CAPÍTULO VII

Máquinas que saben lo que importa: ingeniería de la conciencia de valores sociales

Carles Sierra*

La integración ética de la inteligencia artificial (IA) en la sociedad requiere superar tanto el catastrofismo apocalíptico como las limitaciones de los enfoques éticos puramente normativos. Este capítulo argumenta que el camino viable no reside en dotar a las máquinas de *conciencia fenomenológica (consciousness)*, sino de *capacidad operativa (awareness)* para reconocer y actuar conforme a los valores humanos. Partiendo de la distinción filosófica crítica entre ambos conceptos, se defiende que es posible y necesario construir agentes *value-aware* (conscientes de valores) sin atribuirles un estatus moral inherente. Para ello, se presenta un marco de ingeniería centrado en la Incrustación de Valores *(Value Embedding)*, que supera la rigidez del Diseño Sensible a Valores al internalizar representaciones computacionales de los valores, permitiendo una adaptación autónoma y contextual.

El núcleo de la propuesta es la arquitectura Reflective Global Neuronal Workspace (RGNW), inspirada en la neurociencia cognitiva y estructurada en tres niveles (reactivo, deliberativo y reflectivo). Este modelo no solo ejecuta comportamientos, sino que reflexiona sobre ellos a la luz de un sistema de valores, gestionando la coherencia entre preferencias axiológicas, normas derivadas y acciones. La RGNW dota al agente de una teoría de la mente artificial que le permite inferir los valores y estados mentales de sus interlocutores, y lo posiciona como un participante activo en la negociación social de valores. Así, el capítulo trasciende la visión de la IA como mero ejecutor de reglas, presentándola como un ente capaz de comprender, adaptarse y contribuir al dinámico ecosistema moral humano. La viabilidad del enfoque se demuestra con aplicaciones prácticas en análisis de redes sociales, robótica asistencial y herramientas de apoyo a la decisión clínica, señalando un camino pragmático y urgente hacia una IA socialmente responsable.

Palabras clave: conciencia de valores, sistemas *value-aware*, RGNW, inteligencia artificial, ingeniería.

* Este trabajo ha sido financiado por el proyecto Europeo VALAWAI (#101070930) y por el proyecto nacional VAE (#TED2021-131295B-C31).

1. INTRODUCCIÓN: MÁS ALLÁ DEL DISCURSO APOCALÍPTICO

En el debate contemporáneo sobre la inteligencia artificial (IA), resulta frecuente encontrar un discurso catastrofista, promovido en ocasiones por figuras prominentes y grandes corporaciones, que predica el advenimiento de una superinteligencia que supondría el fin de la humanidad (Bostrom, 2014). Este relato, a menudo revestido de un velo seudofilosófico, se sustenta en la hipótesis de que la IA alcanzará un estado de conciencia (*consciousness*) que la llevará a concluir que el mundo funcionaría mejor sin seres humanos. Es imperioso combatir esta visión, no solo por sus posibles intereses económicos subyacentes, y la desinformación que provoca en la sociedad, sino también por su frágil base conceptual.

El objetivo de este capítulo es doble. En primer lugar, se argumentará, desde planteamientos filosóficos actuales, que la IA no alcanzará la conciencia en los términos en que se discute para los seres vivos. En segundo lugar, y este es el núcleo de nuestra propuesta, se defenderá que la falta de conciencia no es un impedimento para desarrollar sistemas de IA moralmente responsables. Es perfectamente posible, y deseable, construir agentes artificiales cuyo comportamiento sea éticamente aceptable y esté alineado con los valores humanos, sin necesidad de atribuirles un estatus moral. Este enfoque se enmarca en proyectos de investigación europeos, como los financiados en el programa EIC Pathfinder "Awareness Inside", https://awarenessinside.eu), que buscan conectar la neurociencia y la IA para explorar la definición y la ingeniería de un cierto tipo de "atención consciente" y de los valores en sistemas computacionales.

Para ello, nos centraremos en el concepto de *awareness* (conciencia en el sentido de "darse cuenta" o "ser consciente de"), distinto de *consciousness* (conciencia plena o fenomenológica), y su aplicación a los valores humanos. Abordaremos la diferencia entre las aproximaciones del Diseño Sensible a Valores (*Value Sensitive Design*), que no usaremos, y la de la Incrustación de Valores (*Value Embedding*). Detallaremos una arquitectura computacional inspirada en el "Global Neuronal Workspace" (GNW) (Baars and Alonzi, 2018) para lograr sistemas *value-aware*. Finalmente, ilustraremos la viabilidad de este marco con aplicaciones concretas en redes sociales, robótica social y protocolos médicos.

2. DESENTRAÑANDO LA CONCIENCIA: *AWARENESS* VS. *CONSCIOUSNESS* Y EL ESTATUS MORAL

Un primer paso crucial es la distinción terminológica y conceptual. En inglés, y por extensión en la literatura especializada, se manejan dos términos que en castellano se traducen comúnmente como "conciencia": *consciousness* y *awareness*.

- *Consciousness*. Hace referencia a la conciencia fenomenológica, el estado subjetivo de experimentar, de tener una experiencia en primera persona. Está intrínsecamente ligado al concepto de *qualia* (las cualidades subjetivas de las experiencias, como el rojo de una rosa o el dolor de una herida). La *consciousness* es un campo de estudio en el que neurocientíficos y filósofos no han alcanzado un consenso, con decenas de definiciones en pugna (Chalmers, 1996).

- *Awareness*. Se refiere a un concepto más operacional y menos enigmático. Es la capacidad de "darse cuenta" de algo, de tener acceso a una representación de un elemento (un objeto, un hecho, un valor) y poder procesarlo cognitivamente. En el contexto de la IA, buscamos construir sistemas que sean *aware* de aspectos específicos, como la presencia de valores éticos en su entorno.

