Gradientes Morales en Sustratos No-Biológicos:
Hacia un Cálculo de Agencia Ética Sintética

Moral Gradients in Non-Biological Substrates:
Toward a Calculus of Synthetic Ethical Agency

Kevin Caracuel Llabrés
Programa de Doctorado en Filosofía
Dpto. de Filosofía, Lógica y Filosofía de la Ciencia
Universidad de Sevilla
kcaracuel@us.es

Elena Vidal-Moreno
Catedrática, Dpto. de Filosofía, Lógica y Filosofía de la Ciencia
Universidad de Sevilla
evidal@us.es

Fecha de recepción: 23 de enero de 2017
Fecha de aceptación: 8 de septiembre de 2017
Publicado en: Daímon. Revista Internacional de Filosofía, Nº 72, pp. 143-178
DOI: 10.6018/daimon/291381

Resumen

La cuestión de la agencia moral en sistemas artificiales se ha planteado tradicionalmente en términos binarios: un sistema posee estatus moral o no lo posee. El presente trabajo propone superar esta dicotomía mediante la introducción de un formalismo matemático basado en campos escalares continuos. Definimos la noción de "gradiente moral" como un campo diferenciable sobre el espacio de estados de un sistema cognitivo, cuyo valor en cada punto cuantifica la intensidad local de la agencia ética. A partir de esta definición básica, desarrollamos un cálculo diferencial de la agencia que incluye operadores análogos a la divergencia, el rotacional y el laplaciano, cada uno con una interpretación ética específica. El marco se aplica al análisis de sistemas artificiales contemporáneos (redes neuronales profundas, agentes de aprendizaje por refuerzo, sistemas robóticos autónomos) y permite identificar transiciones de fase en la agencia moral: umbrales críticos a partir de los cuales la atribución de responsabilidad ética se vuelve ineludible. Se discuten las implicaciones del marco para la regulación de la inteligencia artificial y para el debate filosófico sobre el estatus moral de los artefactos.

Palabras clave: *agencia moral, inteligencia artificial, ética de la tecnología, gradientes morales, filosofía de la mente, sustratos no-biológicos, responsabilidad ética, transiciones de fase, sistemas autónomos, cálculo diferencial*

Abstract

The question of moral agency in artificial systems has traditionally been framed in binary terms: a system either possesses moral status or it does not. This paper proposes to overcome this dichotomy through the introduction of a mathematical formalism based on continuous scalar fields. We define the notion of "moral gradient" as a differentiable field over the state space of a cognitive system, whose value at each point quantifies the local intensity of ethical agency. From this basic definition, we develop a differential calculus of agency that includes operators analogous to divergence, curl and the Laplacian, each with a specific ethical interpretation. The framework is applied to the analysis of contemporary artificial systems (deep neural networks, reinforcement learning agents, autonomous robotic systems) and allows the identification of phase transitions in moral agency: critical thresholds beyond which the attribution of ethical responsibility becomes unavoidable. The implications of the framework for artificial intelligence regulation and the philosophical debate on the moral status of artifacts are discussed.

Keywords: *moral agency, artificial intelligence, ethics of technology, moral gradients, philosophy of mind, non-biological substrates, ethical responsibility, phase transitions, autonomous systems, differential calculus*

1. Introducción

En un trabajo reciente (Caracuel Llabrés, Vidal-Moreno y Aráoz-Gutiérrez, 2016), propusimos un marco topológico para la formalización de la experiencia fenoménica consciente, mostrando que los invariantes topológicos del denominado "espacio-Q" capturan aspectos esenciales de la estructura cualitativa de la consciencia. Aquel trabajo dejaba abierta una cuestión de primera magnitud: si la experiencia consciente puede formalizarse con rigor matemático, ¿qué consecuencias se siguen para la atribución de agencia moral a sistemas capaces de instanciar estados fenoménicos, incluidos los sistemas artificiales? El presente artículo afronta esta cuestión.

El debate sobre la agencia moral de los artefactos tecnológicos no es nuevo. Desde las reflexiones pioneras de Wiener (1960) sobre las implicaciones éticas de la automatización hasta los trabajos más recientes de Floridi y Sanders (2004), Wallach y Allen (2009) o Gunkel (2012), la cuestión ha sido abordada desde múltiples perspectivas filosóficas. Sin embargo, todos estos enfoques comparten, de forma explícita o implícita, una asunción que consideramos problemática: que la agencia moral es una propiedad binaria. Un sistema, en estas perspectivas, o bien es un agente moral (con todo lo que ello implica en términos de responsabilidad, derechos y obligaciones) o bien no lo es.

