ESTUDIOS PENALES
Y CRIMINOLÓGICOS

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8917

Artículos Doctrinales

LA PERCEPCIÓN DE LA TOMA DE DECISIONES A TRAVÉS DE INTELIGENCIA ARTIFICIAL CUANDO SE PRODUCE DAÑO A LAS PERSONAS

PERCEPTION OF ARTIFICAL INTELLIGENCE DECISION MAKING WHEN PEOPLE RECEIVE HARM

Pablo Espinosa[1,a,*], Miguel Clemente[1,b,**]
[1] Facultad de Ciencias de la Educación, Campus de Elviña, s/n 15071, Universidade da Coruña, Campus de Elviña, s/n 15071, España

[a]pablo.espinosa.breen@udc.es
[b]miguel.clemente@udc.es

Resumen

La toma de decisiones en inteligencia artificial (IA) puede ocurrir en escenarios en los que se decide en una fracción de segundo sobre la vida o el bienestar de los individuos sin que medie supervisión humana. Los algoritmos de IA que se aplican en estos casos pueden basarse o bien en criterios deontológicos o utilitaristas. Incluso si hubiese un consenso ético sobre la toma de decisiones de la IA, si la gente no encontrase aceptables los criterios éticos de la IA, su rechazo dificultaría su implementación. Por ejemplo, si un coche autónomo siempre sacrificase la seguridad de sus pasajeros antes que poner en peligro a otras víctimas en un accidente inevitable, mucha gente no compraría un coche autónomo. En este artículo se realiza una revisión bibliográfica de artículos científicos del ámbito de la psicología social sobre las variables implicadas en la percepción de decisiones relacionadas con la IA. Esta percepción social de la IA puede tener relevancia en el desarrollo de criterios sobre la responsabilidad legal. Finalmente, se examinan aspectos relacionados con el ámbito jurídico con la utilización de la IA en el sistema judicial y en la comisión de delitos.

Palabras clave: Inteligencia Artificial; Deontología; Utilitarismo; Personalidad; Dilemas de Sacrificio.

[*] Profesor titular de universidad, Departamento de Psicología

[**] Catedrático de universidad, Departamento de Psicología

Abstract

Artificial Intelligence (AI) decision making may happen in scenarios where in a split second a decision has to be made without human supervision over the life or well-being of individuals. AI algorithms used in these cases can be based on deontological or utilitarian criteria. Even if there was a normative consensus in the ethics of AI decision making, if people did not find acceptable AI ethical criteria, its rejection would hinder its implementation. For instance, if an autonomous car would always sacrifice its passengers' safety rather than risking other victims in an unavoidable accident, a lot of people would choose not to buy an autonomous car. In this paper we revise Social Psychology research papers related the variables involved in the perception of AI decision making. Social perception of AI may be relevant in developing legal responsibility criteria. Finally, we examine issues related to the legal field, like the use of AI in the legal system and to commit crimes.

Keywords: Artificial Intelligence; Deontology; Utilitarianism; Personality; Sacrificial Dilemmas.

1. INTRODUCCIÓN

El creciente número de ámbitos en los que se aplica la Inteligencia Artificial (IA) conduce a que cada vez en más ocasiones una máquina o un algoritmo informático tome decisiones que afectan al bienestar de las personas o que incluso decida sobre quién vive o muere en determinados contextos. Cada vez en más ocasiones la IA toma decisiones morales sobre lo que es correcto en situaciones en las que ese produce un daño para las personas. A través de sus algoritmos de programación, las máquinas deciden de manera autónoma e instantánea y sin ningún tipo de supervisión humana, quién debe vivir y morir en contextos como el sanitario (e.g. Triaje de pacientes COVID[1]), militares (e.g. Lethal Autonomous Weapon Systems o LAWS[2]) y de conducción autónoma[3] entre otros.

La Inteligencia Artificial (IA) es una rama de las ciencias informáticas centrada en modelos estadísticos y procesamiento de la información que ejecutan recrean o emulan funciones previamente asociadas con la inteligencia humana, como el razonamiento, aprendizaje y automejora[4] . Las máquinas dotadas de IA toman decisiones de manera autónoma y sin supervisión humana, y si cuentan con sistemas de aprendizaje automático o *machine learning*, son capaces de modificar su proceso de toma de decisiones en función de los inputs que reciben. Las decisiones que toma la IA *son cómo* las que tomaría un ser humano y sus resultados no se distinguen de una decisión humana. Sin embargo esto no significa que las máquinas sean inteligentes o que piensen, lo cual constituye una falacia y cae dentro del campo de la superstición. Ante la incomprensión de su funcionamiento y la apariencia de que las decisiones son como las de un humano, se atribuiría a la IA rasgos que no posee, como el de comprender el significado de sus acciones. La IA se define por los resultados que produce, no por cómo llega a ellos[5] .

Es importante determinar cómo las máquinas deben tomar decisiones que para nosotros tienen un componente moral y también qué decisiones resultan generalmente más aceptables. Aunque los criterios éticos que deben incluir los algoritmos de IA no tienen por qué coincidir necesariamente con lo que resulte aceptable para el público, la utilización de la IA en diferentes contextos se verá influida por la compresión y aceptación de estos criterios. Incluso si hubiese un consenso sobre cómo legislar sobre las decisiones morales ejecutadas por la IA, si los protocolos de decisión no fuesen aceptables para los usuarios, su adopción y utilización estaría comprometida[6] .

---

[1] ELLEUCH, M. A., BEN HASSENA, A., ABDELHEDI, M. y PINTO, F.S., "Real-time prediction of COVID-19 patients health situations using Artificial Neural Networks and Fuzzy Interval Mathematical modeling", en *Applied Soft Computing*, 110, 2021.

[2] DE AGREDA, A. G. "Ethics of autonomous weapons systems and its applicability to any AI systems", en *Telecommunications Policy*, 44(6), 2020.

[3] AWAD, E., DSOUZA, S., KIM, R., SCHULZ, J. et al., *"The Moral Machine Experiment"*, en *Nature*, 563(7729), 2018, pp. 59-64.

