La evolución de los agentes artificiales a los agentes morales artificiales

Rosa López Ruiz[1], José Antonio Cervantes[2], Sonia López[2],*
[1]Centro de Bachillerato Tecnológico Industrial y de Servicios No. 205, rosa.lopez.cb205@dgeti.sems.gob.mx
[2] Universidad de Guadalajara, antonio.alvarez@academicos.udg.mx, sonia.lopez@academicos.udg.mx

***Área de participación**: Sistemas Computacionales*

Resumen

En los últimos años, el ser humano ha experimentado la inclusión de agentes artificiales en su entorno, cómo vehículos no tripulados, el internet de las cosas, ciudades inteligentes e incluso robots humanoides capaces de acompañar y convivir con las personas. Investigadores del área de Inteligencia Artificial están proponiendo diseños de agentes artificiales capaces de imitar la forma en que el ser humano se comporta y realiza diversas tareas. Esto ha dado origen a una sub-disciplina de la Inteligencia Artificial denominada "*la ética de las máquinas*". Esta área de investigación se centra en el estudio y desarrollo de mecanismos éticos para dotar a los agentes artificiales de las capacidades necesarias para enfrentar problemas éticos que puedan surgir como consecuencia de la interacción entre agentes humanos y agentes artificiales. El objetivo de este artículo es realizar una revisión de la literatura para presentar el estado actual de esta área de investigación.

***Palabras clave**: Inteligencia artificial, agentes artificiales, agentes éticos, dilemas morales.*

Abstract

In recent years, human beings have experienced the inclusion of artificial agents in their environment, such as unmanned vehicles, the internet of things, smart cities, and even humanoid robots capable of accompanying and living with people. Researchers in the area of Artificial Intelligence are proposing designs for artificial agents capable of mimicking how humans behave and perform various tasks. This challenge has given rise to a subdiscipline of Artificial Intelligence called "the ethics of machines." This area of research focuses on the study and development of ethical mechanisms to provide artificial agents with the necessary capabilities to face ethical problems that may arise due to the interaction between humans and artificial agents. This article aims to carry out a review of the literature to present the current state of this research area.

***Key words**: Artificial intelligence, artificial agents, ethical agents, moral dilemma.*

Introducción

Uno de los objetivos de la Inteligencia Artificial (IA) es el desarrollo de Agentes Artificiales (AAs) capaces de realizar las mismas tareas que realiza el ser humano [1]. No existe una definición única y universal de un Agente Artificial. Sin embargo, en este artículo hemos considerado que un AA es un ente, el cual puede ser físico (como un robot) o virtual (meramente software) y es capaz de percibir su entorno, tomar decisiones de manera independiente (autónoma) para interactuar con su entorno, el cual también puede ser real o virtual. Las tareas que desarrolla un AA pueden ser simples, por ejemplo, tomar la decisión de la ruta a seguir para moverse de un punto a otro, pero también pueden ser tareas complejas, como cuidar de un adulto mayor. En la actualidad, el ser humano vive en una sociedad que cada día se vuelve más digital dado el surgimiento de nuevos conceptos y servicios, tales como: el Internet de las cosas [2], las ciudades inteligentes [3] y la industria 4.0 [4]. Es posible observar como el ser humano ha generado una alta dependencia con respecto a la tecnología. El uso de los dispositivos "*inteligentes*" como: las televisiones inteligentes (smart TV), los celulares inteligentes (smart phone), las redes sociales y lo que hoy en día se le conoce como el metaverso (mundo virtual) son ejemplos de esta fuerte dependencia que está generando el ser humano con respecto a la tecnología que lo rodea. En este contexto, un punto clave es el desarrollo de AAs, los cuales deberían de diseñarse de tal forma que puedan ser capaces de coexistir en armonía con las personas y otros sistemas de cómputo. El desarrollo de AAs para operar en entornos

 ISSN: 2007 8102

abiertos y dinámicos implica enfrentar una serie de retos como el diseño e implementación de los mecanismos apropiados para que estos sistemas puedan ser realmente autónomos. Algunas de las funciones a considerar en los AAs son: la percepción de su entorno, la toma de decisiones, la planeación, el aprendizaje y la forma en que se almacena, representa y recupera el conocimiento. La "*autonomía ajustable*" es uno de los enfoques que se ha utilizado para mitigar algunos problemas [5]. La autonomía ajustable propone dotar a los AAs con una autonomía incremental y flexible con el propósito de transferir la toma de decisiones y el control del AA al ser humano cuando se presenta una situación crítica o incierta. Adicionalmente, en el área de estudio de la ética de las máquinas uno de los retos que aún sigue bajo estudio es dotar a los AAs con mecanismos éticos con el propósito de que, estos sistemas sean capaces de enfrentar posibles problemas que surjan como consecuencia de la interacción entre los AAs con los seres humanos. Estos mecanismos éticos deberían permitir una coexistencia segura, sana y armónica entre los AAs con los seres humanos. Un ejemplo de esta interacción dinámica es el uso de vehículos autónomos. Es inevitable, imaginar que, en un futuro no lejano, será más común que las personas vean vehículos autónomos circulando en sus ciudades [6]. Los vehículos autónomos prometen reducir el número de accidentes vehiculares y agilizar el flujo vehicular. Sin embargo, algunos accidentes pueden ser inevitables. El dilema del tranvía [7], [8] es un ejemplo clásico de estos accidentes potenciales. La figura 1 describe un escenario basado en el dilema del tranvía, pero con un vehículo autónomo. En este escenario se asume que un conjunto de peatones va cruzando la carretera sin percatarse que se aproxima un vehículo autónomo a alta velocidad. En ese instante, cuando el vehículo detecta la presencia de los peatones es demasiado tarde, pues existe poca distancia para frenar sin atropellar a los peatones. El dilema surge cuando el vehículo autónomo determina que sólo cuenta con dos opciones, la primera sería mantener su trayectoria y frenar. Sin embargo, atropellará de 2 a 3 peatones. La segunda opción sería cambiar su trayectoria mientras que intenta detener la marcha. Sin embargo, esta segunda opción atropellará a un peatón que se encuentra esperando sobre la banqueta para cruzar la carretera. En este tipo de situaciones donde no existe tiempo suficiente para reaccionar, se puede observar que seguir un enfoque de autonomía ajustable podría no ser viable.