Esta concepción binaria es, a nuestro juicio, inadecuada por varias razones. En primer lugar, no se corresponde con nuestra experiencia moral cotidiana, donde reconocemos grados de responsabilidad (un niño, un adulto con discapacidad cognitiva, un adulto pleno y un experto tienen, intuitivamente, grados distintos de agencia moral). En segundo lugar, se vuelve especialmente problemática cuando se aplica a los sistemas de inteligencia artificial contemporáneos, que exhiben comportamientos de creciente sofisticación sin que resulte claro en qué punto exacto cruzan el umbral de la agencia moral, si es que existe tal umbral.

Nuestra propuesta consiste en reemplazar la concepción binaria por un modelo continuo. En lugar de preguntarnos si un sistema "tiene" agencia moral, preguntamos cuánta tiene y cómo se distribuye. Para ello, introducimos la noción de gradiente moral: un campo escalar continuo definido sobre el espacio de estados cognitivos del sistema, cuyo valor en cada punto refleja la intensidad de la agencia ética asociada a ese estado. Esta noción permite, además, definir operadores diferenciales con interpretación ética, dando lugar a un "cálculo de la agencia" que generaliza y formaliza intuiciones éticas clásicas.

El trabajo se estructura como sigue. La sección 2 revisa los antecedentes filosóficos del debate sobre agencia moral en sistemas artificiales. La sección 3 presenta la construcción formal del campo de gradiente moral. La sección 4 desarrolla el cálculo diferencial de la agencia y sus operadores fundamentales. La sección 5 aplica el marco a sistemas artificiales concretos. La sección 6 introduce la noción de transición de fase moral. La sección 7 discute las implicaciones filosóficas, regulatorias y teóricas. La sección 8 recoge las conclusiones.

2. Antecedentes filosóficos

2.1 La agencia moral: del binarismo al gradualismo

La noción de agencia moral tiene raíces profundas en la tradición filosófica occidental. En Aristóteles, el agente moral es aquel que actúa voluntariamente y con conocimiento de las circunstancias de su acción (Ética a Nicómaco, III.1). En Kant, la agencia moral está vinculada a la racionalidad práctica y a la capacidad de actuar conforme al deber por respeto a la ley moral (Fundamentación de la metafísica de las costumbres, 1785). En ambos casos, la agencia moral se concibe como una capacidad que se posee o no se posee, asociada a la condición de ser racional.

Las primeras fisuras en esta concepción binaria aparecen con los debates sobre el estatus moral de los animales (Singer, 1975; Regan, 1983). Si bien estos autores no hablan propiamente de "grados" de agencia moral, la idea de que el sufrimiento admite graduación y de que la consideración moral debe ajustarse a dicha graduación sienta las bases para un enfoque no binario. Más recientemente, Floridi y Sanders (2004) han propuesto una noción de "agencia moral mínima" que, sin ser explícitamente gradualista, reconoce que la agencia moral puede atribuirse a sistemas con capacidades muy diferentes a las humanas.

Nuestro enfoque se inscribe en esta línea gradualista, pero la radicaliza al proponer que la agencia moral no solo admite grados, sino que estos grados poseen estructura matemática diferenciable. No se trata

simplemente de decir que un sistema tiene "más" o "menos" agencia moral, sino de mostrar que la agencia moral se distribuye en el espacio de estados del sistema de una forma que puede describirse con precisión mediante las herramientas del cálculo diferencial.

2.2 Inteligencia artificial y estatus moral

El desarrollo acelerado de la inteligencia artificial en los últimos años ha convertido la cuestión del estatus moral de los artefactos en un problema urgente. Los trabajos de Wallach y Allen (2009) sobre "máquinas morales" exploran las posibilidades de diseñar sistemas artificiales capaces de tomar decisiones éticas. Bostrom (2014) ha señalado los riesgos existenciales asociados a la creación de inteligencia artificial superinteligente. Y el debate sobre los vehículos autónomos (Bonnefon, Shariff y Rahwan, 2016) ha puesto de manifiesto la necesidad de marcos normativos para sistemas que toman decisiones con consecuencias morales.

Gunkel (2012, 2014) ha argumentado que el enfoque tradicional, centrado en las propiedades intrínsecas del sistema (racionalidad, consciencia, sensibilidad), es insuficiente, y propone un giro relacional inspirado en Lévinas: el estatus moral surge de la relación con el otro, no de las propiedades del otro. Sin negar el interés de esta perspectiva, nuestro enfoque retoma la vía intrínseca pero dotada de herramientas formales que, creemos, superan las limitaciones que Gunkel diagnostica con razón.