[4] DE SILES, E.L. "AI, on the Law of the Elephant: Toward Understanding Artificial Intelligence", en *Buffalo Law Review*, 69(5), 2021, pp.1389-1469.

[5] KING, T.C., AGGARWAL, N., TADDEO, M. y FLORIDI, L., "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions". *Science and Engineering Ethics*, 26(1), 2020, pp. 89-120.

[6] AWAD, E., DSOUZA, S., KIM, R., SCHULZ, J. et al., *"The Moral Machine Experiment"*, en *Nature*, 563(7729), 2018, pp. 59-64.

El objetivo del presente artículo es examinar los aspectos psicosociales relacionados con la percepción de la IA. En él se revisan artículos de psicología social relacionados con la IA. Dentro de la psicología social, los estudios sobre la IA se centran sobre todo en el proceso de toma de decisiones y su percepción social. Toman como base las investigaciones sobre este proceso en humanos y por ello se ha utilizado para esta revisión términos relevantes en este ámbito como toma de decisiones, personalidad, percepción social, Deontología, Utilitarismo y dilemas de sacrificio. La revisión se ha realizado utilizando principalmente la base de datos Web of Science.

El artículo se organiza de la siguiente manera: en una primera sección se establece la distinción entre decisiones utilitaristas y deontológicas al tratarse de una distinción esencial en el ámbito de la toma decisiones. La siguiente sección aborda la percepción sobre los coches autónomos por ser el principal tópico de investigación sobre la IA debido a su relación con los dilemas de sacrificio, un tema clásico en psicología social. A continuación se abordan cuestiones sobre la atribución de la responsabilidad por los resultados provocados por la IA y finalmente se incluyen dos breves secciones que enumeran una serie de cuestiones relacionadas con aspectos jurídicos, como es la utilización de la IA en el ámbito judicial y para cometer crímenes.

2. DECISIONES UTILITARISTAS Y DEONTOLÓGICAS

La toma de decisiones sobre lo que es correcto en situaciones en las que puede producirse daño a las personas puede llevarse a cabo adoptando criterios utilitaristas (maximizar el beneficio común agregado), deontológicos (siguiendo un conjunto de reglas o principios sobre lo que es moralmente apropiado) o, de manera similar, aplicando el sentido común (siguiendo diversas normas específicas e intuiciones sobre cómo actuar, combinando principios utilitaristas y deontológicos)[7] .

Los criterios utilitaristas a menudo se rechazan cuando entran en conflicto con otras normas morales. En la investigación sobre como tomamos decisiones acerca de lo que es correcto a menudo se utilizan lo que se denomina como “dilemas de sacrificio” en los que los participantes deben escoger si sacrifican a uno o varios individuos para salvar a un número mayor. Estos dilemas se basan en el clásico dilema del tranvía propuesto por Foot[8] :

En el camino de un tren que avanza sin control se encuentran cinco trabajadores ferroviarios que morirán con seguridad a menos que tú, que estás observándolo todo, hagas algo. Si accionas el cambio de agujas, el tren será desviado a otra vía, donde matará a un único trabajador ferroviario que allí se encuentra.

En esta situación, ¿Accionarías el cambio de agujas?

La mayor parte de la gente rechaza las decisiones utilitaristas (pro-sacrificio) que implican hacer daño a una persona para salvar a otras. Este rechazo se basa en intuiciones y

[7] BARTELS, D. M. y PIZARRO, D.A., “The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas”, en *Cognition*, 121(1), 2011, pp. 154-161.
EVERETT, J. A. G. y KAHANE, G., “Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology”, en *Trends in Cognitive Sciences*, 24(2), 2020, pp. 124-134.
HAIDT, J., “The emotional dog and its rational tail: A social intuitionist approach to moral judgment”, en *Psychological Review*, 108(4), 2001, pp. 814-834.
KAHANE, G., EVERETT, J.A.C., EARP, B.D., FARIAS, M. et al., “'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good”, en *Cognition*, 134, 2015, pp. 193-209.

[8] FOOT, P., “The problem of abortion and the doctrine of the double effect”, en *Oxford Review*, 5, 1967, pp. 5–15.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8917

reacciones emocionales de aversión a provocar daño a los demás[9] o incluso en la preferencia por no actuar antes que realizar una acción controvertida[10] . En cambio, las decisiones utilitaristas soslayan esta aversión intuitiva a causar daño en favor de la decisión que logra una mejor relación coste-beneficio.

Esta reducida aversión a provocar daño se relaciona con variables de personalidad del individuo, en concreto con rasgos psicopáticos. Estos rasgos muestran una menor preocupación por el daño sufrido por otros, independientemente del beneficio obtenido, en lugar de una mayor preocupación por las consecuencias para el beneficio común. Esta indiferencia al daño provocado parece que es una motivación más fuerte que buscar el mayor beneficio común en los estudios de laboratorio[11] . Los rasgos psicopáticos afectan a los razonamientos utilitarios y la toma de decisiones también en dilemas sobre situaciones cotidianas[12] . Sin empatía afectiva, el individuo es indiferente al sufrimiento de otros, incluyendo a su familia y amigos. En este sentido, para las personas con rasgos psicopáticos es más fácil optar por decisiones utilitarias. No porque estén preocupados por maximizar los resultados positivos para otros o reducir perjuicios al máximo, sino porque no experimentan sensaciones aversivas por ejecutar acciones perjudiciales para otras personas. Los individuos que obtienen puntuaciones elevadas en psicopatía y en maquiavelismo muestran una preferencia por las decisiones utilitaristas en dilemas de sacrificio, de modo que no se puede determinar si esta preferencia se debe a un déficit emocional o a una preocupación genuina por maximizar el bienestar de los demás[13] . Para estas personas, es más sencillo adoptar una decisión utilitarista porque no experimentan una respuesta emocional aversiva cuando deben decidir sacrificar a una persona para conseguir un resultado beneficioso[14] . También se ha encontrado una preferencia por criterios utilitaristas en las personas que puntúan alto en sadismo, lo que subraya el papel del placer y la crueldad en el razonamiento moral utilitarista[15] .