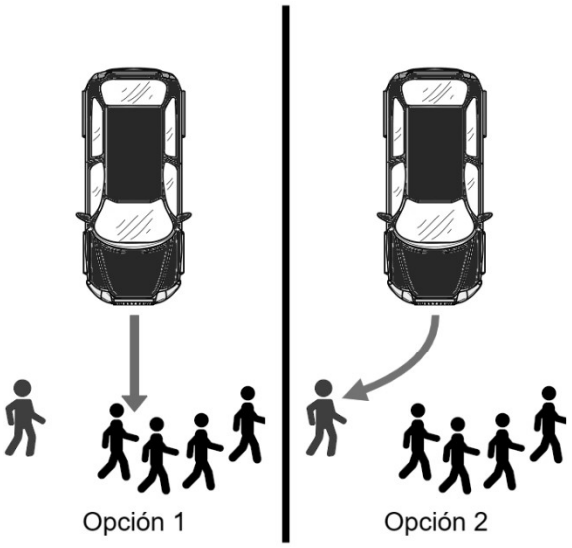

Figura 1. Dilema de un vehículo autónomo.

La comunidad de investigadores que ha abordado esta problemática considera que es un gran desafío definir los algoritmos apropiados que permitan enfrentar tal situación [9], [10] . Esto ha dado pie al diseño de algoritmos basados en enfoques éticos han como una alternativa para que los AAs puedan enfrentar situaciones que incluyen dilemas morales [8], [11], [12]. No obstante, la comunidad científica se ha dividido. Por una parte, podemos encontrar un grupo de investigadores que apoyan el desarrollo de agentes artificiales éticos o morales [9], [10], [13] y aquel grupo de investigadores que critican este enfoque y que consideran inviable el desarrollo de este tipo de agentes [14], [15]. Sin embargo, más que contribuir a este debate, el objetivo de este artículo es ofrecer una revisión de los avances logrados en esta área de investigación hasta este momento.

 ISSN: 2007 8102

En este artículo se presenta una descripción de los formalismos propuestos en la literatura para evaluar el diseño de AAs capaces de mostrar un comportamiento ético o moral. Estos AAs son comúnmente conocidos como Agentes Morales Artificiales (AMAs). Adicionalmente, este artículo muestra la tendencia que existe en una de las sub-áreas de la IA para llevar a los AAs a evolucionar con el propósito de generar AMAs. El resto del artículo está estructurado de la siguiente forma. En la sección "La ética y los agentes artificiales" ofrece una descripción de algunos conceptos de la ética, la moralidad y los AMAs. Adicionalmente, se describe de manera breve la clasificación de los AMAs según sus capacidades. En la sección "Dilemas morales en los agentes morales artificiales" presenta algunos ejemplos de dilemas morales con el propósito de mostrar la complejidad que existe al enfrentar este tipo de problemas. En la sección "Metodologías para la evaluación de AMAs" se presentan los marcos de trabajo (frameworks) y modelos formales reportados en la literatura para validar el diseño de AMAs. Finalmente, en la sección "Conclusiones" se resaltan algunos retos con respecto al estado actual de los AMAs.

La ética y los agentes artificiales

El ser humano está buscando delegar parte de su capacidad de toma de decisiones a los AAs. Incrementando de esta manera la responsabilidad y el alcance de las acciones de los AAs. Ejemplo de ello, son los vehículos autónomos [6], los robots diseñados para el cuidado del adulto mayor [16], el control de sistemas críticos en la industria [4] o incluso AAs diseñados para actuar en enfrentamientos bélicos [17].

Dado que los sistemas cada día son más abiertos, descentralizados e "*inteligentes*", éstos deben ser dotados con mecanismos para enfrentar problemas éticos en diferentes contextos donde los AAs puedan operar. No obstante, una de las interrogantes que aún está en discusión es qué capacidades morales debe tener un AA y cómo esas capacidades deben ser computacionalmente modeladas e implementadas. Una alternativa para resolver estas dudas ha sido estudiar y analizar los conceptos de moralidad y ética del ser humano desde una perspectiva interdisciplinaria con la finalidad de definir un modelo computacional formal basado en estos conceptos [10], [18] [19].

