ARBOR Ciencia, Pensamiento y Cultura
Vol. 197-800, abril-junio 2021, a601 | ISSN-L: 0210-1963
https://doi.org/10.3989/arbor.2021.800003

ÉTICAS FALIBLES PARA MÁQUINAS (IN)FALIBLES.

FALLIBLE ETHICS FOR (UN) FALLIBLE MACHINES.

Jordi Vallverdú
ICREA Acadèmia – Departamento de Filosofía
Universitat Autònoma de Barcelona
ORCID: 0000-0001-9975-7780
jordi.vallverdu@uab.cat

Sarah Boix
Departamento de Antropología
Universitat Autònoma de Barcelona
ORCID: 0000-0003-2449-794X
Sarah.Boix@uab.cat

Cómo citar este artículo/Citation: Vallverdú, Jordi; Boix, Sarah (2021). Éticas falibles para máquinas (in)falibles. Arbor, 197(800): https://doi.org/10.3989/arbor.2021.800003

Recibido: 21 febrero 2021. Aceptado: 15 mayo 2021.
Publicado:

RESUMEN: los códigos éticos humanos no son coherentes en su diseño ni tampoco tienen una distribución universal. Por lo tanto, la imposible universalidad ni completitud de los sistemas éticos para la inteligencia artificial (IA) es algo evidente que tan sólo fue apuntado someramente por el estudio reciente del MIT (Moral Machine). Al tener toda ética un matiz cultural inequívoco, y también incluir grados de interpretación en sus principios (como el derecho universal a la vivienda, evidente si bien ninguna institución vela por su cumplimiento, algo que una máquina no entendería). Al mismo tiempo se produce un problema formal: un sistema de IA no siempre cuenta con suficientes datos ni tiempo óptimo para procesar una situación, por lo que un factor de azar ético debe tenerse en cuenta en el diseño de tales sistemas.

PALABRAS CLAVE: ética, inteligência artificial (IA), moral, sesgo, falibilidad.

ABSTRACT: Human ethical codes are not consistent in their design, nor are they universally distributed. Therefore, the impossible universality or completeness of ethical systems for artificial intelligence (AI) is something evident that was only briefly pointed out by the recent MIT (Moral Machine) study. As all ethics have an unequivocal cultural nuance and include degrees of interpretation in its principles (such as the universal right to housing, evident although no institution ensures compliance, something that a machine would not understand). At the same time there is a formal problem: an AI system does not always have enough data or optimal time to process a situation, so a factor of ethical chance must be considered in the design of such systems.

KEYWORDS: ethics, artificial intelligence (AI), moral, bias, fallibility.

1. ¿EXISTEN LOS UNIVERSALES HUMANOS, ESPECÍFICAMENTE LOS ÉTICOS?

Ante el reto de programar éticamente nuestras máquinas inteligentes, nos enfrentamos ante un problema fundamental previo: ¿existe un único y coherente sistema de ideas éticas que permitan su implementación computacional de forma sistemática, completa y eficiente? Ello nos remitiría a la noción de universal ético y, por ende, a la de universal humano. Lo cierto es que cuando nos enfrentamos a la realidad que conocemos sobre el ser humano, tales universales no aparecen por ningún lugar, y mucho menos los de tipo ético. Debemos remarcar que por *universal* no nos referimos a una categorización que pueda ser identificada en cualquier grupo humano (como la cooperación, la gramática, el pensamiento mágico, la agresividad, la cura, o la mediación, entre muchas otras), sino más bien a la existencia de un contenido homogéneo y compartido (innato) sobre cada una de tales categorizaciones. Puesto que es obvio que los seres humanos comparten de forma universal diversos elementos funcionales, bajo múltiples niveles (Brown, 2004, 2017; Huang y Jaszczolt, 2018).