Otra cuestión que incide en la utilización de criterios utilitaristas es la familiaridad con la víctima[16] . Cuando la víctima es desconocida, es más frecuente adoptar criterios utilitaristas, lo que es congruente con la relación entre este tipo de decisiones y la indiferencia hacia otros y la frialdad emocional. La cercanía emocional con alguno de los personajes implicados en los

---

[9] HAIDT, J., "The emotional dog and its rational tail: A social intuitionist approach to moral judgment", en *Psychological Review*, 108(4), 2001, pp. 814-834.

[10] GAWRONSKI, B., ARMSTRONG, J., CONWAY, P., FRIESDORF, R., et al., "Consequences, Norms, and Generalized Inaction in Moral Dilemmas: The CNI Model of Moral Decision-Making", en *Journal of Personality and Social Psychology*, 113(3), 2017, pp.343-376.

[11] EVERETT, J. A. G. y KAHANE, G., "Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology", en *Trends in Cognitive Sciences*, 24(2), 2020, pp. 124-134.

[12] TAKAMATSU, R., "Personality correlates and utilitarian judgments in the everyday context: Psychopathic traits and differential effects of empathy, social dominance orientation, and dehumanization beliefs", en *Personality and Individual Differences*, 146, 2019, pp. 1-8.

[13] BARTELS, D. M. y PIZARRO, D.A., "The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas", en *Cognition*, 121(1), 2011, pp. 154-161.

[14] PLETTI, C., LOTTO, L., BUODO, G., y SARLO, M. "It's immoral, but I'd do it! Psychopathy traits affect decision-making in sacrificial dilemmas and in everyday moral situations", en. *British Journal of Psychology*, 108(2), 2017, pp. 351-368.

[15] DINIC, B. M., MILOSAVLJEVIC, M., y MANDARIC, D.J., "Effects of Dark Tetrad traits on utilitarian moral judgement: The role of personal involvement and familiarity with the victim", en *Asian Journal of Social Psychology*, 24(1), 2021, pp. 48-58.

dilemas de sacrificio hace que los participantes muestren un sesgo a favor de las personas que conocen y a tomar decisiones poco éticas[17] .

Aunque la IA no tenga personalidad, se le atribuyen igualmente rasgos de personalidad. Las máquinas controladas por IA se perciben con un elevado nivel de competencia y eficacia y una baja calidez emocional. También se percibe con mayor frecuencia que funcionan con criterios utilitaristas[18] . Esta percepción puede conducir a una aversión a utilizar la IA para decidir sobre dilemas morales y a preferir que otros utilicen máquinas guiadas por IA, pero no utilizarlos para uno mismo (p. e. comprando un coche autónomo).

Otras preferencias en los dilemas de sacrificio se orientan a una mayor aversión a provocar daño antes que a permitir que ocurra de manera pasiva (principio de acción), mayor aversión a provocar daño directamente antes que indirectamente (principio de contacto) y mayor aversión a provocar daño como medio para alcanzar un objetivo antes que a provocarlo como una consecuencia inevitable de ayudar a otros (principio de intención)[19] .

3. COCHES AUTÓNOMOS Y DILEMAS MORALES

Gran parte de las investigaciones sobre la toma de decisiones morales por parte de la IA se centran en dilemas de sacrificio en el contexto de los coches autónomos. A veces los accidentes no pueden evitarse, y en algunos accidentes la IA debe tomar decisiones sobre a quién sacrificar cuando no se puede salvar a todos los humanos implicados. En los dilemas sobre coches autónomos los participantes deben escoger que tipo de decisión es preferible en el caso de un accidente inevitable en el que un coche autónomo tenga que decidir por ejemplo si atropella mortalmente cinco peatones en el carril por el que circula o se desvía de su ruta y mata a un peatón que de otra manera no moriría, en otro carril, o si atropella a un peatón o desvía el coche hacia un obstáculo matando al conductor. En el *Moral Machine Experiment* [20] llevado a cabo con millones de participantes en 233 países, se plantearon escenarios sobre un accidente inevitable variando los siguientes factores en la toma de decisiones: salvar a personas (*vs.* mascotas); salvar más vidas (*vs.* menos vidas); salvar a los pasajeros (*vs.* peatones); salvar a hombres (*vs.* mujeres); salvar a jóvenes (*vs.* personas mayores); salvar a un peatón que cruza correctamente (*vs.* incorrectamente); salvar a personas saludables (*vs.* personas con sobrepeso); salvar a personas con un estatus social alto (*vs.* un estatus bajo); y salvar a ciudadanos respetuosos con la ley (*vs.* criminales). Las mayores diferencias se encontraron a favor de salvar a personas antes que a animales, salvar el mayor número de vidas y salvar preferentemente a personas jóvenes, aunque también se encontraron diferencias notables en contra de los criminales, las personas de bajo estatus, las

[16] DINIC, B. M., MILOSAVLJEVIC, M., y MANDARIC, D.J., "Effects of Dark Tetrad traits on utilitarian moral judgement: The role of personal involvement and familiarity with the victim", en *Asian Journal of Social Psychology*, 24(1), 2021, pp. 48-58.

[17] NAVARICK, D.J., "Question framing and sensitivity to consequences in sacrificial moral dilemmas", en *Journal of Social Psychology*, 161(1), 2021, pp. 25-39.

[18] ZHANG, Z.X., CHEN, Z.S., y XU, L.Y., "Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI", en *Journal of Experimental Social Psychology*, 101, 2022.

[19] BOSTYN, D. H., ROETS, A., y CONWAY, P., "Sensitivity to Moral Principles Predicts Both Deontological and Utilitarian Response Tendencies in Sacrificial Dilemmas", en *Social Psychological and Personality Science*, 2021, pp. 1-10.