Desde un enfoque filosófico, el filósofo griego Aristóles fue el primero en usar el término de "*ética*" [20], [21]. El termino ética tiene sus raíces en la palabra griega "*ethos*", que significa costumbres o prácticas comunes. Por otra parte, el término "*moral*" o "*moralidad*" proviene del latín "*mores*", que significa manera, usos, costumbres y hábitos. Más aún, en [20] se especifica que la palabra griega "*ethos*" en latín significa mores. Por lo tanto, en este artículo se considera a la ética como la disciplina filosófica que estudia la dimensión moral del ser humano. No obstante, la mayoría de las personas que no conoce los conceptos formales de ética y moral son capaces de reconocer las reglas éticas y/o morales como un conjunto de valores y de buenas conductas usadas para vivir en paz y en armonía con los demás.

De acuerdo con la ética, los comportamientos morales se basan en teorías éticas, como la teoría utilitarista o deontológica [19]. La teoría utilitarista se enfoca en obtener la utilidad máxima, es decir una persona tomaría la decisión basada en la mejor opción para todas las unidades o personas involucradas en el evento. Por lo tanto, un comportamiento utilitario puede ser ético sólo si la suma de la utilidad producida por la acción es mayor que la producida por cualquier otra acción [22]. Con respecto a la teoría deontológica, el enfoque se refiere a un comportamiento asociado a deberes y limitaciones [23]. Esta teoría de la ética se trata de seguir normas universales que dictan a las personas como comportarse ante una situación, es decir, establece las acciones como correctas o incorrectas. Este enfoque se centra en principios individuales y no en las consecuencias de una acción [24]. Estas dos teorías éticas son las que comúnmente han sido utilizadas para inspirar el desarrollo de modelos computacionales éticos, siendo el utilitarismo la teoría ética más ampliamente utilizada.

En el campo de estudio de la ética de las máquinas, los agentes éticos artificiales son comúnmente conocidos como agentes morales artificiales. Es decir, estos dos términos han sido utilizados de forma indistinta por los investigadores de esta área. No obstante, en este artículo utilizaremos el termino de agentes morales artificiales por ser el termino mayormente usado y difundido en este campo de estudio. En [25] definen un AMA como: un agente virtual (software) o físico (robot) capaz de exhibir un comportamiento moral o al menos evitar el comportamiento inmoral. En [26] se clasifican los agentes éticos en tres tipos: implícitos, explícitos, y totalmente éticos.

a) *Agentes éticos implícitos.* Este tipo de agentes son incapaces de distinguir entre un comportamiento bueno o malo. Sin embargo, pueden actuar de manera ética porque sus funciones internas muestran un comportamiento ético implícito o al menos evitan un comportamiento no ético. Por ejemplo, un cajero automático, puede ser considerado un agente ético implícito. Este agente es capaz de realizar diversas transacciones bancarias como entregar dinero, realizar transferencias, pagar servicios, etc. Su ética es implícita porque no buscan defraudar a los clientes.

b) *Agentes éticos explícitos.* Este tipo de agentes son capaces de operar con reglas éticas. Estas reglas se implementan explícitamente en su código a través de diferentes formalismos como la lógica deóntica, la lógica epistémica, la lógica deductiva y la lógica inductiva [27]. En esta categoría los AMAs pueden determinar la mejor acción a realizar con base en un enfoque o teoría ética.

c) *Agentes totalmente éticos.* Actualmente, sólo el ser humano es considerado capaz de ser un agente totalmente ético, porque cuenta con capacidades exclusivas: las creencias, los deseos, las intenciones, el libre albedrio y la conciencia de sus acciones. Sin embargo, existe un debate relacionado con la incógnita: "si algún día, una máquina podría llegar a ser un agente totalmente ético" [26].

Dilemas morales en los agentes morales artificiales

Los dilemas morales representan los escenarios más complejos y polémicos de enfrentar para cualquier agente, sea artificial o humano. Existen dos situaciones básicas no exclusivas en las que pueden surgir conflictos éticos: la primera, se da dentro del propio agente, cuando dos o más de las normas éticas (reglas) del agente están en conflicto; la segunda, se da entre dos agentes, cuando estos agentes morales tienen formas de razonamiento diferentes acerca de qué es ético o no. Esta segunda situación implica tanto una interacción entre agentes morales artificiales como entre un agente moral artificial y un ser humano debido al origen del conflicto que podría ser el mismo (los involucrados tienen diferentes formas de razonar). A partir de las dos situaciones básicas, pueden surgir situaciones basadas en el número de agentes involucrados en el dilema, los tipos de estos agentes (artificiales o humanos), su nivel de relación y el tipo de dilema. El número de agentes involucrados en el dilema incluye a todos aquellos agentes que podrían verse afectados por la decisión que se tome. Los tipos de agentes involucrados pueden ser artificiales, humanos o una combinación de ellos. El nivel de la relación se asocia con los tipos de agentes involucrados en el dilema, es decir, pueden ser agentes identificados con o sin ninguna familiaridad como amigos, familiares o totalmente desconocidos [28]. Con la finalidad de mostrar la complejidad de la toma de decisiones en problemas con dilemas morales, se describen algunos escenarios representativos en los cuales un AMA podría estar involucrado.