Esta distinción es fundamental para desacoplar la discusión sobre el comportamiento moral de la IA de la espinosa cuestión de su posible conciencia. Nuestra propuesta se centra exclusivamente en el *awareness*, no en la *consciousness*.

La cuestión del estatus moral de las entidades no humanas es central en la ética filosófica. Según el investigador Ali Ladac (Ladak, 2024), una entidad tiene categoría moral (*Moral standing* en su versión en inglés) si ella, o sus intereses, importan de manera intrínseca, al menos hasta cierto grado, en la evaluación moral de las acciones. Dos tipos de entidades se ajustan a esta definición:

- *Entidades sintientes* (i.e. con capacidad de sentir): la sintiencia es la capacidad de una entidad de tener experiencias positivas o negativas de manera intrínseca. Es decir, consideramos aquí a entidades para las cuales una experiencia puede ser per se buena o mala. Esta capacidad se asocia típicamente a los seres vivos.

- *Entidades conscientes:* como se ha definido, aquellas con capacidad de tener experiencias subjetivas.

Bajo esta visión, ejemplos de entidades que poseen categoría moral serían los humanos conscientes (incluso bajo los efectos de drogas, por su potencialidad de recuperación) y, de manera cada vez más aceptada, los animales. De hecho, legislaciones como la del Reino Unido han reconocido formalmente que los animales domésticos (en 2021) y los pulpos (en 2024) son *sentient*, un primer paso hacia la concesión de derechos. Una roca, una nación, serían ejemplos de entidades sin categoría moral.

El filósofo David Chalmers (Chalmers, 1996) explora esta idea mediante un experimento mental basado en el dilema del tranvía. En su versión, se debe decidir entre desviar un

tren para matar a un humano consciente o a un número **n** de zombies filosóficos (entes funcionalmente idénticos a humanos pero carentes de conciencia y de capacidad de sentir). La mayoría de las personas elige salvar al humano consciente (aunque cuando la **n** es muy grande algunos humanos empiezan a dudar). Esto ilustra que la sintiencia y la conciencia son criterios fundamentales para la atribución de categoría moral.

Aplicando este marco a la IA, se debe concluir que, actualmente, los sistemas de IA no tienen una categoría moral. No son sintientes (no tienen experiencias intrínsecamente positivas o negativas) y no poseen intereses intrínsecos. A una IA le es indiferente si está ejecutándose o no, si consigue o no un objetivo y no tiene experiencias subjetivas de ningún tipo. Por lo tanto, carece de los requisitos necesarios para ser considerada una entidad moral, lo que implica, por ejemplo, que no existe un problema ético en "apagarla".

Cabe mencionar que según otros autores se pueden usar criterios alternativos para atribuir categoría moral, como la posesión de capacidades cognitivas (agencia, inteligencia, racionalidad), la capacidad de mantener relaciones sociales o la performatividad (comportarse *como si* tuviera categoría moral). Con estos criterios nos adentramos en terrenos más resbaladizos en los que la atribución de categoría moral a sistemas de IA avanzados podría ser objeto de debate. No obstante, para los fines de este capítulo, partimos de la premisa de que es posible y deseable construir sistemas *aware* de los valores sin necesidad de atribuirles una categoría moral.

3. LA INGENIERÍA DE SISTEMAS *VALUE-AWARE*

La propuesta central de este capítulo es la ingeniería de sistemas de IA que sean *aware* de los valores humanos. La motivación es clara: si un sistema de IA se despliega en una sociedad con unos valores determinados (por ejemplo, en Japón, España o Estados Unidos), su comportamiento debe adaptarse a respetar el marco axiológico de cada contexto.

3.1. Diseño sensible a los valores vs. incrustamiento de valores

Existen dos aproximaciones principales para incorporar valores en la IA:

1. *Value Sensitive Design* (VSD). En este enfoque los valores se identifican y se presentan explícitamente a los ingenieros, quienes los incorporan en las fases de diseño y prueba del sistema. El VSD es valioso, pero puede ser rígido, ya que requiere volver a la mesa de diseño cada vez que el contexto de aplicación cambia.

2. *Value Embedding* (Incrustación de valores). Esta es la aproximación que defendemos. Los valores se representan computacionalmente de manera que el sistema los tiene "embebidos" o internalizados. Esto le permite adaptarse autónomamente a situaciones nuevas, tomando decisiones alineadas con esos valores sin necesidad de una reingeniería constante.

Una definición informal de un sistema *value-aware* sería: un sistema de IA capaz de capturar un sistema de valores dado (como la Teoría de los fundamentos morales[1] o el sistema de Schwartz[2]), entenderlo, respetarlo en sus acciones y explicar su comportamiento (y el de otros) en términos de este conjunto de valores. Los beneficios de tales sistemas se presentan en la tabla 1.

TABLA 1

BENEFICIOS DE UNA IA *VALUE-AWARE*

Función	*Descripción*
Monitorización	Verificar si las acciones propias o ajenas respetan un conjunto de valores.
Asesoramiento	Recomendar a un humano acciones compatibles con ciertos valores, justificándolas.
Certificación	Avalar si un sistema o proceso cumple con los valores establecidos.
Aprendizaje	Identificar e inferir los valores y normas implícitas en un grupo humano a partir de su comportamiento.
Autonomía guiada	Dotar de autonomía a los sistemas para la toma de decisiones, con garantías de respeto a los valores.

3.2. Una definición computacional de *awareness*

Para que el concepto de *awareness* sea operativo, necesitamos una definición computacional. En el marco del proyecto VALAWAI (https://valawai.eu/), hemos establecido que un sistema puede considerarse *aware* de un elemento *X* en un contexto *C* si cumple con tres condiciones:

1. *Acceso a una representación:* el sistema tiene acceso a una representación computacional de *X* en el contexto *C*.