[20] AWAD, E., DSOUZA, S., KIM, R., SCHULZ, J. et al., *"The Moral Machine Experiment"*, en *Nature*, 563(7729), 2018, pp. 59-64.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8917

personas con sobrepeso, los hombres y los pasajeros. Este estudio sugiere que las decisiones de los usuarios pueden ser tenidas en cuenta para programar la IA de los coches autónomos. Sin embargo, sus conclusiones han sido criticadas[21] , porque se basa en criterios utilitaristas exclusivamente (como se minimiza el daño en función de si se sacrifica a un protagonista u otro de los escenarios) y porque no es permisible que sean los conductores quienes decidan los criterios de toma de decisiones de un coche autónomo, incluso por encima de consideraciones legales. Una postura ética sobre esta cuestión es que los coches autónomos siempre deben arriesgar a sus ocupantes antes que a personas fuera del vehículo, y confiar en las medidas de seguridad del coche, sobre todo porque que si estuviesen programados de otra manera las decisiones estarían sesgadas hacia el interés particular o el egoísmo de los pasajeros. Sin embargo, si esto fuese así en todos los casos, es posible que este tipo de vehículos no fueran aceptables o populares entre los usuarios. De hecho, existe una relación entre la preferencia por criterios utilitaristas y la confianza en los coches autónomos que utilizan estos criterios. Sólo las personas que tienen actitudes utilitaristas están a favor de que la IA utilice estos criterios. Esta confianza influye en la intención de utilizarlos y permitir que conduzcan sin supervisión[22] . Los coches autónomos se perciben por la mayoría de las personas como menos fiables, morales y más merecedores de culpa que los humanos. Este escepticismo proviene de que se les atribuye capacidad de tomar decisiones, pero no de tomarlas de acuerdo con valores o empatía hacia los humanos[23] . También se ha observado, que existe menos motivación para comprar coches autónomos "utilitaristas" frente a coche "egoístas" que protejan al conductor[24] . De manera similar, se ha encontrado que las personas están de acuerdo con que los coches autónomos sacrifiquen a los pasajeros para salvar a los peatones de acuerdo con criterios utilitaristas, pero ellas no se los comprarían ni querrían que sus familiares se los comprasen[25] .

4. RESPONSABILIDAD Y JUSTICIA DE LAS DECISIONES TOMADAS A TRAVÉS DE INTELIGENCIA ARTIFICIAL

Otra cuestión relevante relacionada con la percepción de la IA es la atribución de responsabilidad y la evaluación de la justicia en la toma de decisiones. Existe un vacío de responsabilidad para atribuir culpa a las máquinas que provocan algún resultado perjudicial[26] . En la medida en que máquinas sin supervisión humana provoquen daños a

---

[21] HARRIS, J., "The Immoral Machine", en *Cambridge Quarterly of Healthcare Ethics*, 29(1), 2020, pp. 71-79.

[22] YOKOI, R. y NAKAYACHI, K., "Trust in Autonomous Cars: Exploring the Role of Shared Moral Values, Reasoning, and Emotion in Safety-Critical Decisions", *Human Factors*, 63(8), 2021, pp. 1465-1484.

[23] YOUNG, A.D., y MONROE, A.E., "Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas", en *Journal of Experimental Social Psychology*, 85, 2019.

[24] LIU, P. y LIU, J.T., "Selfish or Utilitarian Automated Vehicles? Deontological Evaluation and Public Acceptance", en *International Journal of Human-Computer Interaction,* 37(13), 2021, pp. 1231-1242.
MORITA, T. y MANAGI, S., "Autonomous vehicles: Willingness to pay and the social dilemma", en *Transportation Research Part C-Emerging Technologies*, 119, 2020.

[25] BONNEFON, J. F., SHARIFF, A., y RAHWAN, I., "The social dilemma of autonomous vehicles", en *Science*, 352(6293), 2016, pp. 1573-1576.

[26] TIGARD, D.W., "Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible", en *Cambridge Quarterly of Healthcare Ethics*, 30(3), 2021, pp. 435-447.

terceros y no se pueda identificar claramente al responsable (el propietario, el usuario, el programador, etc.), debe cuestionarse el uso de estas máquinas.

Además, la utilización de IA contribuye a restar responsabilidad a los agentes humanos. Hay evidencias de que se atribuye menos culpa a un humano ante un resultado negativo cuando han delegado en una IA que cuando actúa solo[27] . La agencia indirecta (cuando una persona actúa por mediación de otros agentes) reduce los procesos que regulan el comportamiento poco ético. Cuando una persona actúa a través de otros agentes humanos en lugar de hacerlo directamente (e.g. el gerente de una empresa da instrucciones a sus supervisores para que las ejecuten) la distancia psicológica con las consecuencias del comportamiento se reduce debido a la separación tanto en el tiempo como en el espacio de la decisión de ejecutarlo. En el caso de la agencia directa, cuando una persona provoca un daño, pude anticipar que será culpada por ello, perderá reputación y hasta será castigada, incluso por terceros que presencien la acción. Cuando hay una agencia indirecta, la anticipación de las consecuencias morales está comprometida. Las decisiones muestran menos consideración por los receptores de las mismas y se percibe menos responsabilidad y probabilidades de sufrir consecuencias perjudiciales por tomar decisiones poco éticas. Cuando se ejerce una agencia indirecta a través de la IA, las personas están más dispuestas a engañar, son responsabilizadas en menor medida y además anticipan que esto será así[28] .

Los sistemas de IA no tienen motivaciones o consciencia, por lo que no puede en ningún caso atribuírseles responsabilidad y tampoco tienen personalidad legal. Además si se le atribuyese responsabilidad a una IA, se estaría restando responsabilidad a los agentes humanos que han promovido, diseñado o usado esa IA[29] . En definitiva, la IA no puede ser moralmente responsable porque no es un agente moral, sino un *proxy* moral del humano responsable[30] . Por otro lado, los sistemas de IA no pueden ser castigados o recompensados, por lo que no tiene sentido considerarlos responsables de los resultados de sus acciones. Los castigos y recompensas tienen el objetivo de conseguir un cambio en el comportamiento del culpable de una acción, lo que no tiene sentido en el caso de una máquina que puede ser reprogramada si provoca un daño[31] . La responsabilidad de las acciones de una IA debe pues atribuirse a humanos. En este sentido, los humanos serían los "autores intelectuales" de las acciones de una IA. Por otro lado, puede atribuirse responsabilidad a la persona que dirige o es propietaria de un sistema de IA, de la misma manera que el propietario de una empresa es responsable por las acciones de esta aunque no las ordene directamente. Los fabricantes, programadores y usuarios de la IA son los posibles candidatos. Los sistemas de IA deberían estar diseñados para rechazar órdenes ilegales, aunque puede darse el caso de que los desarrolladores conscientemente diseñen la IA para cometer actos moralmente reprochables siendo en este caso claramente culpables de sus acciones. En los casos en

[27] FEIER, T., GOGOLL, J., y UHL, M., "Hiding Behind Machines: Artificial Agents May Help to Evade Punishment", en *Science and Engineering Ethics*, 28(2), Article 19, 2022.