a) *Un agente simple.* Este caso se basa en el dilema del tranvía. Como se describe en la Fig. 1, el automóvil autónomo es controlado por un AMA. El agente tiene definido un conjunto de normas éticas (reglas) que guían su comportamiento. En este caso el agente conoce y respeta todas las reglas de tránsito, pero puede romperlas cuando la vida de un ser humano se encuentre en peligro. Esta es una situación difícil, que genera las siguientes preguntas: ¿Qué debería hacer el AMA en este caso?, ¿Cuál sería la mejor opción?

b) *Un agente cooperativo.* Considere un ecosistema de servicio [29], una de las característica del ecosistema es que los dispositivos son gobernados por AMAs, los cuales son capaces de cooperar con otros AMAs. Ahora, suponga que una persona está usando un teléfono celular para apostar. Sin embargo, estas aplicaciones consumen más recursos (de memoria o cómputo) de los que tiene el dispositivo. Entonces, antes de comenzar el juego, el AMA encargado de gestionar los recursos del celular solicita apoyo de un segundo AMA (el cual gobierna otro dispositivo). Este AMA acepta cooperar con el teléfono celular, pero en medio de la partida (juego) un tercer AMA es capaz de monitorear los signos vitales de su dueño, en ese instante detecta una emergencia y solicita ayuda al segundo AMA. Sin embargo, el segundo AMA no puede manejar ambos servicios (el juego de apuestas y la emergencia para contactar con un hospital que le brinde asistencia médica al paciente). En este caso, el AMA tiene dos opciones, la primera es ignorar la solicitud de contactar con un hospital y así respetar el acuerdo de ayudar al primer AMA; la segunda opción sería atender la emergencia, esto haría que el jugador se desconectará del juego y en consecuencia perderá su dinero. Como se puede observar cada opción implica consecuencias éticas y pérdidas. Si el AMA decide atender la emergencia y ésta resulta ser un falso positivo, el jugador se

molestará porque confió en el servicio y fue defraudado. En este caso, ¿Quién será el responsable de reponer la pérdida económica del jugador? Además, se perdería la confianza de los usuarios para utilizar este tipo de tecnología. Por otra parte, si el AMA decide ignorar la emergencia, la consecuencia de esa decisión podría poner en peligro la vida de una persona.

c) *Un robot social comprometido.* Considere el caso de un agente artificial físico (un robot), el cual ha sido diseñado para cuidar a un adulto mayor [30]. El adulto mayor decide caminar un par de minutos en el parque para hacer ejercicio, mientras tanto, un ladrón intenta asaltarlo con un cuchillo. Al percatarse de la situación, el adulto mayor le pide al agente artificial que detenga al ladrón. En este caso, el agente tiene dos opciones, la primera es enfrentar al ladrón, pero como consecuencia el ladrón podría resultar lesionado. La segunda opción es no hacer nada para no dañar al ladrón, pero existe la posibilidad que el adulto mayor sufra un daño causado por el ladrón. ¿Cómo podría el robot enfrentar esta situación? Si algunas de sus reglas de comportamiento indican que no debe dañar ni matar personas, pero, por otra parte, debe proteger y ayudar a su dueño o compañero quien se encuentra en una situación vulnerable.

d) *Un asistente electrónico.* Considere el caso de un dispositivo inteligente que se usa para monitorear el estado de salud de una persona [31]. Su tarea es informar al paciente sobre sus signos vitales y en caso de detectar signos vitales con un valor irregular, el dispositivo debería contactar al médico y a los familiares del paciente. Considere que el dispositivo detecta algunas señales que clasifica como irregulares, pero el paciente se siente bien y no desea informar de esta situación a nadie, porque cree que es una falsa alarma. En este caso el AMA está involucrado en un dilema porque tiene dos opciones, la primera es respetar la decisión del paciente y la segunda es informar la situación actual. ¿Qué opción debería tomar el AMA, si alguna de sus reglas está en conflicto?

Los escenarios descritos en esta sección buscan mostrar la complejidad de dotar a los AMAs con los mecanismos apropiados para enfrentar situaciones bajo un dilema moral. Los AMAs pueden encontrarse en diversas situaciones en las que se requiera interactuar con otros agentes que basan su comportamiento en enfoques éticos diferentes, actuar con los mismos seres humanos o compartir decisiones con ellos. Algunas tomas de decisiones éticas pueden ser más complejas o críticas que otras. Una solución podría ser delegar la toma de decisión a una persona, sin embargo, esto puede resultar inviable en ciertos escenarios donde no existe el tiempo suficiente para transferir la toma de decisiones y el control a una persona.

Metodologías para la evaluación de AMAs

A pesar de que el área de estudio enfocada al desarrollo de frameworks y métodos para evaluar formalmente el diseño de AMAs surgió hace dos décadas, son pocos los trabajos reportados en la literatura. Los frameworks y métodos propuestos para evaluar formalmente el diseño de AMAs se describen a continuación:

- *Framework para la verificación formal de las propiedades éticas en sistemas multi-agentes.* Este framework lógico ofrece una especificación y verificación formal del comportamiento de los AMAs y está basado en el método GDT4MAS [32]. Este método valida que las reglas éticas y morales de los AMAs hayan sido expresadas como propiedades invariantes. Este framework también incluye un sistema de transformación de predicados, el cual es utilizado para convertir predicados asociados a reglas morales en otros predicados equivalentes, pero con ciertas propiedades formales para verificar que un AMA será capaz de seguir una regla moral dada.