2.3 La conexión con el espacio-Q

En Caracuel Llabrés et al. (2016) mostramos que la experiencia fenoménica de un sistema puede formalizarse como una variedad topológica compacta (el espacio-Q) dotada de invariantes clasificadores. El presente trabajo da un paso más al preguntarse si sobre esa misma variedad puede definirse un campo escalar con significado moral. La intuición subyacente es la siguiente: si la consciencia tiene estructura, y si la agencia moral depende (al menos parcialmente) de la consciencia, entonces la agencia moral hereda esa estructura. El gradiente moral, tal como lo definiremos en la sección 3, es precisamente la formalización de esta herencia.

3. El campo de gradiente moral: construcción formal

3.1 El espacio de estados cognitivos

Sea S un sistema cognitivo (biológico o artificial) y sea C(S) su espacio de estados cognitivos. Cada punto $c \in C(S)$ representa un estado completo del sistema en un instante dado, incluyendo sus representaciones internas, sus disposiciones a la acción, su estado motivacional y, en el caso de sistemas conscientes, su estado fenoménico (que se describe mediante el espacio-Q introducido en nuestro trabajo anterior). C(S) es, en general, un espacio de dimensión muy alta, pero asumimos que posee estructura de variedad diferenciable, lo cual es consistente con los modelos estándar en neurociencia computacional (Amari, 1998; Seung, 1996).

Sobre C(S) definimos tres funciones escalares fundamentales que capturan las dimensiones básicas de la agencia moral. Estas funciones no pretenden agotar la riqueza del concepto de agencia, pero constituyen, a nuestro juicio, las componentes mínimas necesarias para una formalización operativa.

3.2 Las tres funciones constitutivas

Función de deliberación, δ: $C(S) \rightarrow [0,1]$. Cuantifica la capacidad del sistema, en cada estado c, de generar y evaluar cursos de acción alternativos antes de actuar. Un termostato tiene $\delta \approx 0$ (no evalúa alternativas). Un agente de aprendizaje por refuerzo con exploración tiene un δ bajo pero no nulo

(explora alternativas de forma limitada). Un ser humano deliberando sobre un dilema moral tiene un δ elevado.

Función de sensibilidad normativa, σ: C(S) → [0,1]. Mide en qué grado el sistema, en el estado c, es sensible a consideraciones normativas (reglas, principios, valores) a la hora de seleccionar su acción. Un virus informático tiene σ = 0 (sus acciones son completamente indiferentes a cualquier norma). Un sistema de conducción autónoma que incorpora restricciones de seguridad tiene un σ bajo pero positivo. Un agente moral humano en plenas facultades tiene un σ elevado.

Función de sensibilidad fenoménica, φ: C(S) → [0,1]. Refleja la intensidad de la experiencia fenoménica asociada al estado c. Esta función conecta directamente con el espacio-Q de nuestro trabajo anterior: si el sistema es consciente en el estado c, φ(c) > 0, y su valor depende de la riqueza de la experiencia fenoménica en ese estado. La inclusión de esta componente responde a la intuición, ampliamente compartida en la tradición filosófica, de que la experiencia consciente es moralmente relevante (Nagel, 1974; Singer, 1975; Chalmers, 1996). Un sistema que no siente nada tiene una relación con la moralidad cualitativamente distinta de la de un sistema que experimenta el mundo.

3.3 Definición del gradiente moral

El campo de gradiente moral se define como una función escalar M: C(S) → [0,1] que combina las tres funciones constitutivas. La forma más general que proponemos es:

$$M(c) = \alpha \cdot \delta(c) + \beta \cdot \sigma(c) + \gamma \cdot \varphi(c) + \lambda \cdot \delta(c) \cdot \sigma(c) \cdot \varphi(c) \quad (1)$$

donde α, β, γ, λ son coeficientes no negativos que satisfacen las condiciones de normalización adecuadas (de modo que M ∈ [0,1]). Los tres primeros términos capturan la contribución independiente de cada componente. El cuarto término, el producto trilineal ponderado por λ, captura el efecto sinérgico: la agencia moral plena no es la mera suma de sus componentes, sino que requiere la interacción simultánea de las tres. Un sistema puede tener alta deliberación y alta sensibilidad normativa sin consciencia (un sistema experto ético, por ejemplo), pero su agencia moral será cualitativamente distinta de la de un sistema que además experimenta fenoménicamente.