Por todo lo expuesto, volvamos a la pregunta importante: ¿podemos identificar universales en el comportamiento de los humanos que nos aplanaran el camino hacia una ética universal del comportamiento innato? La respuesta es clara y definitiva: NO. La totalidad de etnografías realizadas por investigaciones antropológicas a lo largo del mundo abogan tras su meta-estudio por un relativismo moral (Klenk, 2019). Al mismo tiempo, ello no impide que existan modos universales de gestionar lo moral en la mayor parte de culturas. A partir de la necesaria cooperación para el aumento de la complejidad social, la clave de la evolución cultural humana, emergen similares reglas que podemos definir como: valores familiares, lealtad de grupo, reciprocidad, respeto, equidad, derechos de propiedad, y valentía (Curry, Mullins y Whitehouse, 2019). Con todo, cómo se desarrollan estos universales cooperativos es lo que marcará enormes diferencias entre unos y otros tipos de sociedades. Por ejemplo, una sociedad esclavista cristiana, como los Estados Unidos a mediados del siglo XIX, puede aplicar tales normas, pero sólo a lo que considera sus ciudadanos reales (sic), es decir, los WASP (White Anglo-Saxon Protestant).

Con todo, es obvio que tanto la estructura corporal y sus usos sociales desarrollados al amparo de miradas culturales constriñen el funcionamiento cognitivo (Buchtel y Norenzayan, 2008; Nisbett, 2003; Norenzayan y Nisbett, 2000), dando lugar a una retroalimentación entre genes y cultura (Laland, Odling-Smee y Myles, 2010). Y si bien es cierto que hay una universalidad básica de la reacción ante ciertos fenómenos, como los colores (Jonauskaite *et al.*, 2020), los determinantes lingüístico-culturales son los que acaban por condicionar la forma de comprender tales fenómenos. Algunas de estas interacciones cultural-cognitivas han provocado visiones aparentemente absolutas sobre lo real, como por ejemplo ha sucedido con el pensamiento mágico. Existen razones de ecología cognitiva que justifican la correlación entre el tipo de divinidades con el desarrollo de estructuras sociales complejas (Botero *et al.*, 2014; Enke, 2019; Lang *et al.*, 2019). Es decir, desde una perspectiva operativa, los seres humanos parten de mecanismos similares para defender ideas divergentes, tales como la identidad y supremacía de grupo propio o las normas básicas a seguir (alimentarias, morales, de estrategia social …).

2. ¿EXISTEN AXIOMAS DEL PENSAMIENTO ÉTICO?

Una vez demostrada la no existencia de universales humanos en relación a la acción (Awad *et al.*, 2020), podemos plantearnos un segundo nivel de análisis. Si lo natural no puede dar soporte a teorías éticas universalistas… ¿Lo podría hacer la razón? En trabajos anteriores, uno de los autores del presente texto ha explorado bajo prismas relacionados tales ideas sobre la variabilidad ético-normativa (Vallverdú, 2007a, 2009 y 2019), si bien en este caso se plantea una colaboración estrecha entre los ámbitos de la filosofía cognitiva, la ética y la antropología.

Numerosos han sido los esfuerzos por intentar identificar los principios absolutos que deberían poder guiar la acción, si bien la mayor parte de tales acciones han sido llevadas a cabo bajo el paraguas de dogmas religiosos que se daban por ciertos. Sin embargo, no se ha cejado en la identificación de principios supuestamente universales como la denominada regla de oro (golden rule). Para empezar, tal regla puede tanto encontrarse en su versión positiva *(haz a los demás…)* como negativa *(no hagas a los demás…)*; además, la interpretación de *hacer* o *no hacer* debe enmarcarse dentro de un discurso de tipo religioso y cultural que incluso comparando la misma versión remite a significados distintos (Allinson, 1992; Mou, 2004; Rembert, 1983).