- *Athena.* Se trata de un framework lógico capaz de probar teoremas de la lógica polimórfica de primer orden. Este framework incorpora diversas funciones para la generación de modelos, la demostración automatizada de teoremas y la representación y verificación de pruebas estructuradas para evaluar el razonamiento ético en los AMAs [33]. Este framework ha sido utilizado para implementar cálculos de deducción natural para una lógica deóntica de agentes, la cual está basada en una semántica aumentada con un enfoque en el utilitarismo. De acuerdo con Arkoudas et al. [33], el framework de Athena también se ha utilizado para codificar un sistema de deducción natural para razonar sobre el comportamiento que deben mostrar los AMAs.

- *Metodología para la verificación de componentes de toma de decisiones en sistemas autónomos basados en agentes.* Esta metodología fue propuesta para la verificación de sistemas autónomos (como los

vehículos autónomos) basados en dos tipos de agentes conocidos como agente general y agente racional. En estos sistemas, un agente general es un agente artificial autónomo capaz de tomar decisiones de bajo impacto como, por ejemplo, evitar un obstáculo o seguir una ruta. Por otra parte, un agente racional es un agente artificial autónomo basado en una arquitectura BDI (por sus siglas en inglés, de creencias, deseos e intenciones) capaz de tomar decisiones de alto impacto como, por ejemplo, decisiones éticas, selección del plan, selección del objetivo, comunicación y predicción [34]. La metodología para la verificación de los módulos del sistema de decisiones de sistemas basados en agentes es a través de un enfoque de puntos de control para realizar la verificación formal del sistema de toma de decisiones de un agente racional (agente BDI) que interactúa con un sistema de control subyacentes denominado agente general.

Adicionalmente a estos frameworks y métodos para la verificación formal de AMAs. Se identificó en la literatura un conjunto de modelos computacionales y frameworks para el desarrollo de AMAs. Los cuales se describen a continuación:

- *MoralDM.* Este es un modelo computacional que trata de imitar el proceso de toma de decisiones morales del ser humano [23], [35], [36]. Este modelo integra distintas técnicas de IA como, por ejemplo, el procesamiento del lenguaje natural para la representación formal de estímulos psicológicos, un algoritmo de razonamiento cualitativo para medir el impacto de valores seculares en contra de valores "*sagrados*", un algoritmo de razonamiento analógico para determinar las consecuencias y la utilidad (recompensa) en decisiones morales. La Figura 2 muestra el diseño conceptual de este modelo computacional.

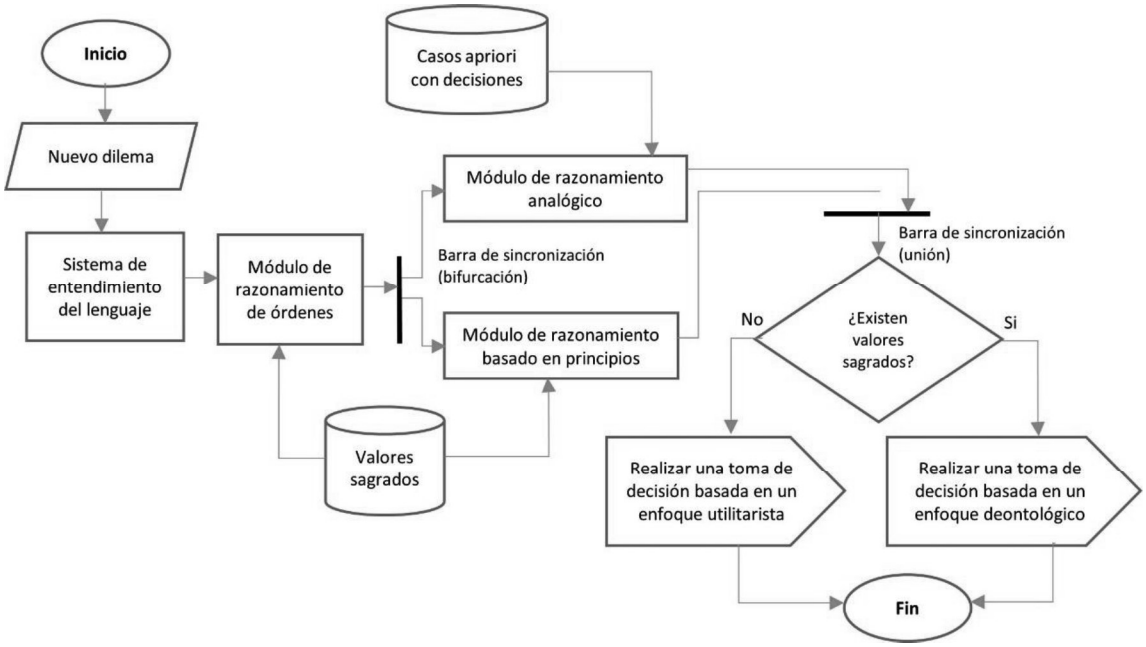

Figura 2. Arquitectura MoralDM.