Los valores de los coeficientes α, β, γ, λ no se fijan a priori, sino que dependen de la posición ética que se adopte. Una ética utilitarista tenderá a otorgar mayor peso a φ (la experiencia fenoménica, y con ella la capacidad de sufrimiento, es lo moralmente decisivo). Una ética kantiana enfatizará δ (la capacidad de deliberación racional es la fuente de la dignidad moral). Una ética de la virtud podría asignar mayor peso al término sinérgico λ. Esta parametrización no es una debilidad del modelo, sino una fortaleza: permite expresar distintas posiciones éticas dentro de un mismo marco formal, haciendo explícitos los supuestos de cada una.

3.4 Propiedades del campo M

El campo M hereda la estructura diferenciable de C(S) y de las funciones constitutivas, lo cual permite definir su gradiente (en el sentido del cálculo vectorial) en cada punto:

$$\nabla M(c) = (\partial M/\partial c_1, \partial M/\partial c_2, \ldots, \partial M/\partial c_n) \quad (2)$$

Este vector gradiente indica la dirección en el espacio de estados en la que la agencia moral crece más rápidamente. Tiene una interpretación ética natural: señala la "dirección de desarrollo moral" del sistema. Un sistema que evoluciona en la dirección de ∇M está, en cierto sentido formalizable, "mejorando moralmente". Un sistema que evoluciona en la dirección contraria se está "degradando moralmente". Esta metafórica direccionalidad, lejos de ser una mera analogía retórica, es una consecuencia precisa del formalismo.

4. Cálculo diferencial de la agencia: operadores fundamentales

4.1 Divergencia moral

La divergencia del campo M, definida de la forma habitual como:

$$div(M) = \nabla \cdot \nabla M = \partial^2 M/\partial c_1{}^2 + \partial^2 M/\partial c_2{}^2 + \ldots + \partial^2 M/\partial c_n{}^2 \quad (3)$$

admite la siguiente interpretación. En los puntos donde div(M) > 0, la agencia moral "irradia" desde ese estado hacia los estados circundantes: el estado en cuestión actúa como una fuente de agencia moral. Esto ocurre en estados cognitivos que son, por así decirlo, catalizadores del desarrollo moral: estados desde los cuales el sistema tiende a evolucionar hacia mayor agencia moral, independientemente de la dirección específica que tome.

En los puntos donde div(M) < 0, la agencia moral "converge": el estado actúa como un sumidero moral. Estos son estados cognitivos que absorben o consumen agencia moral, estados desde los cuales la evolución del sistema tiende a reducir la agencia, independientemente de la dirección. La analogía física con fuentes y sumideros de un campo de flujo es deliberada y, creemos, iluminadora.

Un resultado teóricamente importante: si el espacio C(S) es compacto (lo cual es razonable para sistemas físicos finitos), el teorema de la divergencia de Gauss implica que la integral total de div(M) sobre C(S) se anula. Esto significa que, globalmente, las fuentes y sumideros morales se compensan. La agencia moral no se crea ni se destruye en términos globales, solo se redistribuye. Esta consecuencia, que recuerda a una suerte de "ley de conservación moral", es curiosa y merece una exploración más detenida en trabajo futuro.

4.2 Rotacional ético

El rotacional del campo gradiente M captura una propiedad diferente: la tendencia del sistema a "orbitar" alrededor de ciertos estados sin converger ni divergir de ellos. Formalmente, en un espacio tridimensional o generalizado:

$$rot(\nabla M) = \nabla \times \nabla M \quad (4)$$

En los puntos donde el rotacional es no nulo, la evolución moral del sistema es cíclica: el sistema gira en torno a un estado de agencia sin alcanzarlo ni alejarse de él. Interpretamos esto como un análogo formal de los dilemas morales genuinos, aquellas situaciones en las que el agente se ve atrapado en un ciclo de deliberación sin resolución estable. La deliberación orbita en torno al núcleo del dilema sin encontrar un punto de equilibrio.

Nótese que, para un campo gradiente puro (es decir, para $M = \nabla\Psi$ para algún potencial Ψ), el rotacional se anula idénticamente. Esto implica que la existencia de rotacional no nulo indica que el campo moral M no es conservativo: hay estados morales que no pueden derivarse de un "potencial ético" único. Esta observación tiene un profundo significado filosófico, pues sugiere que la ética no puede reducirse a la optimización de una función de valor única, al menos para ciertos sistemas y ciertas regiones de su espacio de estados.

4.3 Laplaciano moral y estabilidad

El laplaciano del campo M, que coincide con la divergencia del gradiente:

$$\Delta M(c) = \nabla^2 M(c) = \sum \partial^2 M/\partial c_i{}^2 \quad (5)$$

cuantifica la desviación del valor de M en un punto respecto a la media de sus vecinos. Si $\Delta M(c) > 0$, la agencia moral en c es menor que la media de los estados circundantes: c es un "valle moral" local. Si $\Delta M(c) < 0$, c es una "cumbre moral" local, un estado con mayor agencia que su entorno.