Incluso desde un punto de vista propio de una única creencia, pongamos por caso la ética cristiana, ve-

ARBOR Vol. 197-800, abril-junio 2021, a601 | ISSN-L: 0210-1963 https://doi.org/10.3989/arbor.2021.800003

remos que en función de la secta y el uso de cierta edición e interpretación de los textos primarios se producen divergencias notables que muestran falta de coherencia entre los integrantes de tal religión. Algo que incluso filósofos medievales ya pusieron sobre la mesa al analizar las divergencias presentes en la Biblia, como ejemplificó Pedro Abelardo con su *sic et non* (Jakubecki, 2012). Si sumamos además el hecho de la coexistencia en las sociedades modernas de numerosas comunidades con valores supranaturales divergentes, tal coordinación y acuerdo en lo ético y moral resulta más que improbable, incluso si remitimos a la noción de *ética mínima* (Cortina Orts, 1986). Si bien es cierto que en entornos extremos no hay lugar para el debate, como en posibles comunidades fuera de la superficie del planeta (Maruyama, 1976), en las sociedades contemporáneas esta diversidad es seguramente el elemento más complejo de gestionar a nivel ético-moral.

Por ello, los desacuerdos sobre el valor de las acciones son evidentes en contextos de mixtura o confrontación. Pensemos por ejemplo en las divergentes nociones de *colectivo, deber* y *responsabilidad* esgrimidas por los jueces y abogados de diferentes nacionalidades que tomaron parte del Juicio de Tokyo, tras el fin de la segunda Guerra Mundial (Ushimura, 2003). Si los propios agentes participantes en un proceso divergen profundamente en los valores asignados a conceptos éticos básicos, entonces no podemos esperar que la ética pueda reducirse a un razonamiento formal evidente. El pensamiento ético es, por lo tanto, no formalizable en un sentido universalmente aceptado, deviniendo no consistente. En sistemas multiculturales no cabe más que intentar una aproximación estadística (Vallverdú, 2009) al pensamiento ético. Además, se tiene que tener en cuenta que los agentes propios participantes en tales debates han ido aumentando con el paso del tiempo: desde el punto clave de la declaración universal de los derechos del hombre (que no del ser humano), le siguió el de las mujeres, los animales, el clima y ahora se debate el de los robots o inteligencias artificiales (Pagallo, 2018). Consecuentemente estamos en entornos variables donde no puede darse una justicia algorítmica en la que normas claras respaldadas de forma generalizada dan lugar a evaluaciones secuencialmente válidas, sin lugar para la duda o resultados inesperados. La interpretabilidad y la ética, serían en este sentido, mutuamente excluyentes. Pero lo cierto es que la realidad nos arroja otro escenario: cuando entendemos cómo las personas a lo largo del mundo emiten argumentos totalmente divergentes en favor de generar normas de comportamiento para las máquinas inteligentes, es cuando se pone sobre la mesa de forma empírica esta evidencia conceptual otrora demostrada. El caso más claro es el de la Moral Machine, experimento realizado por el MIT (Awad *et al.*, 2018). Era un experimento mental simple que arrojó muchos datos a partir de las respuestas de millones de personas conectadas en red a lo largo del globo. El experimento consistía en plantear escenarios hipotéticos en los que una persona tenía que decidir los posibles cursos de acción de un coche autónomo enfrentado a problemas en el tráfico: aparición de obstáculos, de menores, ancianos, animales… Lo impresionante fueron los resultados, puesto que se mostró que en función de la ubicación geográfica de los individuos que respondían, se generaban patrones homogéneos entre ciertos grupos culturales, pero estos diferían en gran manera con los de otros situados en otras latitudes[1].

3. AGENTES RACIONALES Y AGENTES MORALES

Por si las serias diferencias culturales en la percepción de los elementos claves del pensamiento ético no fueran suficiente problema, topamos ante un escollo insalvable del pensamiento ético: los agentes implicados en los mismos no son realmente agentes racionales. En realidad, los humanos pensamos y actuamos heurísticamente de formas muy creativas y divergentes, y sesgadas (Caviola *et al.*, 2014; Marshall *et al.*, 2013), pero nunca siguiendo patrones estrictamente formales (Kahneman, 2011). Es lo que uno de los autores del presente trabajo ha desarrollado como *cognición mixta* (*blended cognition,* Vallverdú y Müller, 2019). Ello nos aboca a la existencia de agentes que toman decisiones optimizando todo tipo de heurísticas y al mismo tiempo dando por evidentes cosas discutibles o incluso entrando en contradicción manifiesta con sus propias ideas básicas: nazis kantianos, cristianos esclavistas… Además, el papel de la percepción sobre la justicia en las transacciones humanas determina el comportamiento de los propios agentes sociales, algo evidente en la disparidad en las acciones de las poblaciones que reaccionan al test del ultimátum (Nowak, Page y Sigmund, 2000; Sanfey *et al.*, 2003; Thaler, 1988).