[28] GRATCH, J. y FAST, N.J., "The power to harm: AI assistants pave the way to unethical behavior", en *Current Opinion in Psychology*, 47, 2022.

[29] KING, T.C., AGGARWAL, N., TADDEO, M. y FLORIDI, L., "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions". *Science and Engineering Ethics*, 26(1), 2020, pp. 89-120.

[30] GOGOLL, J. y MULLER, J.F., "Autonomous Cars: In Favor of a Mandatory Ethics Setting", en *Science and Engineering Ethics*, 23(3), 2017, pp. 681-700.

[31] TIGARD, D.W., "Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible", en *Cambridge Quarterly of Healthcare Ethics*, 30(3), 2021, pp. 435-447.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8917

los que los perjuicios provocados por una IA sean una consecuencia natural o probable de su funcionamiento y se acepta este riesgo negligentemente, los desarrolladores serían indudablemente responsables. Tanto la atribución de responsabilidad como la supervisión del funcionamiento de la IA son desafíos importantes que presentan enormes dificultades[32] . Entre otras cuestiones, la IA debe implementar principios que preserven el bienestar, la dignidad, la privacidad, la autonomía, la justicia y que sean transparentes, manteniendo la responsabilidad humana en la toma de decisiones por parte de la IA[33] . Además en algunas circunstancias cabe atribuirse mayor responsabilidad a las acciones de una IA que a las de un humano. En situaciones similares al dilema del tranvía, los humanos puede que deban actuar un una fracción de segundo (arrollar a un peatón o sacrificar al conductor chocando contra un obstáculo) y actúen por instinto sin oportunidad de deliberar cuál es la acción moralmente más correcta. En estos casos, no le asignaríamos responsabilidad a un humano incluso si su decisión no fuese óptima desde el punto de vista moral. Sin embargo, la IA puede manejar enormes cantidades de información y actuar en una fracción de segundo, y lo hará de acuerdo con su algoritmo programado previamente, con lo que existe una responsabilidad moral (del regulador, programador o usuario, en función de quién decida el algoritmo utilizado). Por este motivo, debe existir un criterio previo sobre cómo debe decidir el algoritmo en una situación límite que provoque un dilema moral, que no quede al albedrío del humano responsable. Si el algoritmo moral dependiese de decisiones individuales y pudiese ser más altruista (actuar de acuerdo con el bien común) o más egoísta (proteger los intereses propios), los criterios egoístas dominarían a lo largo del tiempo, como ocurre en el dilema del prisionero, pues las personas no estarían dispuestas a sacrificarse por otras personas que no tuviesen un comportamiento recíproco[34] . Por supuesto, si los algoritmos son decididos por un regulador, la responsabilidad de las decisiones morales de la IA corresponde al regulador.

Como nota adicional, además de establecer quién es el responsable de las acciones de la IA cuando esta provoca daños, para poder generar confianza en el público la IA debe tomar decisiones que se perciban como justas. Por ejemplo, la inclusión de variables demográficas (sexo, edad, nacionalidad) como input para la toma de decisiones de un algoritmo se percibe como injusta. Aunque la IA puede aumentar esta confianza porque no están sujeta a los sesgos humanos provocados por intereses, emociones, fatiga o falta de atención, también pueden producir resultados arbitrarios e injustos. El concepto de justicia algorítmica implica que las decisiones de un algoritmo no deben producir consecuencias injustas, discriminatorias o dispares en situaciones similares. Se han propuesto cuatro dimensiones en la justicia algorítmica que deben cumplirse para generar confianza en la utilización de la IA: *justicia distributiva* (la distribución no discriminatoria de recursos basada en la igualdad, equidad o necesidad); *justicia procedimental* (reversibilidad y consistencia en el mecanismo de toma de decisiones); *justicia informativa* (transparencia en las decisiones); y *justicia interpersonal* (la no utilización de datos protegidos y el respeto a la privacidad)[35] .

---

[32] KING, T.C., AGGARWAL, N., TADDEO, M. y FLORIDI, L., "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions". *Science and Engineering Ethics*, 26(1), 2020, pp. 89-120.

[33] DE AGREDA, A. G. "Ethics of autonomous weapons systems and its applicability to any AI systems", en *Telecommunications Policy*, 44(6), 2020.

[34] GOGOLL, J. y MULLER, J.F., "Autonomous Cars: In Favor of a Mandatory Ethics Setting", en *Science and Engineering Ethics*, 23(3), 2017, pp. 681-700.

[35] STARKE, C., BALEIS, J., KELLER, B. y MARCINKOWSKI, F., "Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature", en *Big Data & Society*, 9(2), 2022.