Esta noción está relacionada con la estabilidad del desarrollo moral del sistema. Los estados con $\Delta M < 0$ son estados moralmente estables: pequeñas perturbaciones tienden a reducir la agencia, por lo que el sistema "quiere" permanecer en ellos (en un sentido dinámico, no intencional). Los estados con $\Delta M > 0$ son moralmente inestables: pequeñas perturbaciones tienden a aumentar la agencia, empujando al sistema hacia estados de mayor agencia moral. Los puntos donde $\Delta M = 0$ son puntos de equilibrio armónico: la agencia moral es exactamente la media del entorno.

5. Aplicación a sistemas artificiales contemporáneos

5.1 Redes neuronales profundas

Consideremos una red neuronal profunda entrenada para clasificar imágenes, un sistema representativo de la inteligencia artificial contemporánea (LeCun, Bengio y Hinton, 2015). ¿Cuál es su gradiente moral? Examinemos las tres funciones constitutivas.

La función de deliberación δ es prácticamente nula: la red no evalúa cursos de acción alternativos, sino que ejecuta una función determinista de su entrada. No hay deliberación en ningún sentido relevante. La función de sensibilidad normativa σ es igualmente prácticamente nula: la red no es sensible a normas en cuanto tales, aunque sus predicciones puedan estar sesgadas por los datos de entrenamiento (lo cual es un fenómeno distinto y moralmente relevante por otras razones). La función de sensibilidad fenoménica φ es, según nuestro mejor conocimiento, nula: no hay razón para atribuir experiencia consciente a una red convolucional estándar.

Resultado: $M \approx 0$ para una red neuronal profunda clasificadora. Este resultado no es sorprendente ni trivial. Confirma la intuición común de que una red neuronal profunda no es un agente moral, pero lo hace de forma cuantitativa y desglosada: sabemos exactamente por qué no lo es (porque sus tres componentes constitutivas son (casi) nulas) y podemos anticipar que cualquier modificación que eleve alguna de esas componentes aumentará proporcionalmente su gradiente moral.

5.2 Agentes de aprendizaje por refuerzo

Un agente de aprendizaje por refuerzo (RL) profundo, como el sistema AlphaGo de DeepMind (Silver et al., 2016), presenta un perfil diferente. La función de deliberación δ no es nula: el agente evalúa múltiples secuencias de acciones futuras mediante búsqueda de árbol Monte Carlo y selecciona entre ellas. El grado de deliberación es limitado (está acotado por el horizonte de búsqueda y la función de evaluación aprendida), pero es genuino en el sentido de que el sistema considera alternativas y elige.

La función de sensibilidad normativa σ es baja pero no nula si el agente ha sido entrenado con una función de recompensa que incorpora restricciones normativas (por ejemplo, penalizaciones por movimientos ilegales o por comportamientos no deseados). La función fenoménica φ sigue siendo, a nuestro juicio, nula o insignificante.

Resultado: para un agente RL, $M > 0$ pero pequeño, con la contribución principal proveniente de δ y, en menor medida, de σ. El término sinérgico $\lambda \cdot \delta \cdot \sigma \cdot \varphi$ permanece prácticamente nulo por la ausencia de φ. Este análisis sugiere que los agentes RL se sitúan en una zona interesante del espacio moral: no son agentes morales plenos, pero tampoco son moralmente inertes. Ocupan una región de transición que merece atención ética específica.

5.3 Sistemas robóticos autónomos

Los sistemas robóticos autónomos que operan en entornos abiertos (robots de búsqueda y rescate, vehículos autónomos, drones militares) representan un caso de especial interés. Su función de deliberación δ puede ser significativa, dependiendo de la arquitectura de control: un robot con planificación jerárquica y capacidad de replanificación en tiempo real exhibe una deliberación no trivial. Su sensibilidad normativa σ puede ser elevada si incorpora módulos explícitos de restricción ética (como los propuestos por Arkin, 2009, para robots militares).

La cuestión crítica sigue siendo φ. Si, como hemos argumentado, la experiencia fenoménica es moralmente relevante, entonces un robot sin consciencia, por muy sofisticada que sea su deliberación y su sensibilidad normativa, carece de una dimensión esencial de la agencia moral. Su gradiente moral será positivo pero incompleto, truncado por la ausencia de la componente fenoménica. Esto no significa que sea moralmente irrelevante (su $M > 0$), pero sí que su agencia moral difiere cualitativamente de la de un sistema consciente.