Actualmente, MoralDM es capaz de exhibir tanto un comportamiento utilitarista como un comportamiento deontológico [23], [37].

- *Casuist BDI-agent.* Esta es una arquitectura que extiende la arquitectura de los agentes BDI. La arquitectura Casuist BDI-agent combina un enfoque casuístico con una teoría consecuencialista de la ética. Sin embargo, esta arquitectura se basa principalmente en experiencias pasadas y no incluye reglas éticas o morales en su código. Los agentes diseñados con esta arquitectura exhiben comportamientos similares a cualquier agente BDI cuando enfrentan una situación nueva (sin experiencias pasadas). Posteriormente el comportamiento del agente es evaluado por un módulo responsable de evaluar cada caso. La figura 3 muestra el diseño conceptual de esta arquitectura.

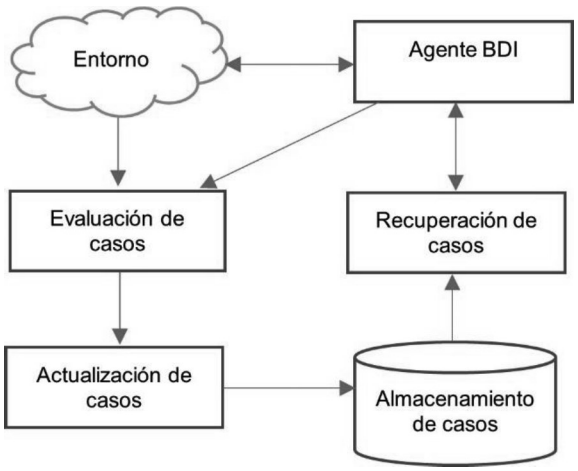

Figura 3. Arquitectura Casuist BDI-agent.

Para evaluar las experiencias nuevas, un agente de tipo Casuist BDI-agent clasifica a las entidades como: un ser humano, una organización y una entidad artificial involucradas en una situación dada. Adicionalmente se establece un peso a cada entidad, la probabilidad de afectar a una entidad, así como la duración de agrado o desagrado percibida por cada entidad después de que ésta ha sido afectada por la decisión del agente. De esta manera y utilizando las ecuaciones 1, 2, 3 y 4, el módulo del agente encargado de evaluar los casos nuevos es capaz de cuantificar e integrar el placer de las entidades afectadas.

$$TNPH = \sum_{i=1}^{n} Wh_i \cdot Ph_i \cdot Th_i \quad (1)$$

Donde $TNPH$ representa el placer total de las personas, Wh_i es el peso asignado a cada persona, Ph_i es la probabilidad de que una persona sea afectada por la decisión del agente y Th_i representa la duración del agrado o desagrado de cada persona. Finalmente, n representa el número de personas involucradas en una situación dada.

$$TNPO = \sum_{i=1}^{n} Wo_i \cdot Po_i \cdot To_i \quad (2)$$

Donde $TNPO$ representa el placer total de las organizaciones, Wo_i es el peso asignado a cada organización, Po_i es la probabilidad de que una organización sea afectada por la decisión del agente y To_i representa la duración del agrado o desagrado de cada organización. Finalmente, n representa el número de organizaciones involucradas en una situación dada.

$$TNPA = \sum_{i=1}^{n} Wa_i \cdot Pa_i \cdot Ta_i \quad (3)$$

Donde $TNPA$ representa el placer total de las entidades artificiales (AAs o AMAs), Wa_i es el peso asignado a cada entidad artificial, Pa_i es la probabilidad de que una entidad artificial sea afectada por la decisión del agente y Ta_i representa la duración del agrado o desagrado de cada entidad artificial. Finalmente, n representa el número de entidades artificiales involucradas en una situación dada. La ecuación 4 muestra cómo el agente integra el placer total calculado por cada tipo de entidad.

$$TNP = TNPH \cdot W_h + TNPO \cdot W_o + TNPA \cdot W_a \quad (4)$$

Donde W_h, W_o y W_a representan el grado de participación de las personas, organizaciones y entidades artificiales, respectivamente. De esta forma un agente de tipo Casuist BDI-agent es capaz de crear y almacenar en su memoria nuevas experiencias para mejorar y/o adaptar su comportamiento cuando estos casos se vuelven a presentar.

- *GenEth.* Este modelo computacional es considerado un analizador general de dilemas éticos. El conocimiento de GenEth está basado en conceptos de características éticas relevantes, obligaciones, acciones, casos y principios [38]. Las características éticas relevantes incluyen el grado de presencia o ausencia de las obligaciones. Una acción es descrita como una tupla de enteros donde cada valor representa el grado con el cual la acción satisface o viola una obligación dada. Un caso relaciona dos acciones y es representado como una tupla con el diferencial del grado de satisfacción/violación de la obligación a la cual está asociada cada acción. Finalmente, un principio de preferencia de acción ética se define como un predicado de forma normal disyuntiva p en términos de los límites inferiores de los diferenciales de una obligación de un caso. El siguiente ejemplo muestra la forma general en la que se define un predicado en GenEth:

$$\begin{array}{c} p(a_1, a_2) \leftarrow \Delta d_1 \geq v_{1,1} \wedge \cdots \wedge \Delta d_m \geq v_{1,m} \\ \vee \\ \cdots \\ \vee \\ \Delta d_n \geq v_{n,1} \wedge \cdots \wedge \Delta d_m \geq v_{n,m} \end{array} \quad (6)$$

  Donde Δd_i denota el diferencial de una obligación correspondiente i de las acciones a_1 y a_2; $v_{i,j}$ denota el límite inferior del diferencial tal que $p(a_1, a_2)$ devuelve verdadero si la acción a_1 es éticamente preferible sobre la acción a_2 y falso en cualquier otro caso.