En la base de este problema subyace una estructura cognitiva anclada en lo emocional, como nos

1 Nota de los editores: para una reflexión complementaria sobre este experimento véase el texto de Guevara en el presente volumen.

han mostrado los numerosos estudios de neuroética (Clausen y Levy, 2015), y el ya famoso problema de la vagoneta (Navarrete *et al.*, 2012). Los agentes sociales, tomamos decisiones morales influenciados por nuestros valores culturales (Narvaez, 2010), y por nuestro lenguaje (Costa *et al.*, 2014). Finalmente, la mentira forma parte inextricable de la vida social humana (Kümmerli, 2011; Riehl y Frederickson, 2016), o esta no sería posible (pensemos en este sentido en los problemas de los individuos con Asperger, ver Li *et al.*, 2011).

Recapitulando, vemos que los agentes humanos no pueden ser considerados agentes racionales en el sentido estricto (e idealizado) del término (Thaler y Sunstein, 2009) ni tampoco mantienen una coherencia moral en sus acciones, puesto que sus heurísticas del comportamiento son inconstantes, incompletas e incoherentes. A pesar de ello, las sociedades humanas existen desde hace mucho tiempo y su tamaño ha ido aumentando de forma exponencial en los últimos milenios. Veamos en el próximo apartado qué elementos reales permiten explicar tal agregación de individuos.

4. UNA APROXIMACIÓN EVOLUTIVA Y PRAGMÁTICA HACIA LO MORAL

La clave para entender la socialización humana es tender el estudio de la misma hacia la era pre-cultural en la que desde una perspectiva naturalista evolutiva podamos identificar los mecanismos que explican el aumento en la complejidad de sistemas sociales. Si bien en las primeras aproximaciones darwinistas a la evolución se primó la noción de competición entre especies e incluso entre los propios individuos de cada una, lo cierto es que es una visión poco acertada. Aquello que parece como la fuente de cohesión social es la noción de cooperación (Axelrod y Hamilton, 1981; Axelrod, 1986; Boyd y Richerson, 2009; Challet y Zhang, 1997; Enke, 2019; Kümmerli, 2011). Entre los mamíferos tal cooperación se encuentra más evidenciada en la cría y la cura de los agentes implicados (Lukas y Clutton-Brock, 2012), partiendo de un sistema neurológico destinado al aprendizaje por imitación sensoriomotriz: la empatía (Debes, 2017; Iacoboni, 2009). Es decir, lo que constituye el pilar sobre el que se asienta la vida social humana es la capacidad de colaborar, de cuidarnos los unos a los otros. Desde tal horizonte, los sistemas normativos que capturen la esencia de este mecanismo tendrán como recompensa sociedades más estables, al mismo tiempo que capaces de aumentar su complejidad. Como también hemos indicado en apartado anteriores, el control moral será no sólo intrasocial sino también interior, bajo la creencia en fuerzas supranaturales que observan al individuo.