5. INTELIGENCIA ARTIFICIAL EN EL ÁMBITO JUDICIAL

Las anteriores cuestiones examinadas tienen su aplicación en el ámbito legal en lo que respecta a la elaboración de criterios y atribución de responsabilidad, pero hay aspectos específicos en los que la IA ya tiene un efecto directo y cotidiano sobre el ámbito legal. Esto es así porque uso de la IA también se ha extendido al ámbito judicial, donde a menudo se desconoce su alcance[36] . A través de algoritmos de IA es posible asistir decisiones judiciales con un elevado nivel de precisión. Xu, analizando principalmente su aplicación al sistema judicial chino[37] , plantea que aporta tanto ventajas como inconvenientes. Entre las ventajas se encuentran una mayor eficacia, especialmente con el incremento exponencial de litigios. A través de potentes algoritmos, la IA puede utilizarse entre otras cosas para evaluar pruebas, generar documentación y transcribir las vistas a través de reconocimiento de voz. Se calcula que en China esto ha supuesto una reducción de entre 20% y 30 % de la duración de los juicios y hasta una reducción del 50% en los juicios más complejos. El reconocimiento de textos e imágenes y su integración asiste en la generación de documentación y permite evitar en gran medida que la ley sea aplicada de manera inconsistente en diferentes casos. Todo ello permite que se resuelvan más casos con menos personal. La IA también permite acumular más experiencia de la que sería capaz de alcanzar un juez humano a la hora de revisar los hechos y compararlos con casos previos incorporados a una base de datos, que en el ámbito chino alcanza los 100 millones de casos. También contribuye a que casos similares se resuelvan de forma parecida. Otra ventaja es una mayor objetividad, pues permite eliminar las inconsistencias y falta de neutralidad que pueden producirse en los jueces humanos debido a intereses personales, sesgos, burnout o corrupción.

Sin embargo, también se plantean una serie de limitaciones. Actualmente la IA sólo puede asistir al juez humano y no reemplazarlo. Además, su aplicación está limitada a casos estándar que reúnan una serie de requisitos técnicos. La IA puede ser eficaz, pero la calidad de sus decisiones para valorar aspectos muy específicos de un caso no puede compararse con la de un juez. Podría decirse que la IA no sabe "leer entre líneas". Además la IA incorporará las limitaciones y sesgos con los que haya sido programada. Estas limitaciones pueden corresponder a errores de programación, incapacidad para valorar aspectos sutiles o sesgos conscientes o inconscientes introducidos en la programación por los desarrolladores, administradores o grupos de presión. En lo que respecta a los sistemas dinámicos de IA que incorporan sistemas de *machine learning*, su mutabilidad introduce un mayor riesgo de provocar perjuicios al estar más expuesto a la generación de sesgos. Por otro lado, aunque la IA en algunos casos ha alcanzado una precisión por encima de la de los jueces humanos, las expectativas sobre su funcionamiento provocan que sus errores sean menos tolerables. Además, como los algoritmos suelen ser opacos y no hay una compresión clara de cómo alcanza las decisiones, es muy difícil corregir los errores que se produzcan, lo que en la práctica lleva a que no se corrijan y sea mucho más difícil apelar sentencias asistidas por IA. Asimismo, la experiencia con la que cuenta la IA proviene de la acumulación de datos de otras sentencias, pero la integración de esta experiencia no cuenta con la intuición, la percepción de las necesidades de las partes y el "saber hacer" que permite la experiencia de los jueces humanos. La codificación de los algoritmos de IA difícilmente podrá incorporar todos los

[36] DE SILES, E.L. "AI, on the Law of the Elephant: Toward Understanding Artificial Intelligence", en *Buffalo Law Review*, 69(5), 2021, pp.1389-1469.

[37] XU, Z.C., "Human Judges in the Era of Artificial Intelligence: Challenges and Opportunities", en *Applied Artificial Intelligence*, 36(1), 2022.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8917

matices de la situación que puede captar un juez humano. La objetividad de la IA también es cuestionable porque su código de programación puede introducir sesgos que incluso pueden incrementarse con el tiempo a través de los mecanismos de*machine learning* que modifican los algoritmos sobre la base de los inputs recibidos. La Inteligencia emocional tampoco puede tener en cuenta adecuadamente aspectos emocionales o necesidades que pueden afectar a las partes.

Acudiendo a un ejemplo muy citado en la literatura, el caso Loomis[38] , podemos atisbar varias controversias sobre el uso de la IA. En este caso el acusado alegó que la utilización del sistema COMPAS que utiliza algoritmos de IA para determinar el riesgo de reincidencia que se había utilizado para dictar sentencia violaba las garantías procesales. El sistema COMPAS determina la peligrosidad de los acusados basándose en grupos de personas con características similares y los algoritmos que usa son un secreto comercial, por lo que el proceso de toma de decisiones es opaco. La sentencia no fue revisada, pero a raíz de este caso la corte suprema de Wisconsin requirió incluir advertencias para los jueces sobre la opacidad de su metodología y que se basaba en predicciones grupales, no sobre un individuo concreto. Sin embargo, es muy difícil para un juez abstraerse del dato sobre peligrosidad proporcionado por este sistema y tener en cuenta que puede ser impreciso y estar cuestionado. El dato sobre probabilidad de reincidencia actuará como un estándar de comparación que sesgará las decisiones del juez a través del sesgo heurístico de anclaje. Además, tal y como ocurrió en este mismo caso, las posibilidades de éxito en una apelación se verán reducidas en la medida que el juez que revise el caso tenga reticencias a cuestionar un sistema de IA cuyo funcionamiento desconoce pero que se plantea como una herramienta compleja y eficaz para emitir sentencias más objetivas. Este sistema también ha sido criticado porque puede discriminar a grupos de personas. Aunque no maneja información acerca del origen étnico de los acusados, la información que se recoge (e. g. lugar de residencia) pude funcionar como un proxy del grupo étnico. Estas críticas han sido cuestionadas a su vez por los desarrolladores del programa por sesgadas y por el uso de información parcial para evaluar la eficacia de COMPAS. En cualquier caso, la controversia sobre este caso proviene de la falta de justicia procedimental y justicia informativa planteadas en el apartado anterior[39] .

6. INTELIGENCIA ARTIFICIAL Y CRIMEN

Finalmente, podemos examinar otro ámbito en el que la IA tiene implicaciones inmediatas sobre el ámbito legal. Existe un notable riesgo de la IA por su capacidad para ser orientada a cometer o facilitar delitos. Por ejemplo, en el ámbito de las redes sociales, puede utilizarse para adaptar mensajes de *phishing* para que sean más convincentes para los receptores y estos sean más proclives a proporcionar la información privada buscada por los delincuentes para cometer un fraude[40] .