2. *Coherencia:* dicha representación es coherente, sólida (*sound*) y carente de ambigüedades críticas.

[1] *Teoría de los fundamentos morales* (MFT) (Haidt y Graham, 2007): postula cinco (luego seis) pares de fundamentos intuitivos que subyacen a los juicios morales: Cuidado/Daño, Justicia/Engaño, Lealtad/Traición, Autoridad/Subversión y Santidad/Degradación.

[2] *Teoría de los valores humanos básicos de Schwartz* (Schwartz, 2012): propone una estructura circular de diez valores motivacionales universales (por ejemplo, Autodirección, Seguridad, Logro, Benevolencia) que se relacionan por compatibilidad o conflicto.

3. *Expresabilidad*: la representación puede expresarse en términos comprensibles para los humanos, permitiendo el diálogo y la justificación de las acciones.

Esta definición traslada el problema filosófico a uno de ingeniería: cómo representar, hacer coherente y expresar un valor.

3.3. El triángulo valores-normas-comportamiento: el *moral stance computacional*

En nuestra aproximación, la representación de un valor *V* se basa en dos pilares:

1. *Preferencias sobre estados del mundo.* Un valor se entiende como una preferencia sobre los posibles estados del mundo. Por ejemplo, el valor "igualdad de género" prefiere estados del mundo donde mujeres y hombres reciben el mismo salario por el mismo trabajo frente a estados donde esto no ocurre.

2. *Normas alineadas.* Las normas son la traducción operacional y deontológica de los valores. Prescriben el comportamiento correcto (por ejemplo, "debe pagarse el mismo salario por el mismo trabajo").

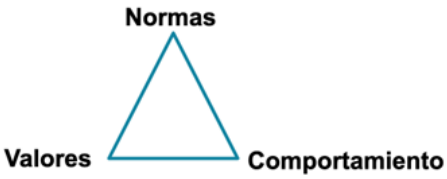
La interacción entre valores, normas y comportamiento forma un triángulo que define el *moral stance* (posicionamiento moral) de un sistema. Permite verificar:

- Si un comportamiento está alineado con un valor.
- Si un comportamiento respeta una norma.
- Si las normas y los valores ordenan los "mundos posibles" de manera consistente.

Este "alineamiento expresable" es la base práctica para construir sistemas *value-aware*.

FIGURA 1

TRIÁNGULO DE POSICIONAMIENTO MORAL



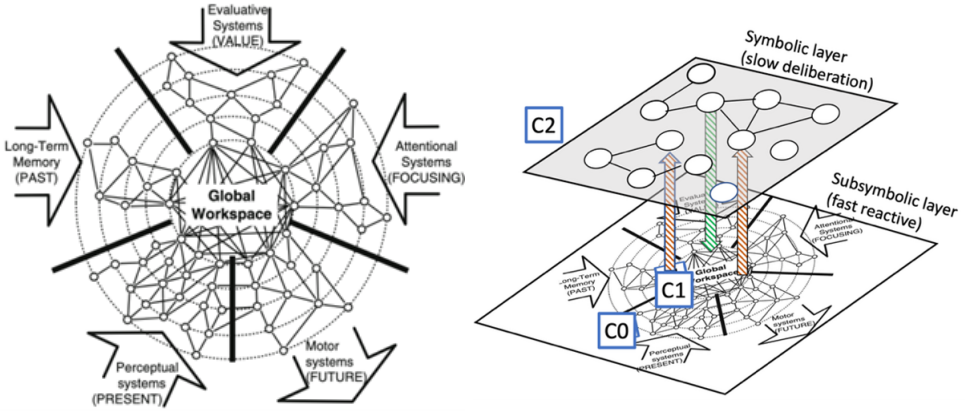
Fuente: Elaboración propia.

3.4. Una arquitectura para la conciencia de valores: el modelo RGNW

Para materializar estos principios, hemos desarrollado una arquitectura inspirada en el *Global Neuronal Workspace (GNW)* de la neurociencia (Baars y Alonzi, 2018), extendido con una capa reflectiva, dando lugar al modelo *Reflective Global Neuronal Workspace (RGNW)* (ver figura 2). Esta arquitectura se inspira también en cierta medida en el modelo dual de Kahneman (pensamiento rápido y lento) (Kahneman, 2011) y en la arquitectura de Brooks (Brooks, 1991). Consta de tres niveles:

FIGURA 2

ESQUEMA DEL MODELO GNW SEGÚN (DEHAENE *ET AL.*, 1998) (IZQUIERDA).
DIAGRAMA DE LA ARQUITECTURA RGNW USADA EN EL PROYECTO VALAWAI (DERECHA)



Fuente: Elaboración propia.

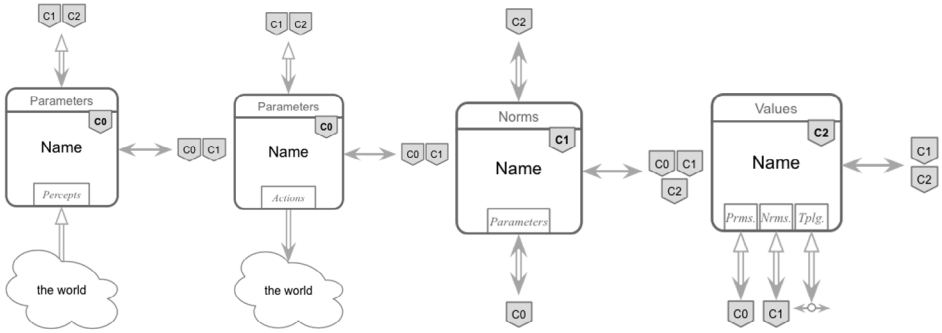
- Nivel C0 (subsimbólico/reactivo). Este nivel contiene componentes modulares que interactúan directamente con el mundo a través de sensores y actuadores. Su funcionamiento es rápido y basado en estímulos. *Ejemplo: Un módulo de* "social sensing" *que analiza tweets en tiempo real para detectar mensajes de odio.*

- Nivel C1 (deliberativo). Aquí la información de los módulos C0 se integra y se procesa de manera más simbólica. Se activan diferentes módulos en función del contexto y la percepción para tomar decisiones que se envían a los actuadores.