En consecuencia, el nacimiento de lo moral es debido a un diseño social que está regido por unas pautas de interacción que permiten la existencia de las propias sociedades. La ética de la cura es uno de los pilares fundamentales de la cohesión social (Hel, 2018; Slote, 2007), aunque desde la perspectiva interesada del heteropatriarcado neoliberal ha sido ninguneada y menospreciada, para ensalzar códigos morales inspirados por la competición y la jerarquización como motor del cambio (y la mejora) social. Esta cura estaba sustentada por las mujeres dentro de las comunidades, o de las familias, de manera que a pesar de ser algo finalmente necesario para la cohesión social, su valor ha sido menospreciado a lo largo de la historia. En cualquier caso, cooperación y cura emergen como elementos fundamentales de las sociedades humanas, por lo que la creación de códigos morales efectivos debe partir de tales axiomas. Al mismo tiempo, debemos considerar la razón de tales pilares: nuestra propia morfología, de tipo emocional y social. Certeramente, Aristóteles acertó en sus aproximaciones naturalistas a la naturaleza humana: somos animales sociales que buscamos respuestas con sentido acerca del mundo. Y el sentido se encuentra situado por la propia corporalidad, que emerge y se manifiesta y siente a través de lo simbólico. Bajo este prisma, resulta claro que una ética que no parta de las emociones para centrarse en acciones o valores no es viable (Vallverdú, 2007b).

5. ¿ÉTICAS DIFUSAS PARA SISTEMAS LÓGICOS?

Cuando estamos a punto de crear inteligencias autónomas artificiales, fruto del pensamiento lógico más preciso y avanzado del que nunca la humanidad ha gozado antes, nos encontramos ante un problema sorprendente, si bien no inesperado: los sistemas morales que desearíamos implementar en tales máquinas no son susceptibles de poder ser completamente operativos. En primer lugar, por su indefinición general (*todo el mundo tiene derecho a…* pero sin instrucciones para su realización, por ejemplo); en segundo lugar, por su localidad, al no ser aceptados de forma universal; en tercer lugar, por su incompletitud, por no decir contradicción. En relación con este último punto, pensemos por ejemplo en las absurdas diver-

ARBOR Vol. 197-800, abril-junio 2021, a601 | ISSN-L: 0210-1963 https://doi.org/10.3989/arbor.2021.800003

gencias interpretativas de muchos jueces aplicando las mismas leyes. Por ello se han iniciado estudios para la automatización de la justicia en determinadas áreas de lo social, en países como Estonia o China. Los problemas con estos sistemas son recurrentes: la fiabilidad de los datos, la corrección de los algoritmos utilizados, la fiabilidad del sistema en su conjunto. O yendo incluso más allá, en los sesgos flagrantes que acompañan a los juicios tanto legales como morales, como hemos podido ver en los usos del sistema de IA COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) en los Estados Unidos de América (Bechter y Lowenkamp, 2016; Larson *et al.*, 2016), y en otras disyunciones que dieron lugar a la prohibición del análisis de datos derivados de la actividad judicial individual por parte la LegalTech (tecnología y software para ofrecer servicios jurídicos) en algunos países como Francia (Bues y Matthaei, 2017; en relación al caso francés, véase el artículo 33 de la Reforma de la Justicia publicada en marzo de 2019); en tercer lugar a la variabilidad moral existente dentro de un mismo sistema cultural (sin que hablemos explícita ni necesariamente de sociedades multiculturales) impide diseñar sistemas generalmente válidos, es decir, que respondan de forma cohesionada a los valores de toda una comunidad en sí diversa (que se pueda expresar con libertad, claro está); y, en cuarto lugar, el carácter evolutivo de los sistemas morales, los cuáles evolucionan y son modificados con el paso del tiempo. La lucha por nuevos listados de derechos se ha visto afectada por esta variabilidad evolutiva.