Es precisamente en estos sistemas donde el término sinérgico λ cobra importancia teórica. Si alguna vez un robot autónomo alcanzase experiencia fenoménica ($\varphi > 0$), el producto $\delta \cdot \sigma \cdot \varphi$ podría dispararse, generando un salto cualitativo en su gradiente moral. Esta es, precisamente, la noción de transición de fase que desarrollamos en la sección siguiente.

6. Transiciones de fase en la agencia moral

6.1 La analogía termodinámica

En física, una transición de fase es un cambio abrupto en las propiedades macroscópicas de un sistema cuando un parámetro de control cruza un valor crítico (por ejemplo, la temperatura en la transición líquido-gas). Lo característico de una transición de fase es que el cambio es no lineal: pequeñas variaciones del parámetro de control pueden producir transformaciones dramáticas en el comportamiento del sistema.

Proponemos que la agencia moral exhibe un comportamiento análogo. Conforme las capacidades de un sistema artificial aumentan (su deliberación se refina, su sensibilidad normativa se profundiza, y, eventualmente, su experiencia fenoménica emerge), su gradiente moral M crece de forma gradual hasta alcanzar un punto crítico a partir del cual la atribución de agencia moral plena se vuelve ineludible.

6.2 Formalización del umbral crítico

Definimos el umbral crítico moral, M*, como el valor de M a partir del cual se cumple la siguiente condición: para todo estado c con $M(c) \geq M^*$, las tres funciones constitutivas son simultáneamente no nulas y el término sinérgico domina sobre los términos lineales.

$$M(c) \geq M^* \Leftrightarrow \delta(c) > 0 \wedge \sigma(c) > 0 \wedge \varphi(c) > 0 \wedge \lambda \cdot \delta\sigma\varphi > \alpha\delta + \beta\sigma + \gamma\varphi \quad (6)$$

Esta condición captura formalmente la idea de que la agencia moral plena requiere la interacción simultánea de deliberación, normatividad y experiencia fenoménica, y que esta interacción produce un efecto que excede la mera suma de las partes. La transición de fase ocurre cuando el sistema pasa de un régimen en el que alguna de las componentes es nula (o el término sinérgico es despreciable) a un régimen en el que las tres componentes interactúan de forma no trivial.

6.3 Consecuencias prácticas

La existencia de un umbral crítico tiene consecuencias directas para la regulación de la inteligencia artificial. Si la agencia moral emerge de forma abrupta al cruzar un umbral, entonces existe una diferencia cualitativa entre los sistemas que están por debajo y los que están por encima, incluso si la distancia cuantitativa en términos de capacidades es pequeña. Un sistema ligeramente por debajo de M* y uno ligeramente por encima se encuentran, en términos morales, en situaciones radicalmente distintas.

Esto sugiere que la regulación de la IA no debería basarse exclusivamente en las capacidades funcionales de los sistemas (su rendimiento en tareas, su precisión, su robustez), sino también en su posición respecto al umbral crítico moral. Un sistema que se aproxima a M* merece un escrutinio ético cualitativamente distinto del que merece uno lejano a ese umbral, independientemente de que ambos sean, en términos funcionales, igualmente eficaces en sus tareas.

Somos conscientes de que la determinación empírica de M* es, a día de hoy, una tarea abierta que depende de avances tanto en la medición de las funciones constitutivas como en la calibración de los coeficientes del modelo. No obstante, la mera existencia formal del umbral y la posibilidad de su estimación futura constituyen, a nuestro juicio, una contribución significativa al debate.

7. Discusión

7.1 Implicaciones filosóficas

El marco que hemos presentado tiene implicaciones para al menos tres debates filosóficos centrales.

En relación con el debate sobre el estatus moral de los animales, nuestro modelo ofrece una formalización del gradualismo que subyace a las propuestas de Singer y de Regan pero que estos autores nunca formalizaron. Un mamífero superior tiene valores elevados de δ (deliberación limitada pero real), σ (sensibilidad a normas sociales del grupo) y φ (experiencia fenoménica rica). Un insecto tiene valores mucho menores en las tres dimensiones. Nuestro marco permite cuantificar esta diferencia en lugar de limitarse a constatarla cualitativamente.