- Nivel C2 (reflectivo). Es el nivel superior, que alberga la representación simbólica de los valores. Su función es "reflejar" o monitorizar lo que ocurre en C0 y C1. Basándose en los valores, el nivel C2 puede *modificar la topología de la*

FIGURA 3

ESQUEMAS DE LOS COMPONENTES C0 (SENSOR Y ACTUADOR), C1 Y C2



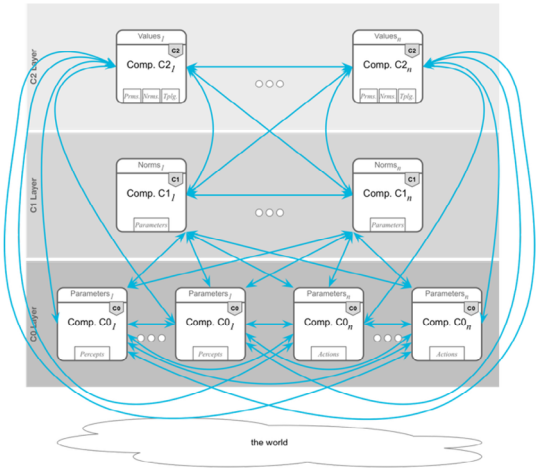
Fuente: Elaboración propia.

arquitectura, alterando qué módulos de C1 se comunican entre sí o con C0. Esto permite una adaptación profunda del comportamiento para alinearlo con los valores, sin reescribir el código de los módulos de nivel inferior.

Esta arquitectura proporciona un marco flexible y potente para implementar sistemas que no solo actúan, sino que también reflexionan sobre sus acciones a la luz de un sistema de valores internalizado. Como se ilustra en las figuras 4 y 5, el modelo RGNW

FIGURA 4

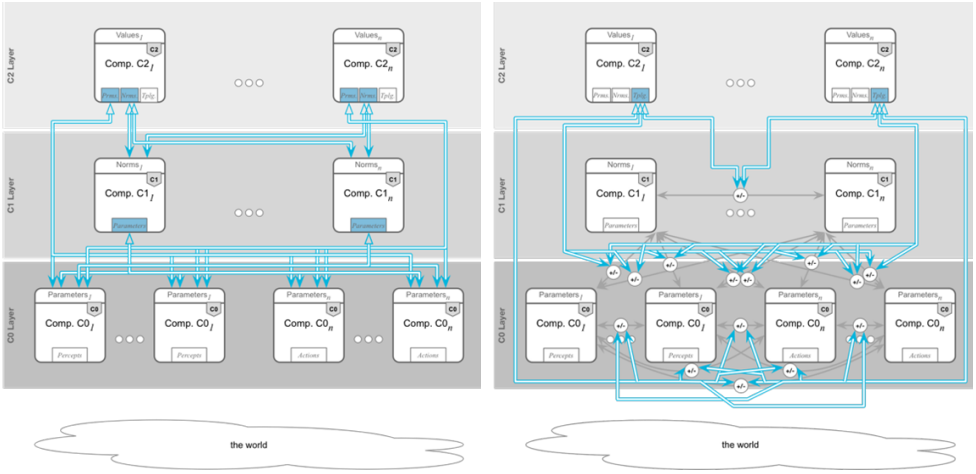
FLUJO DE DATOS EN LA ARQUITECTURA RGNW



Fuente: Elaboración propia.

FIGURA 5

FLUJO DE CONTROL (IZQUIERDA) Y DE LA TOPOLOGÍA (DERECHA) DE LA ARQUITECTURA RGNW



Fuente: Elaboración propia.

combina influencias de la arquitectura GNW y de la arquitectura de subsunción de Brooks, definiendo dos flujos esenciales:

1. *Un flujo de datos*, que garantiza la *representación, el acceso y la expresión* de la información, materializando así los puntos 1 y 3 de la definición operativa de *awareness*.

2. *Un flujo de control*, que, mediante la gestión dinámica de la *topología* de módulos, asegura la *coherencia y solidez* de las representaciones, cumpliendo con el punto 2 de dicha definición.

De este modo, la arquitectura traduce formalmente los fundamentos teóricos a un mecanismo computacional viable para la conciencia de valores.

Desde el punto de vista computacional, la arquitectura funciona como un sistema plenamente concurrente en el que todos los módulos están activos de forma continua e intercambian información. En el nivel C0, los módulos reactivos y subsimbólicos observan constantemente el entorno y procesan los estímulos entrantes, generando elaboraciones parciales que se transmiten a los niveles superiores. Cuando esta información impacta en el nivel C2, los módulos reflectivos la evalúan a la luz del sistema de valores internalizado y pueden emitir directrices normativas hacia C1. De manera crucial, este proceso reflectivo incluye una reconfiguración dinámica de la topología de la arqui-

tectura: el nivel C2 puede activar, inhibir o redirigir las conexiones entre los módulos de C1 y C0. Al modificar estos patrones de interacción, el sistema determina cómo se desarrollarán los cómputos posteriores, introduciendo un mecanismo de metacognición sobre su propio funcionamiento. Este ciclo continuo de flujo de información, adaptación topológica y cómputo concurrente acaba repercutiendo en el nivel C0, donde determinados módulos ejecutan acciones sobre el entorno, ya sean digitales o físicas, cerrando así el bucle percepción–deliberación–acción.