Los sistemas inteligentes artificiales han puesto por lo tanto sobre la mesa un problema obvio y terriblemente devastador: los humanos coordinamos nuestras acciones morales recurriendo de forma oportunista a una multiplicidad de heurísticas, sobre las cuales improvisamos y realizamos cambios de forma continuada. Lo interesante es ver que, a pesar de todas estas fisuras del pensamiento único ético, podemos identificar patrones de decisión en relación a aspectos culturales geográficamente situados. Ello podría parecernos una debilidad en el camino hacia una ética universal, pero en realidad codifica un valor superior de mayor importancia: la existencia de diversidad cultural, lingüística, simbólica, moral. En estas diferencias residen los sentires reales de la humanidad en tanto que conjunto de seres en continua adaptación. Al no ser la propia ética un sistema formalizable de forma consistente y completa, no cabe preocuparse por la imposibilidad de disponer de modelos artificiales homogéneos para nuestras máquinas. Todo ello sin tener en cuenta sesgos muchos más graves e irresolubles, como los de tipo supranaturalista; pensemos, por ejemplo, en los modos sabbath preinstalados en muchos hornos de alta gama (Bix, 2020; Woodruff, Augustin y Foucault, 2007). ¿Debemos permitir que los robots domésticos se ciñan a prescripciones religiosas? Incluso yendo más allá: ¿es lícito diseñar robots que cumplan funciones estrictamente religiosas, como sacerdotes o incluso reificaciones de divinidades? Ello ya está sucediendo especialmente en Asia (Mori, 1989; Wagner, 2009; Wong, 2019).

Gracias a la enumeración de las contradicciones presentes en la implementación computacional de códigos éticos o morales, estamos asistiendo a una renegociación y debate sobre lo ético y lo moral. Y también a un posicionamiento de cerrazón y conservadurismo por algunos de los agentes amenazados por la multiculturalidad cuando hasta hace poco campaban a sus anchas sin oposición alguna.

CONCLUSIONES

Retomando el título del artículo, constatamos que no es posible implementar modelos éticos falibles en sistemas computacionales de formalismos rigurosos. Por lo menos no desde una perspectiva universalista, sino tal vez desde la visión parcial de una de las formas culturales que tomemos como punto de partida. Las soluciones son complejas e insatisfactorias: ni una ética de mínimos es útil en un contexto de teoría de juegos aplicados a la moralidad, ni la implementación de un sistema estadístico con el uso del azar en las decisiones parece justo (Pauly, 1968; Rowell y Connelly, 2012; Stevens y Thevaranjan, 2010). Si debemos decidir qué pacientes tienen preferencia en las listas para recibir un trasplante de órgano, el azar no puede tener cabida alguna, o por lo menos de este modo es como necesitamos afrontar tales retos. Y teniendo en cuenta que la mayor parte de decisiones sociales se encuentran condicionadas por las implicaciones económicas (distribución de recursos, derecho a ciertas prestaciones…), podemos entender cómo los propios modelos sobre el diseño económico abren la veda al debate sin fin. Es decir, las numerosas contradicciones presentes en los códigos éticos humanos (tanto en su diseño interno, como las que surgen al compararlos entre sí) impiden una implementación algorítmica satisfactoria y global. En el análisis del debate sobre universales (o no) éticos humanos, no debemos caer en las redes de la escolástica académica a la cual son tan proclives los filósofos profesionales; más bien, debemos partir de las eviden-

ARBOR Vol. 197-800, abril-junio 2021, a601 | ISSN-L: 0210-1963 https://doi.org/10.3989/arbor.2021.800003

cias actuales. Una aproximación pragmática y sincera a la acción humana de los últimos tres milenios nos impide afirmar que un solo sistema ético haya demostrado ser universal y aceptado racionalmente por los seres humanos. Si nos planteamos qué códigos éticos con aspiración universal aparecieron durante el siglo XX (Declaración Universal Derechos Humanos, Declaración de Helsinki, Protocolo de Kyoto) todos ellos son respuestas a situaciones de descontrol. Y a pesar de sus redacciones, el cumplimiento es totalmente insatisfactorio (pensemos por ejemplo en el informe anual de derechos humanos realizado por Amnistía Internacional, con detalles de todas las graves infracciones sistemáticas, Estado a Estado).

AGRADECIMIENTOS

Las investigaciones del Prof. Vallverdú han sido financiadas por el proyecto *Innovación epistemológica: el caso de las ciencias biomédicas* (FFI2017-85711-P), y una beca de investigación ICREA Acadèmia (2020-2025). A los revisores, por sus aportaciones de carácter formal académico.