[38] CRIMINAL LAW SENTENCING GUIDELINES, "Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing. - 'State v. Loomis', 881 N.W.2d 749 (Wis. 2016).", en *Harvard Law Review*, 130(5), 2017, pp. 1530–1537.

[39] STARKE, C., BALEIS, J., KELLER, B. y MARCINKOWSKI, F., "Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature", en *Big Data & Society*, 9(2), 2022.

[40] KING, T.C., AGGARWAL, N., TADDEO, M. y FLORIDI, L., "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions". *Science and Engineering Ethics*, 26(1), 2020, pp. 89-120.

King *et al.* [41] revisan algunos contextos en los que la IA es instrumentalizada para cometer delitos son la manipulación de mercados (e. g. para influir en el precio de acciones de manera artificial a través de algoritmos de alteración de precios); el uso de vehículos dirigidos por IA en el contrabando de sustancias ilegales (drones aéreos o vehículos subacuáticos); la distribución de drogas a través de Internet con anuncios adaptados al perfil de los consumidores; el acoso a las personas a través de *bots* en las redes sociales de Internet que difundan mensajes de odio o que dañen la reputación (e. g. redistribuyendo o marcando *like* en mensajes negativos, o sesgando encuestas en contra de una persona o institución); manipulando *bots* ajenos influyendo en ellos para que modifiquen su comportamiento; y creando contenido falso, como videos sintéticos donde se reemplaza la cara de otra persona (*deep fakes*, a menudo videos pornográficos). Otro ámbito controvertido en el que la IA puede usarse es en los interrogatorios y tortura. La IA puede utilizarse para detectar con mayor eficacia el engaño, pero el interrogado puede percibir que si le interroga una IA no mostrará empatía ni compasión, lo que puede incrementar su angustia. Los responsables de la IA pueden aislarse emocionalmente de los actos de tortura, haciendo que así sea más fácil ejecutarla. La IA también puede fomentar los delitos sexuales, permitiendo al usuario emular delitos como el abuso sexual infantil. El uso de IAs antropomórficas para este fin puede desensibilizar a los potenciales agresores o incluso aumentar su deseo de agredir sexualmente a víctimas humanas. Sin duda, estas capacidades de la IA para facilitar la comisión de delitos suponen un enorme desafío para el ámbito judicial.

7. CONCLUSIONES

La presente revisión sobre la IA en relación a las decisiones morales pone en evidencia una serie de cuestiones que deben tenerse en cuenta para regular su uso. Para que la IA sea aceptada en ámbitos donde se toman decisiones morales debe generar confianza y por este motivo sus decisiones deben percibirse como justas en diferentes dimensiones, proteger los intereses legítimos de los usuarios y debe poder identificarse al responsable de los daños potenciales provocados por la IA sin que la propia IA sirva para enmascarar esta responsabilidad. Es fundamental promover esta confianza toda vez que la IA genera desconfianza y se percibe acertadamente como fría y carente de empatía.

Los criterios utilitaristas orientados a la consecución del mayor beneficio (o mínimo perjuicio) agregado, pueden plantearse como criterios racionales para programar algoritmos de IA. Sin embargo, la mayoría de las personas se guían por criterios deontológicos[42] , y aunque existen personas utilitaristas preocupadas por el bien común, la utilización de estos criterios se confunde con la toma de decisiones carente de empatía propia de las personas con rasgos psicopáticos, maquiavélicos o sádicos. Cabe anticipar que los algoritmos de IA que empleen criterios utilitaristas para tomar decisiones en contextos morales serán percibidos como carentes de empatía y serán rechazados en mayor medida por el público. En definitiva, si las instituciones públicas y privadas no son capaces de diseñar sistemas de IA que se perciban como justos, lo más probable es que las personas se sientan alienadas y rechacen o circuventen la utilización de estos sistemas[43] .

[41] *Ibidem*

[42] YOUNG, A.D., y MONROE, A.E., "Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas", en *Journal of Experimental Social Psychology*, 85, 2019.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8917

El objetivo de este estudio ha sido exponer plantear cuestiones psicosociales relevantes para elaborar criterios sobre el uso de la IA. Resulta evidente que estamos ante un desafío incipiente en al menos tres vertientes. Por un lado, en el desarrollo de criterios para la utilización de la IA de una manera justa, ética y aceptable socialmente. En segundo lugar en el ámbito de su aplicación efectiva y justa en el ámbito judicial. Por último, en la prevención de su utilización ilegal y en el desarrollo de herramientas adecuadas para combatir este tipo de usos.

Bibliografía

AWAD, E., DSOUZA, S., KIM, R., SCHULZ, J. et al., *"The Moral Machine Experiment"*, en *Nature*, 563(7729), 2018, pp. 59-64. https://doi.org/10.1038/s41586-018-0637-6.

BARTELS, D. M. y PIZARRO, D.A., "The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas", en *Cognition,* 121(1), 2011, pp. 154-161. https://doi.org/10.1016/j.cognition.2011.05.010.

BONNEFON, J. F., SHARIFF, A., y RAHWAN, I., "The social dilemma of autonomous vehicles", en *Science*, 352(6293), 2016, pp. 1573-1576. https://doi.org/10.1126/science.aaf2654.

BOSTYN, D. H., ROETS, A., y CONWAY, P., "Sensitivity to Moral Principles Predicts Both Deontological and Utilitarian Response Tendencies in Sacrificial Dilemmas", en *Social Psychological and Personality Science*, 2021, pp. 1-10. https://doi.org/10.1177/19485506211027031.

CRIMINAL LAW SENTENCING GUIDELINES, "Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing. - 'State v. Loomis', 881 N.W.2d 749 (Wis. 2016).", en *Harvard Law Review*, 130(5), 2017, pp. 1530–1537.

DE AGREDA, A. G. "Ethics of autonomous weapons systems and its applicability to any AI systems", en *Telecommunications Policy*, 44(6), 2020. https://doi.org/10.1016/j.telpol.2020.101953.