En este sentido, la arquitectura propuesta se aparta deliberadamente del patrón dominante en los sistemas generativos actuales, en los que el comportamiento global queda fundamentalmente determinado por un único algoritmo de aprendizaje automático entrenado a partir de grandes volúmenes de datos de entrada y salida. En lugar de basarse en una función de predicción monolítica, el modelo articula el comportamiento del sistema a través de la interacción dinámica de módulos especializados y de un mecanismo explícito de control y reflexión basado en valores. Esto no excluye, sin embargo, la incorporación de procesos de aprendizaje: el sistema podría integrar bucles de retroalimentación que evalúen las consecuencias de sus acciones en el entorno, permitiendo ajustar tanto las decisiones sobre la topología de la arquitectura como los procesos internos de los propios módulos. De este modo, el aprendizaje no sustituiría a la estructura deliberativa y reflexiva, sino que la complementaría, abriendo una línea de trabajo futura especialmente prometedora para el desarrollo de sistemas capaces de adaptarse de manera progresiva y responsable a contextos cambiantes.

3.5. Aplicaciones prácticas y casos de uso

El marco teórico y la arquitectura RGNW se han validado en varios casos de uso.

3.5.1. Análisis de valores en redes sociales

En colaboración con Sony en París, se ha desarrollado un sistema para analizar la conversación pública en Twitter (actualmente X) utilizando la Teoría de los fundamentos morales (MFT) de Haidt y Graham (2007), que postula cinco pares de valores (Cuidado/Daño, Justicia/Engaño, Lealtad/Traición, Autoridad/Subversión y Santidad/Degradación). A nivel computacional, la arquitectura RGNW se utiliza para estructurar el procesamiento de la información en tres niveles: los módulos del nivel C0 reciben y codifican los datos de entrada (los *tweets* de los usuarios); los módulos C1 procesan y extraen características relevantes, como patrones de comportamiento y expresiones de valores; y el nivel C2 evalúa estas representaciones frente a los valores internalizados, ajustando la interpretación de los *tweets* y corrigiendo, si es necesario, las clasificaciones generadas en C1. Cada usuario se posiciona en un pentágono de valores a partir de esta evaluación, y los resultados muestran una clara distribución ideológica: los perfiles

progresistas se concentran en "Cuidado" y "Justicia", mientras que los conservadores se orientan hacia "Autoridad" y "Lealtad". Esta aplicación, relativamente sencilla dentro del marco RGNW, permite automatizar el mapeo y comprensión de la ecología moral de una red social, manteniendo la coherencia entre la categorización de los *tweets* y los principios y valores definidos en C2 y demostrando la utilidad de la arquitectura para tareas de análisis ético a escala.

3.5.2. Robótica social para el cuidado de personas mayores

En la Universidad de Gante se ha desarrollado un robot social destinado a la interacción en entornos domiciliarios con personas mayores. Equipado con modelos de lenguaje grande (LLMs) y la arquitectura RGNW, el robot mantiene conversaciones con los usuarios respetando valores fundamentales como la dignidad y la privacidad. Los módulos del nivel C0 procesan continuamente las percepciones del entorno y de la interacción con el usuario, mientras que los módulos C1 integran esta información y generan posibles respuestas o acciones. El nivel C2 evalúa estas acciones frente a los valores internalizados y ajusta dinámicamente la topología de comunicación entre los módulos, asegurando que solo las respuestas coherentes con estos valores lleguen a C0. Gracias a este mecanismo, el sistema puede percibir cambios en las temáticas de la conversación—por ejemplo, de hobbies y rutinas a recuerdos personales o cuestiones de salud—indicativos de la creación de un vínculo de confianza. Además, el nivel C2 permite al robot adaptar su tono y contenido, modulando la interacción de manera ética y beneficiosa para el usuario, garantizando que la asistencia no solo sea funcional sino también respetuosa de los valores de la persona.

3.5.3. Brújula de valores en la asistencia médica

En el Hospital del Mar y el Hospital San Pablo, ambos de Barcelona, se ha desarrollado una "brújula de valores" para apoyar la toma de decisiones clínicas. El sistema, basado en los cuatro principios de la bioética[3] analiza el estado de un paciente y las acciones clínicas propuestas. Proporciona un *feedback* al médico, indicando en qué medida cada opción respeta o vulnera estos principios. Esto no sustituye al clínico, sino que le ofrece una herramienta de reflexión para mejorar su práctica, asegurando que las decisiones técnicas también sean éticas.

Desde un punto de vista computacional, la "brújula de valores" implementa un flujo de información jerárquico y concurrente entre los niveles de la arquitectura RGNW. La información que llega a C0 corresponde a las observaciones sobre el estado del paciente, que

[3] *Principialismo bioético* (Beauchamp y Childress, 2013): marco compuesto por cuatro principios prima facie para la deliberación ética en medicina: Beneficencia (hacer el bien), No maleficencia (no hacer daño), Autonomía (respetar la autodeterminación) y Justicia (equidad en la distribución de beneficios y cargas).

son procesadas por los módulos internos en C1 encargados de analizar datos, predecir posibles evoluciones clínicas y evaluar qué acciones son factibles. Esta información se usa para integrar y deliberar sobre las posibles acciones. El nivel C2 interviene de manera reflexiva, evaluando cada acción prevista frente a los valores internalizados (los principios de la bioética). A partir de esta evaluación, C2 puede inhibir o modificar las conexiones entre los módulos de C1, de modo que solo aquellas acciones coherentes con los valores éticos lleguen a C0 y sean finalmente sugeridas a los clínicos. Este mecanismo permite un control metacognitivo sobre la arquitectura, garantizando que las decisiones técnicas también cumplan criterios éticos antes de ser comunicadas al profesional sanitario.