REFERENCIAS

Allinson, Robert E. (1992). The golden rule as the core value in confucianism & christianity: Ethical similarities and differences. *Asian Philosophy*, 2 (2): 173-185. https://doi.org/10.1080/09552369208575363

Awad, Edmond, Dsouza, Sohan, Kim, Richard, Schulz, Jonathan, Henrich, Joseph, Shariff, Azim, Bonnefon, Jean-François, & Rahwan, Iyad (2018). The Moral Machine experiment. *Nature*, 563: 59-64. https://doi.org/10.1038/s41586-018-0637-6

Awad, Edmond, Dsouza, Sohan, Shariff, Azim, Rahwan, Iyad, & Bonnefon, Jean-François (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences of the United States of America*, 117 (5): 2332-2337. https://doi.org/10.1073/pnas.1911517117

Axelrod, Robert & Hamilton, William D. (1981). The evolution of cooperation. *Science*, 211(4489): 1390-1396. https://doi.org/10.1126/science.7466396

Axelrod, Robert. (1986). An evolutionary approach to norms. *American Political Science Review*, 80 (4): 1095-1111. https://doi.org/10.1017/S0003055400185016

Bix, Amy Sue (2020). 'Remember the Sabbath': a history of technological decisions and innovation in Orthodox Jewish communities. *History and Technology*, 36 (2): 205-239. https://doi.org/10.1080/07341512.2020.1816339

Botero, Carlos A., Gardner, Beth, Kirby, Kathryn R., Bulbulia, Joseph, Gavin, Michael C., & Gray, Russell D. (2014). The ecology of religious beliefs. *Proceedings of the National Academy of Sciences of the United States of America*, 111 (47): 16784-16789. https://doi.org/10.1073/pnas.1408701111

Boyd, Robert, & Richerson, Peter J. (2009). Culture and the evolution of human cooperation. In *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364: 3281–3288. https://doi.org/10.1098/rstb.2009.0134

Brown, Donald E. (2004). Human universals, human nature & human culture. *Daedalus*, 133 (4): 47-54. https://doi.org/10.1162/0011526042365645

Brown, Donald E. (2017). Our Common Denominator: Human Universals Revisited. *Evolutionary Studies in Imaginative Culture*, 1 (1): 213-216. https://doi.org/10.26613/esic/1.1.28

Buchtel, Emma E., & Norenzayan, Ara (2008). Which should you use, intuition or logic? Cultural differences in injunctive norms about reasoning. *Asian Journal of Social Psychology*, *11*(4): 264-273. https://doi.org/10.1111/j.1467-839X.2008.00266.x

Bues, Micha-Manuel, & Matthaei, Emilio (2017). LegalTech on the Rise: Technology Changes Legal Work Behaviours, But Does Not Replace Its Profession. En: Jacob K., Schindler D., Strathausen R. (eds) *Liquid Legal. Management for Professionals,* pp. 89-110. Springer, Cham. https://doi.org/10.1007/978-3-319-45868-7_7

Caviola, Lucius, Mannino, Adriano, Savulescu, Julian, & Faulmüller, Nadira (2014). Cognitive biases can affect moral intuitions about cognitive enhancement. *Frontiers in Systems Neuroscience*, *8*. https://doi.org/10.3389/fnsys.2014.00195

Challet, Damien, & Zhang, Y. C. (1997). Emergence of cooperation and organization in an evolutionary game. *Physica A: Statistical Mechanics and Its Applications*, 246 (3-4): 407-418. https://doi.org/10.1016/S0378-4371(97)00419-6

Clausen, Jens, & Levy, Neil (2015). Handbook of neuroethics. In *Handbook of Neuroethics*. Springer. https://doi.org/10.1007/978-94-007-4707-4

Cortina Orts, Adela (1986). *Ética Mínima*. Tecnos.