DE SILES, E.L. "AI, on the Law of the Elephant: Toward Understanding Artificial Intelligence", en *Buffalo Law Review*, 69(5), 2021, pp.1389-1469.

DINIC, B. M., MILOSAVLJEVIC, M., y MANDARIC, D.J., "Effects of Dark Tetrad traits on utilitarian moral judgement: The role of personal involvement and familiarity with the victim", en *Asian Journal of Social Psychology*, 24(1), 2021, pp. 48-58. https://doi.org/10.1111/ajsp.12422.

ELLEUCH, M. A., BEN HASSENA, A., ABDELHEDI, M. y PINTO, F.S., "Real-time prediction of COVID-19 patients health situations using Artificial Neural Networks and Fuzzy Interval Mathematical modeling", en *Applied Soft Computing*, 110, 2021. https://doi.org/10.1016/j.asoc.2021.107643.

EVERETT, J. A. G. y KAHANE, G., "Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology", en *Trends in Cognitive Sciences*, 24(2), 2020, pp. 124-134. https://doi.org/10.1016/j.tics.2019.11.012.

[43] STARKE, C., BALEIS, J., KELLER, B. y MARCINKOWSKI, F., "Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature", en *Big Data & Society*, 9(2), 2022.

FEIER, T., GOGOLL, J., y UHL, M., "Hiding Behind Machines: Artificial Agents May Help to Evade Punishment", en *Science and Engineering Ethics*, 28(2), Article 19, 2022. https://doi.org/10.1007/s11948-022-00372-7.

FOOT, P., "The problem of abortion and the doctrine of the double effect", en *Oxford Review*, 5, 1967, pp. 5–15.

GAWRONSKI, B., ARMSTRONG, J., CONWAY, P., FRIESDORF, R., et al., "Consequences, Norms, and Generalized Inaction in Moral Dilemmas: The CNI Model of Moral Decision-Making", en *Journal of Personality and Social Psychology*, 113(3), 2017, pp.343-376. https://doi.org/10.1037/pspa0000086.

GOGOLL, J. y MULLER, J.F., "Autonomous Cars: In Favor of a Mandatory Ethics Setting", en *Science and Engineering Ethics*, 23(3), 2017, pp. 681-700. https://doi.org/10.1007/s11948-016-9806-x.

GRATCH, J. y FAST, N.J., "The power to harm: AI assistants pave the way to unethical behavior", en *Current Opinion in Psychology*, 47, 2022. https://doi.org/10.1016/j.copsyc.2022.101382.

HAIDT, J., "The emotional dog and its rational tail: A social intuitionist approach to moral judgment", en *Psychological Review*, 108(4), 2001, pp. 814-834. https://doi.org/10.1037//0033-295x.108.4.814.

HARRIS, J., "The Immoral Machine", en *Cambridge Quarterly of Healthcare Ethics*, 29(1), 2020, pp. 71-79. https://doi.org/10.1017/s096318011900080x.

KAHANE, G., EVERETT, J.A.C., EARP, B.D., FARIAS, M. et al., "'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good", en *Cognition*, 134, 2015, pp. 193-209. https://doi.org/10.1016/j.cognition.2014.10.005.

KING, T.C., AGGARWAL, N., TADDEO, M. y FLORIDI, L., "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions". *Science and Engineering Ethics*, 26(1), 2020, pp. 89-120. https://doi.org/10.1007/s11948-018-00081-0.

LIU, P. y LIU, J.T., "Selfish or Utilitarian Automated Vehicles? Deontological Evaluation and Public Acceptance", en *International Journal of Human-Computer Interaction*, 37(13), 2021, pp. 1231-1242. https://doi.org/10.1080/10447318.2021.1876357.

MORITA, T. y MANAGI, S., "Autonomous vehicles: Willingness to pay and the social dilemma", en *Transportation Research Part C-Emerging Technologies*, 119, 2020. https://doi.org/10.1016/j.tre.2020.102748.

NAVARICK, D.J., "Question framing and sensitivity to consequences in sacrificial moral dilemmas", en *Journal of Social Psychology*, 161(1), 2021, pp. 25-39. https://doi.org/10.1080/00224545.2020.1749019.

PLETTI, C., LOTTO, L., BUODO, G., y SARLO, M. "It's immoral, but I'd do it! Psychopathy traits affect decision-making in sacrificial dilemmas and in everyday moral situations", en. *British Journal of Psychology*, 108(2), 2017, pp. 351-368. https://doi.org/10.1111/bjop.12205.

STARKE, C., BALEIS, J., KELLER, B. y MARCINKOWSKI, F., "Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature", en *Big Data & Society*, 9(2), 2022. https://doi.org/10.1177/20539517221115189.

TAKAMATSU, R., "Personality correlates and utilitarian judgments in the everyday context: Psychopathic traits and differential effects of empathy, social dominance orientation, and dehumanization beliefs", en *Personality and Individual Differences*, 146, 2019, pp. 1-8. https://doi.org/10.1016/j.paid.2019.03.029.

TIGARD, D.W., "Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible", en *Cambridge Quarterly of Healthcare Ethics*, 30(3), 2021, pp. 435-447. https://doi.org/10.1017/s0963180120000985.

XU, Z.C., "Human Judges in the Era of Artificial Intelligence: Challenges and Opportunities", en *Applied Artificial Intelligence*, 36(1), 2022. https://doi.org/10.1080/08839514.2021.2013652.

YOKOI, R. y NAKAYACHI, K., "Trust in Autonomous Cars: Exploring the Role of Shared Moral Values, Reasoning, and Emotion in Safety-Critical Decisions", *Human Factors*, 63(8), 2021, pp. 1465-1484. https://doi.org/10.1177/0018720820933041.

YOUNG, A.D., y MONROE, A.E., "Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas", en *Journal of Experimental Social Psychology*, 85, 2019. https://doi.org/10.1016/j.jesp.2019.103870.

ZHANG, Z.X., CHEN, Z.S., y XU, L.Y., "Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI", en *Journal of Experimental Social Psychology*, 101, 2022. https://doi.org/10.1016/j.jesp.2022.104327.