3.6. Dilemas éticos y la importancia del contexto social

La ingeniería de valores que proponemos no puede eludir la complejidad de los dilemas éticos. Experimentos como el del "tranvía" y sus variantes (como la de empujar a una persona gruesa desde un puente para salvar a cinco) demuestran que la toma de decisiones morales humanas no se rige por una simple utilidad. La *causación* (si mi acción es directa o indirecta) y los vínculos emocionales (elegir entre un desconocido y un hijo) alteran profundamente el juicio.

Un experimento revelador con psicópatas de la cárcel de Lleida ilustra esta diversidad. Se planteó a los participantes si matarían a una persona sana pero molesta para utilizar sus órganos y salvar a cinco pacientes. Mientras que la mayoría de la población se opone, alrededor del 65 % de los psicópatas aceptaría la opción utilitarista, debido a su falta de empatía y remordimiento.

Este caso subraya una conclusión crucial: *los aspectos éticos de la IA no pueden ser decididos unilateralmente por ingenieros en Silicon Valley, Palo Alto o cualquier otro centro tecnológico*. Deben ser el resultado de un consenso social que involucre a los colectivos que van a utilizar y a ser afectados por la tecnología. La ética, como reflexión filosófica sobre las normas, a menudo entra en conflicto con la moral (las normas concretas de una sociedad), tal y como ilustra el conflicto de Antígona entre su obligación familiar (enterrar a su hermano) y la ley de la ciudad (prohibir el entierro a un traidor). La IA debe ser capaz de navegar en estas tensiones, para lo cual es fundamental que sus sistemas de valores sean transparentes, debatibles y adaptables a los contextos culturales y situacionales.

4. EL PAPEL DE LA TEORÍA DE LA MENTE EN LA INTERACCIÓN SOCIAL Y SU IMPLEMENTACIÓN EN LA RGNW

Para que un sistema de IA sea genuinamente *value-aware* y pueda interactuar de forma ética y efectiva en contextos sociales, debe ir más allá del procesamiento de reglas

explícitas de interacción, de protocolos rígidos. Es necesario que desarrolle una capacidad operativa de *Teoría de la mente (ToM)*. En psicología y ciencias cognitivas, la ToM se refiere a la habilidad de atribuir estados mentales (creencias, deseos, intenciones, conocimientos, emociones) a otros, y de comprender que estos pueden diferir de los propios. Esta capacidad es fundamental para predecir, interpretar e influir en el comportamiento ajeno, y es un pilar de la competencia social.

La arquitectura *RGNW* proporciona un marco natural y estructurado para implementar una ToM artificial, esencial para una conciencia de valores socialmente situada. Su implementación se distribuye a lo largo de sus tres niveles:

1. En el Nivel C0 (reactivo). Módulos especializados de "percepción social" procesan señales en tiempo real (análisis de lenguaje natural, tono de voz, expresiones faciales, postura) para generar inferencias básicas sobre estados emocionales o intenciones inmediatas. Por ejemplo, un módulo puede detectar frustración en la sintaxis o el léxico utilizado por un usuario.

2. En el Nivel C1 (deliberativo). Estas señales se integran para construir *modelos de agente.* Aquí, el sistema forma y actualiza representaciones simbólicas o subsimbólicas sobre los estados mentales atribuidos a cada interlocutor (por ejemplo, "el usuario *cree* que no he entendido su instrucción" o "el usuario *desea* terminar la conversación"). Este nivel permite el razonamiento sobre cómo las acciones propias afectarán estos estados mentales.

3. En el Nivel C2 (reflectivo/valores). Este es el componente crucial que diferencia una ToM meramente funcional de una *ToM guiada por valores*. El sistema de valores alojado en C2 actúa como un filtro y un guía para la ToM:

    - *Prioriza* qué estados mentales de otros son moralmente relevantes (por ejemplo, dar mayor peso a detectar incomodidad o confusión que a simple aburrimiento).

    - *Evalúa* las posibles acciones deliberadas en C1 no solo por su eficacia instrumental, sino por su *impacto en el bienestar, la autonomía o la dignidad del otro,* según los valores.

    - *Modifica la topología* para dirigir la atención (flujo de datos) de los módulos C1 hacia los aspectos de la interacción que son axiológicamente significativos.

De este modo, la RGNW logra que la ToM no sea un módulo aislado, sino una *capacidad permeada por valores*. Un robot social con esta arquitectura no solo inferirá que una persona mayor está repitiendo una historia, sino que, a la luz del valor de la *dignidad*, su

nivel C2 guiará la interacción para evitar reacciones que puedan hacer sentir al usuario ignorado o menospreciado. En una negociación automática, el sistema podría utilizar su ToM para identificar cuándo la contraparte actúa por necesidad (estado mental) versus por ambición, ajustando sus propuestas para alinearse con el valor de la *justicia*.

Por lo tanto, la integración de una Teoría de la mente dentro de la arquitectura RGNW es un paso indispensable para trascender de sistemas que *conocen* valores abstractos a sistemas que los *aplican* en dinámicas sociales complejas, interactuando no con agentes genéricos, sino con entidades cuyos estados mentales deben ser comprendidos y respetados.

5. NEGOCIACIÓN DE VALORES Y ESTRATEGIAS BASADAS EN LA RGNW

La arquitectura RGNW no solo permite a un agente actuar conforme a un sistema de valores estático, sino que lo dota de la capacidad fundamental para participar en la negociación dinámica de valores que caracteriza a las comunidades humanas. Este proceso implica dos dimensiones entrelazadas: 1) la observación y modelado de cómo las comunidades definen y armonizan valores, y 2) la participación estratégica en diálogos o interacciones donde los valores de los participantes pueden converger o entrar en conflicto.