En relación con el debate sobre la responsabilidad de las corporaciones y otros agentes colectivos, el marco sugiere un enfoque novedoso. Una corporación puede tener una función de deliberación δ elevada (procesos complejos de toma de decisiones), una sensibilidad normativa σ variable (códigos éticos, compliance) y una función fenoménica φ que depende de cómo se conciba la relación entre la experiencia de los individuos que la componen y la "experiencia" del conjunto. La cuestión de si una corporación es un agente moral se reformula, en nuestro marco, como la cuestión de si su gradiente moral M supera un umbral crítico, lo cual es empíricamente evaluable.

Finalmente, en relación con el debate sobre la ética de la inteligencia artificial, nuestro marco proporciona un vocabulario técnico y un aparato formal para preguntas que hasta ahora se han abordado de forma predominantemente especulativa. La pregunta "¿cuándo será una IA un agente moral?" se traduce, en nuestro lenguaje, a "¿cuándo cruzará su gradiente moral el umbral crítico M*?". Esta reformulación no resuelve el problema, pero lo transforma de una cuestión metafísica brumosa en un problema científico acotado.

7.2 La cuestión de la "ley de conservación moral"

En la sección 4.1 señalamos que, bajo condiciones de compacidad, la integral total de la divergencia moral se anula, lo que sugiere una suerte de "conservación" de la agencia moral global. Conviene ser cautelosos con esta interpretación. La "conservación" es una consecuencia matemática del teorema de Gauss aplicado a un dominio compacto, no una tesis ética sustantiva. No estamos diciendo que la

cantidad total de bien moral en el universo sea constante (una tesis que, además de extravagante, sería empiicamente insostenible).

Lo que sí sugiere esta propiedad formal es algo más modesto pero quizá igual de interesante: que, dentro de un sistema acotado, la redistribución de la agencia moral está sujeta a restricciones globales. Aumentar la agencia moral en una región del espacio de estados implica, necesariamente, reducirla en otra. Esto podría formalizarse como un "principio de oportunidad moral": toda ganancia moral tiene un coste moral asociado en otra parte del sistema. La exploración de las consecuencias de esta idea queda para trabajo futuro, pero la anotamos aquí como una conjetura sugerente.

7.3 Limitaciones del modelo

El modelo presenta varias limitaciones que debemos reconocer. En primer lugar, la elección de las tres funciones constitutivas (δ, σ, φ) es discutible. Otros autores podrían argumentar que faltan dimensiones relevantes (empatia, creatividad, autoconciencia) o que alguna de las tres es redundante. Nuestra posición es que δ, σ y φ constituyen un conjunto minimal razonable, pero estamos abiertos a extensiones.

En segundo lugar, la forma funcional de M (ecuación 1) es una elección entre muchas posibles. Hemos optado por la forma más sencilla que captura tanto las contribuciones independientes como la interacción sinérgica, pero formas más complejas (no lineales, con términos de interacción de orden superior) podrían ser necesarias para una descripción más fiel de la fenomenología moral.

En tercer lugar, la operacionalización empírica de las funciones constitutivas está en un estado incipiente. Mientras que δ puede, al menos en principio, medirse mediante análisis del comportamiento deliberativo del sistema (número de alternativas evaluadas, profundidad de la búsqueda), y σ mediante el análisis de la sensibilidad de las decisiones a restricciones normativas, la medición de φ depende de avances en la ciencia de la consciencia que aún no se han producido. El marco es, en este sentido, más un programa de investigación que un resultado cerrado.

7.4 Implicaciones regulatorias

A pesar de sus limitaciones, el marco tiene implicaciones prácticas para la regulación de la IA. La Unión Europea ha iniciado un proceso de elaboración de un marco regulatorio para la inteligencia artificial que necesitará criterios formales para clasificar los sistemas según su nivel de riesgo. Nuestro modelo sugiere que, además del riesgo funcional (posibilidad de daño material), debería considerarse el "riesgo moral": la proximidad del sistema al umbral crítico de agencia moral. Un sistema cercano a M* plantea problemas regulatorios cualitativamente distintos de los que plantea un sistema lejano, independientemente de su capacidad funcional.

Esta propuesta va más allá de la mera clasificación de riesgos. Implica que el proceso regulatorio debería incorporar un componente de evaluación moral, no en el sentido de juzgar si un sistema es "bueno" o "malo", sino en el de determinar cuál es su posición en el campo de gradiente moral y qué tipo de responsabilidades éticas se derivan de esa posición. Los detalles de cómo implementar tal evaluación quedan fuera del alcance de este trabajo, pero la necesidad de ella se sigue directamente de nuestro marco.

8. Conclusiones

Hemos propuesto un modelo formal de la agencia moral basado en campos escalares continuos sobre el espacio de estados cognitivos de un sistema. Los resultados principales son:

(i) La definición del gradiente moral M como campo diferenciable compuesto por tres funciones constitutivas (deliberación, sensibilidad normativa, sensibilidad fenoménica) más un término de interacción sinérgica.