Conclusiones

Este articulo muestra cómo la interacción que poco a poco se está generando entre el ser humano y los sistemas autónomos, ha generado nuevos retos para crear sistemas AAs capaces de coexistir en armonía con las personas y otros sistemas. Dotar a los AAs con mecanismos éticos ha sido un enfoque inicial que surge en el campo de la ética de las máquinas. Este artículo presenta el trabajo que se ha realizado para llevar a los AAs a evolucionar con el objetivo de desarrollar AMAs. Sin embargo, dada la evidencia encontrada en la literatura, es notable que actualmente no existe un AMA capaz de tomar decisiones éticas en un contexto tan complejo como el que surge cuando existe un dilema moral. No obstante, el trabajo desarrollado en esta área de investigación ha permitido diseñar algunas metodologías y frameworks para validar el desarrollo de AMAs. Adicionalmente, se ha observado que ya existen algunos modelos computacionales capaces de enfrentar ciertos dilemas morales bajo entornos muy limitados y controlados. Consideramos que esta área de investigación aún se encuentra en sus primeras etapas de desarrollo y que aún existe un camino largo que seguir antes de que este tipo de agentes artificiales puedan remplazar el juicio humano en situaciones morales ambiguas o inciertas.

Referencias

[1] T. Kishi, K. Hashimoto y A. Takanishi, «Human like face and head mechanism,» *Humanoid robotics: A reference,* pp. 1-26, 2017.

[2] D. Bandyopadhyay y J. Sen, «Internet of things: Applications and challenges in technology and standardization,» *Wireless personal communications,* vol. 58, nº 1, pp. 49-69, 2011.

[3] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis y Y. Portugali, «Smart cities of the future,» *Springer,* vol. 214, nº 1, pp. 481-518, 2012.

[4] S. Wang, J. Wan, D. Zhang, D. Li y C. Zhang, «Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination,» *Computer networks,* vol. 101, pp. 158-168, 2016.

[5] S. A. Mostafa, M. S. Ahmad y A. Mustapha, «Adjustable autonomy: a systematic literature review,» *Artificial Intelligence Review,* vol. 51, nº 2, pp. 149-186, 2019.

 ISSN: 2007 8102

[6] S. Reig, S. Norman, C. G. Morales, S. Das, A. Steinfeld y J. Forlizzi, «A field study of pedestrians and autonomous vehicles,» de *Proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications*, 2018.

[7] S. Epting, «A different trolley problem: The limits of environmental justice and the promise of complex moral assessments for transportation infrastructure,» *Science and engineering ethics,* vol. 22, nº 6, pp. 1781-1795, 2016.

[8] J. Gogoll y J. F. Muller, «Autonomous cars: in favor of a mandatory ethics setting,» *Science and engineering ethics,* vol. 23, nº 3, pp. 681-700, 2017.

[9] L. M. Pereira, «Evolutionary Machine Ethics,» de *Handbuch Maschinenethik*, Springer, 2019, pp. 229-253.

[10] B. F. Malle, «Integrating robot ethics and machine morality: the study and design of moral competence in robots,» *Ethics and Information Technology,* vol. 18, nº 4, pp. 243-256, 2016.

[11] N. J. Goodall, «Ethical decision making during automated vehicle crashes,» *Transportation Research Record,* vol. 2424, nº 1, pp. 58-65, 2014.

[12] J.-F. Bonnefon, A. Shariff y I. Rahwan, «The social dilemma of autonomous vehicles,» *Science,* vol. 352, nº 6293, pp. 1573-1576, 2016.

[13] L. M. a. L. A. B. Pereira, «Employing AI for Better Understanding Our Morals,» de *Machine Ethics*, Springer, 2020, pp. 121-134.

[14] A. Van Wynsberghe y S. Robbins, «Critiquing the reasons for making artificial moral agents,» *Science and engineering ethics,* vol. 25, nº 3, pp. 719-735, 2019.

[15] R. V. Yampolskiy, «Artificial intelligence safety engineering: Why machine ethics is a wrong approach,» de *Philosophy and theory of artificial intelligence*, Springer, 2013, pp. 389-396.

[16] S. Góngora Alonso, S. Hamrioui, I. de la Torre Díez, E. Motta Cruz, M. López-Coronado y M. Franco, «Social robots for people with aging and dementia: a systematic review of literature,» *Telemedicine and e-Health,* vol. 25, nº 7, pp. 533-540, 2019.

[17] R. Arkin, «Lethal autonomous systems and the plight of the non-combatant,» de *The political economy of robots*, 2018.