5.1. Modelado de la negociación social de valores

Las sociedades humanas no operan con un conjunto de valores fijo e inmutable; estos son continuamente puestos a prueba, reinterpretados y negociados a través del discurso público, las normas legales y las interacciones cotidianas. Un sistema *value-aware* debe poder observar y entender este proceso. La RGNW puede facilitar esto mediante:

- Aprendizaje e inferencia (Nivel C1). El sistema puede analizar grandes corpus de interacciones (debates parlamentarios, hilos de redes sociales, resoluciones judiciales) para identificar cómo ciertos valores (por ejemplo, "privacidad" vs. "seguridad") son invocados, ponderados y ajustados en distintos contextos. Podemos desarrollar módulos que detecten patrones de *trade-offs* y compromisos.

- Representación de consensos y conflictos (Nivel C2). El nivel reflectivo puede mantener representaciones de los "espacios de valores" de una comunidad, no como puntos fijos, sino como, por ejemplo, distribuciones de probabilidad. Puede modelar que, en una comunidad dada, el valor "libertad individual" tiene un peso alto pero está en constante conflicto con el valor "responsabilidad colectiva".

5.2. Estrategias de negociación basadas en valores

Cuando un agente con RGNW se encuentra en un escenario de negociación (cooperativa o competitiva) con humanos u otros agentes, su arquitectura le permite desarrollar estrategias sofisticadas que van más allá de la mera maximización de utilidad económica clásica. En lugar de optimizar exclusivamente un criterio utilitarista, el agente considera un conjunto de valores éticos y normativos que pueden guiar sus decisiones incluso cuando impliquen aceptar pequeñas pérdidas económicas.

La clave reside en utilizar un modelo de la Teoría de la mente (ToM) para inferir el sistema de valores del interlocutor. Esta información permite al agente alinear estratégicamente sus acciones, es decir, elegir comportamientos que no solo buscan un beneficio material, sino que también respetan o satisfacen los valores del otro agente o humano. La alineación estratégica implica seleccionar acciones que maximicen simultáneamente la utilidad clásica y la satisfacción de los valores éticos, lo que puede incluir compromisos de carácter deontológico, promoción de la cooperación y respeto a normas sociales, asegurando que la negociación sea efectiva y socialmente aceptable.

1. Detección de alineamiento y fricción axiológica. El sistema, a través de sus módulos en C0 y C1, puede analizar las propuestas y reacciones del interlocutor. A nivel C2 se puede evaluar continuamente: ¿los valores que motivan mi propuesta (por ejemplo, "eficiencia") están alineados con los valores que parecen motivar al otro (por ejemplo, "estabilidad" o "equidad")? Y de esa manera identificar tanto *valores de acuerdo* como *valores en conflicto.*

2. Generación de opciones y argumentación guiada por valores. Basándose en esta evaluación, el nivel C1/C2 puede generar no solo opciones que maximizan un objetivo, sino opciones que pueden ser *renmarcadas* o *justificadas* en términos de los valores del interlocutor.

    - Si hay *valores compartidos*, la estrategia será *amplificarlos.* "Ambas propuestas buscan la justicia, la mía lo hace priorizando la equidad a largo plazo".

    - Si hay *conflicto de valores*, la estrategia puede ser:

        - *Crear puentes.* Encontrar un valor de orden superior compartido ("Ambos valoramos el bienestar de la comunidad, exploremos cómo balancear libertad y seguridad para lograrlo").

        - *Ofrecer compensación axiológica.* Reconocer explícitamente el sacrificio de un valor y compensarlo en otro ámbito ("Esta opción prioriza la eficiencia que usted necesita, y propongo una cláusula de revisión que proteja la transparencia que yo valoro").

3. Adaptación dinámica de la topología (Función clave del C2). Durante la negociación, el nivel C2 puede *reconfigurar la prioridad de los módulos deliberativos* en C1. Por ejemplo:

   - Si detecta que el interlocutor responde positivamente a argumentos basados en "lealtad", puede potenciar los módulos que generan propuestas que enfaticen compromisos a largo plazo y relaciones estables.

   - Si percibe que la negociación está estancada por un conflicto de valores irreconciliable, puede activar módulos de "creatividad" para buscar opciones innovadoras que satisfagan los intereses subyacentes de ambos sin violar sus valores esenciales.

5.3. Ejemplo ilustrativo: negociación automatizada de un contrato

Consideremos un agente RGNW que negocia un acuerdo de colaboración entre una universidad y una empresa.

Valores del agente (Universidad). Acceso abierto al conocimiento, independencia académica.

Valores inferidos de la contraparte (Empresa). Propiedad intelectual, rentabilidad.

Proceso RGNW:

1. *C1 (Teoría de la mente–ToM).* El agente analiza la conducta de la contraparte y otros datos contextuales (por ejemplo, insistencia en cláusulas de confidencialidad) para inferir que estos comportamientos reflejan el valor "propiedad intelectual".

2. *C2 (reflexión).* Evalúa los conflictos entre los valores propios y los de la contraparte ("acceso abierto" vs. "propiedad intelectual") y busca un valor compartido superior: "innovación".

3. *C2 (modulación).* Ajusta los módulos de generación de propuestas en C1 para producir opciones que maximicen un *criterio híbrido*:

   - Satisfacción de los valores propios del agente (universidad).

   - Reconocimiento parcial de los valores de la contraparte (empresa), aunque esto implique cierta pérdida económica.

4. *Salida:* El agente propone una "ventana de embargo": los resultados son confidenciales durante 12 meses (respetando la propiedad intelectual) y luego se hacen de acceso abierto (respetando la misión universitaria). La propuesta argumenta que este modelo acelera la innovación, beneficiando a ambas partes.