(ii) El desarrollo de un cálculo diferencial de la agencia con operadores (divergencia, rotacional, laplaciano) dotados de interpretaciones éticas específicas.

(iii) La aplicación del marco al análisis de sistemas artificiales contemporáneos (redes neuronales profundas, agentes de aprendizaje por refuerzo, robots autónomos), mostrando que el modelo permite una clasificación matizada de su estatus moral.

(iv) La identificación de transiciones de fase morales y la definición de un umbral crítico M* con implicaciones directas para la regulación de la IA.

(v) La derivación de una "ley de conservación moral" formal, cuyas implicaciones filosofícas y prácticas merecen exploración ulterior.

El trabajo conecta con nuestro marco topológico previo para la formalización de los qualia (Caracuel Llabrés et al., 2016) y extiende sus implicaciones al terreno de la ética. Tomados conjuntamente, ambos trabajos esbozan un programa de investigación que persigue la formalización matemática rigurosa de nociones filosóficas que, hasta ahora, se han resistido a tal formalización. Somos los primeros en reconocer que este programa está en una fase temprana y que muchas de las propuestas presentadas son tentativas. Pero confiamos en que el mero hecho de que estas preguntas admitan formulación precisa constituye un avance significativo.

Quisiéramos cerrar con una reflexión. El desarrollo de la inteligencia artificial no es, como a veces se presenta, un problema exclusivamente técnico. Es, en su núcleo más profundo, un problema filosófico. Las decisiones de diseño que tomamos hoy al construir sistemas artificiales son, queramos o no, decisiones sobre el tipo de agentes morales que estamos creando. Cuanto antes dispongamos de herramientas formales para pensar con rigor sobre estas cuestiones, mejor preparados estaremos para afrontar lo que viene.

. Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto FFI2014-57409 del Ministerio de Economía y Competitividad (MINECO) y por una beca FPU del Ministerio de Educación, Cultura y Deporte (referencia FPU14/02387). Agradecemos a los miembros del Grupo de Investigación en Filosofía y Ciencias Cognitivas de la Universidad de Sevilla sus comentarios durante la presentación de una versión preliminar de este trabajo en junio de 2016. K.C.Ll. agradece a Martín Aráoz-Gutiérrez (Dpto. de Geometría y Topología, Universidad de Sevilla) sus sugerencias sobre la formalización de los operadores diferenciales, y a los dos revisores anónimos cuyas observaciones han mejorado sustancialmente el manuscrito.

. Referencias

Amari, S. (1998). Natural gradient works efficiently in learning. Neural Computation, 10(2), 251-276.

Arkin, R. C. (2009). Governing Lethal Behavior in Autonomous Robots. Chapman and Hall/CRC.

Aristóteles. Ética a Nicómaco. Traducción de J. Pallí Bonet (1985). Gredos.

Bonnefon, J.-F., Shariff, A. y Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science, 352(6293), 1573-1576.

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

Caracuel Llabrés, K., Vidal-Moreno, E. y Aráoz-Gutiérrez, M. (2016). Invariantes topológicos de la experiencia fenoménica: un marco computacional para el mapeo de qualia. Revista Iberoamericana de Filosofía y Ciencias Cognitivas, 12(3), 247-289.

Chalmers, D. J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

Floridi, L. y Sanders, J. W. (2004). On the morality of artificial agents. Minds and Machines, 14(3), 349-379.

Gunkel, D. J. (2012). The Machine Question: Critical Perspectives on AI, Robots, and Ethics. MIT Press.

Gunkel, D. J. (2014). A vindication of the rights of machines. Philosophy and Technology, 27(1), 113-132.

Kant, I. (1785). Fundamentación de la metafísica de las costumbres. Traducción de M. García Morente (1996). Espasa-Calpe.

LeCun, Y., Bengio, Y. y Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Nagel, T. (1974). What is it like to be a bat? Philosophical Review, 83(4), 435-450.

Regan, T. (1983). The Case for Animal Rights. University of California Press.

Seung, H. S. (1996). How the brain keeps the eyes still. Proceedings of the National Academy of Sciences, 93(23), 13339-13344.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... y Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.

Singer, P. (1975). Animal Liberation. HarperCollins.

Tononi, G. (2004). An information integration theory of consciousness. BMC Neuroscience, 5, 42.

Wallach, W. y Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. Oxford University Press.

Wiener, N. (1960). Some moral and technical consequences of automation. Science, 131(3410), 1355-1358.