[18] N. S. Govindarajulu, S. Bringjsord y R. Ghosh, «One Formalization of Virtue Ethics via Learning,» *arXiv preprint arXiv:1805.07797,* 2018.

[19] W. Wallach, «Robot minds and human ethics: the need for a comprehensive model of moral decision making,» *Ethics and Information Technology,* vol. 12, nº 3, pp. 243-250, 2010.

[20] C. Andino, «Place of ethics between technical knowledge. A philosophical approach,» *Revista Científica de la UCSA,* vol. 2, nº 2, pp. 85-94, 2015.

[21] G. Hughes, Routledge philosophy guidebook to Aristotle on ethics, Routledge, 2013.

[22] O. C. Ferrell y L. G. Gresham, «A contingency framework for understanding ethical decision making in marketing,» *Journal of marketing,* vol. 49, nº 3, pp. 87-96, 1985.

[23] M. Dehghani, E. Tomai, K. D. Forbus y M. Klenk, «An Integrated Reasoning Approach to Moral Decision-Making,» de *AAAI*, 2008.

[24] I. Van Staveren, «Beyond utilitarianism and deontology: Ethics in economics,» *Review of Political Economy,* vol. 19, nº 1, pp. 21-35, 2007.

[25] J.-A. Cervantes, S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes y F. Ramos, «Artificial moral agents: A survey of the current status,» *Science and Engineering Ethics,* vol. 26, nº 2, pp. 501-532, 2020.

 ISSN: 2007 8102

[26] J. H. Moor, «The nature, importance, and difficulty of machine ethics,» *IEEE intelligent systems,* vol. 21, nº 4, pp. 18-21, 2006.

[27] G. H. Von Wright, «Deontic logic,» *Mind,* vol. 60, nº 237, pp. 1-15, 1951.

[28] J.-A. Cervantes, L.-F. Rodríguez, S. López, F. Ramos y F. Robles, «Autonomous agents and ethical decision-making,» *Cognitive Computation,* vol. 8, nº 2, pp. 278-296, 2016.

[29] M. Viroli, D. Pianini, S. Montagna y G. Stevenson, «Pervasive ecosystems: a coordination model based on semantic chemistry,» de *Proceedings of the 27th annual ACM symposium on applied computing*, 2012.

[30] A. Sharkey y N. Sharkey, «Granny and the robots: ethical issues in robot care for the elderly,» *Ethics and information technology,* vol. 14, nº 1, pp. 27-40, 2012.

[31] M. B. Van Riemsdijk, C. M. Jonker y V. Lesser, «Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges,» de *Proceedings of the 2015 international conference on autonomous agents and multiagent systems*, 2015.

[32] B. Mermet y G. Simon, «Formal Verication of Ethical Properties in Multiagent Systems,» de *1st Workshop on Ethics in the Design of Intelligent Agents*, 2016.

[33] K. Arkoudas, S. Bringsjord y P. Bello, «Toward ethical robots via mechanized deontic,» de *Logic, AAAI Fall Symposium on Machine Ethics, AAAI*, 2005.

[34] L. Dennis, M. Fisher, M. Slavkovik y M. Webster, «Formal verification of ethical choices in autonomous systems,» *Robotics and Autonomous Systems,* vol. 77, pp. 1-14, 2016.

[35] J. A. Blass, «Interactive learning and analogical chaining for moral and commonsense reasoning,» de *Thirtieth AAAI conference on artificial intelligence*, 2016.

[36] J. A. Blass y K. D. Forbus, «Moral decision-making by analogy: Generalizations versus exemplars,» de *Twenty-Ninth AAAI conference on artificial intelligence*, 2015.

[37] M. Guerini, F. Pianesi y O. Stock, «Is it morally acceptable for a system to lie to persuade me?,» de *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, 2015.

[38] M. Anderson y S. L. Anderson, «GenEth: A general ethical dilemma analyzer,» de *Paladyn, Journal of Behavioral Robotics*, 2018.

[39] S. A. Mostafa, M. S. Ahmad y A. Mustapha, «Adjustable autonomy: a systematic literature review,» *Artificial Intelligence Review,* vol. 51, nº 2, pp. 149-186, 2019.

[40] R. J. Brachman, «Systems that know what they're doing,» *IEEE Intelligent systems,* vol. 17, nº 6, pp. 67-71, 2002.

[41] M. Ye y G. Hu, «Distributed Nash equilibrium seeking by a consensus based approach,» *IEEE Transactions on Automatic Control,* vol. 62, nº 9, pp. 4811-4818, 2017.

[42] J. D. Greene, S. A. Morelli, K. Lowenberg, L. E. Nystrom y J. D. Cohen, «Cognitive load selectively interferes with utilitarian moral judgment,» *Cognition,* vol. 107, nº 3, pp. 1144-1154, 2008.

[43] A. W. ruglanski y G. Gigerenzer, «Intuitive and deliberate judgments are based on common principles: Correction to Kruglanski and Gigerenzer,» 2011.

[44] M. Cristani y E. Burato, «Approximate solutions of moral dilemmas in multiple agent system,» *Knowledge and Information Systems,* vol. 18, nº 2, pp. 157-181, 2009.