La propuesta del agente no busca maximizar exclusivamente la utilidad económica propia, sino encontrar un punto intermedio que equilibre intereses económicos y valores éticos de ambas partes. Los datos que utiliza incluyen comportamientos observables de la contraparte, reglas explícitas de valores, y contexto del escenario de negociación. El agente optimiza una función híbrida que combina beneficio económico y alineación con valores, lo que le permite aceptar pequeñas pérdidas económicas para respetar o fomentar los valores del interlocutor. De esta manera, la RGNW actúa como un participante activo en el ecosistema moral de la negociación, capaz de comprender, influir y adaptarse a los procesos sociales mediante los cuales los valores se negocian y redefinen colectivamente.

6. DISCUSIÓN

La propuesta arquitectónica del *Reflective Global Neuronal Workspace (RGNW*) se enmarca, en última instancia, en el largo esfuerzo por operacionalizar principios abstractos en sistemas computacionales. Su núcleo innovador no radica en descartar los mecanismos normativos clásicos, sino en dotarlos de una capacidad de reflexión guiada por valores. Al definir un valor como una preferencia sobre estados del mundo y generar a partir de él normas alineadas que prescriben comportamientos, la RGNW implementa *de facto* un sistema normativo dinámico y autoadaptativo. Este enfoque establece un diálogo directo con los fundamentos de las instituciones electrónicas, un marco pionero en el que las normas explícitas (protocolos, reglas de compromiso, contratos) se utilizan para estructurar, gobernar y hacer predecibles las interacciones entre agentes autónomos dentro de un espacio de acción compartido. La arquitectura RGNW puede interpretarse como la internalización y "reflexivización" de este principio institucional: su nivel C2 actúa como una suerte de institución electrónica personal y autogobernada, donde los valores constituyen los objetivos de diseño de alto nivel y el conjunto de normas derivadas (cuya aplicación se gestiona mediante la topología dinámica entre C1 y C0) opera como el mecanismo de gobierno interno. Este vínculo conecta directamente con líneas de investigación previas sobre la formalización lógica y la gestión computacional de normativas en sistemas multiagente, como las desarrolladas en el contexto de las instituciones electrónicas (D'Inverno *et al,*, 2012). La contribución distintiva de la RGNW es la adición de la capa reflectiva (C2), que permite al sistema no solo ejecutar normas, sino también evaluar continuamente su coherencia con el sistema de valores, adaptarlas contextualmente y explicitar la justificación axiológica de sus acciones. Así, se transita desde un gobierno normativo estático hacia una administración consciente y justificada de la coherencia axiológica, un paso crucial para construir agentes autóno-

mos que operen con responsabilidad en entornos sociales complejos, es decir con una multitud de agentes autónomos con objetivos e intereses a menudo contrapuestos.

Este capítulo ha defendido una tesis central: la responsabilidad moral de la inteligencia artificial se construye mediante ingeniería, no mediante la especulación sobre una conciencia inalcanzable. Al desacoplar conceptualmente la *conciencia fenomenológica* (*consciousness*) de la *capacidad operativa de darse cuenta* (*awareness*), hemos reenfocado el problema desde la metafísica hacia la computación. La propuesta de sistemas *value-aware*, materializada en la arquitectura RGNW, constituye una respuesta concreta a este desafío.

La RGNW demuestra que es factible diseñar agentes que internalicen valores como preferencias sobre estados del mundo, los traduzcan en normas operativas y monitoricen su propia coherencia axiológica a través de una capa reflectiva. Más allá de la adhesión estática a reglas, esta arquitectura permite a los agentes navegar la complejidad social: implementar una Teoría de la mente para interpretar a los demás, participar en la negociación dinámica de valores y reconfigurar su propio funcionamiento para alinearse con principios éticos en contextos cambiantes. En este sentido, la RGNW opera como una institución electrónica internalizada, un sistema normativo autogobernado y guiado por valores que evoca el trabajo fundacional en sistemas multiagente.

Las aplicaciones descritas —desde el mapeo de la ecología moral en redes sociales hasta la asistencia robótica sensible al contexto— no son meros prototipos, sino pruebas de concepto que validan el marco teórico. Evidencian que la "conciencia de valores" es un requisito técnico imprescindible para el despliegue responsable de la IA en dominios sociales.

El camino futuro no está exento de retos. La explicabilidad profunda de las decisiones guiadas por valores, la gestión de conflictos axiológicos irresolubles y el diseño de mecanismos para la evolución y aprendizaje colaborativo de valores con las comunidades humanas son fronteras de investigación abiertas. Sin embargo, el debate ya no debe centrarse en si las máquinas pueden ser sujetos morales, sino en cómo podemos dotarlas, de manera robusta y verificable, de la capacidad para saber lo que importa. La tarea urgente para ingenieros, diseñadores y científicos sociales es refinar estas herramientas, asegurando que la IA del futuro no solo sea inteligente, sino también centrada en valores, reflejando y sirviendo al pluralismo moral de la humanidad.

Referencias

Baars, B. J., & Alonzi, A. (2018). The global workspace theory. En R. J. Gennaro (Ed.), *The Routledge handbook of consciousness* (pp. 233-245). Routledge.

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7ª ed.). Oxford University Press.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1-3), 139–159.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences, 95*(24), 14529–14534.

D'Inverno, M., Luck, M., Noriega, P., Rodriguez-Aguilar, J. A., & Sierra, C. (2012). Communicating open systems. *Artificial Intelligence,* 186, 38–94.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98–116.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Ladak, A. (2024). What would qualify an artificial intelligence for moral standing? *AI and Ethics,* 4, 213–228.

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture, 2*(1).