sociotechnical systems. In *Proc. Int. Conf.on Autonomous Agents and Multiagent Systems (AAMAS'23)*, 1774–1780.

Montes, N., and Sierra, C. (2022). Synthesis and properties of optimally value-aligned normative systems. *J. Artif. Intell. Res.*, 74, 1739–1774.

Moreno Rebato, M., and Rodríguez García, J. A. (2025). El acceso a los datos en los procesos de asignación de plazas escolares con fines de investigación científica: La utilización de la inteligencia artificial en este contexto. *Revista Española de la Transparencia*, 21, 129–156.

Namazi, E., Li, J., and Lu, CH. (2019). Intelligent intersection management systems considering autonomous vehicles: A systematic literature review. *IEEE Access*, 7, 91946–91965.

Osman, N. (2025). Value-aware multiagent systems. In S. Cranefield, L. G. Nardin, and Lloyd N. (editors), *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of MultiAgent Systems XVII*, 32–37. Springer Nature Switzerland.

Sandholm, T. W. (1999). *Distributed rational decision making*, 201–258. Cambridge, MA, USA: MIT Press.

Santos, J-A., Fernández, A., Moreno-Rebato, M., Billhardt, H., Rodríguez-García, J. A., and Ossowski, S. (2020). Legal and ethical implications of applications based on agreement technologies: The case of auction-based road intersections. *Artif. Intell. Law, 28*(4), 385–414.

Schumacher, M., and Ossowski, S. (2006). The governing environment. In D. Weyns, H. Van Dyke Parunak, and F. Michel (editors), *Environments for Multi-Agent Systems II*, 88–104. Springer.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, 1–65. Elsevier.

Soares, N. (2019). The Value Learning Problem. In R. V. Yampolskiy (editor), *Artificial Intelligence Safety and Security*, chapter 7, 89–98. CRC Press.

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

Van de Poel, I. (2021). Design for value change. *Ethics and Information Technology, 23*(1), 27–31.

Vasirani, M., and Ossowski, S. (2012). A market-inspired approach for intersection management in urban road traffic networks. *J. Artif. Int. Res., 43*(1), 621–659.

Vasirani, M., and Ossowski, S. (2013). A proportional share allocation mechanism for coordination of plug-in electric vehiclecharging. *Engineering Applications of Artificial Intelligence, 26*(3), 1185–1197.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance, 16*(1), 8–37.

Vrtic, M., and Axhausen, K. W. (2002). The impact of tilting trains in Switzerland. a route choice model of regional and long distance public transport trips. Technical report, IVT - ETH Zurich. DOI: 10.3929/ethz-a-004403563.

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems*, 2nd edition. Wiley.