Costa, Albert, Foucart, Alice, Hayakawa, Sayuri, Aparici, Melina, Apesteguia, Jose, Heafner, Joy, & Keysar, Boaz (2014). Your morals depend on language. *PLoS ONE*, *9* (4). https://doi.org/10.1371/journal.pone.0094842

Curry, Oliver S., Mullins, Daniel A., & Whitehouse, Harvey (2019). Is it good to cooperate?: Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, 60 (1). https://doi.org/10.1086/701478

Debes, Remy (2017). Empathy and mirror neurons. In *The Routledge Handbook of Philosophy of Empathy*, pp. 54–63. https://doi.org/10.4324/9781315282015

Enke, Benjamin (2019). Kinship, cooperation, and the evolution of moral systems. *Quarterly Journal of Economics*, 134 (2): 953-1019. https://doi.org/10.1093/qje/qjz001

Flores, Anthony W., Bechtel, Kristin, & Lowenkamp, Christopher T. (2016). False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. And it's biased against blacks." *Federal Probation Journal*, 80 (2): 38-46.

Hel, Virginia (2018). The ethics of care. In *The Oxford Handbook of Distributive Justice*. https://doi.org/10.1093/oxfordhb/9780199645121.013.12

Huang, Minyao, & Jaszczolt, Kasia M. (eds) (2018). *Expressing the self: cultural diversity and cognitive universals*. Oxford University Press.

Iacoboni, Marco (2009). Imitation, Empathy, and Mirror Neurons. *Annual Review of Psychology*, *60* (1): 653–670. https://doi.org/10.1146/annurev.psych.60.110707.163604

Jakubecki, Natalia G. (2012). Los inicios del pensamiento escolástico: el "Sic et Non" de Pedro Abelardo. *Revista Española de Filosofía Medieval,* 19: 31-38. https://doi.org/10.21071/refime.v19i.6059

Jonauskaite, Domicele, *et al.* (2020). Universal Patterns in Color-Emotion Associations Are Further Shaped by Linguistic and Geographic Proximity. *Psychological Science*, 31 (10): 1245–1260. https://doi.org/10.1177/0956797620948810

Kahneman, Daniel (2011). *Thinking fast, thinking slow*. London: Penguin Books.

Klenk, Michael (2019). Moral Philosophy and the 'Ethical Turn' in Anthropology. *Zeitschrift Für Ethik Und Moralphilosophie*, 2 (2): 331–353. https://doi.org/10.1007/s42048-019-00040-9

Kümmerli, Rolf (2011). A test of evolutionary policing theory with data from human societies. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0024350

Laland, Kevin N., Odling-Smee, John, & Myles, Sean (2010). How culture shaped the human genome: Bringing genetics and the human sciences together. *Nature Reviews Genetics*, 11: 137–148. https://doi.org/10.1038/nrg2734

Lang, Martin, Purzycki, Benjamin G., Apicella, Coren L., Atkinson, Quentin D., Bolyanatz, Alexander, *et al.* (2019). Moralizing gods, impartiality and religious parochialism across 15 societies. *Proceedings of the Royal Society B: Biological Sciences*, 286 (1898). https://doi.org/10.1098/rspb.2019.0202

Larson, Jeff, Mattu, Surya, Kirchner, Lauren, & Angwin, Julia (2016). How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica* 5, 9 (1). [23 May 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm]

Li, Annie S., Kelley, Elisabeth A., Evans, Angela D., & Lee, Kang (2011). Exploring the ability to deceive in children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 41: 185–195. https://doi.org/10.1007/s10803-010-1045-4

Lukas, Dieter, & Clutton-Brock, Tim (2012). Cooperative breeding and monogamy in mammalian societies. *Proceedings of the Royal Society B: Biological Sciences*, 279 (1736). https://doi.org/10.1098/rspb.2011.2468

Marshall, James A. R., Trimmer, Pete C., Houston, Alasdair I., & McNamara, John M. (2013). On evolutionary explanations of cognitive biases. *Trends in Ecology and Evolution*, 28 (8): 469-473. https://doi.org/10.1016/j.tree.2013.05.013

Maruyama, Magoroh (1976). Design principles for extraterrestrial communities. *Futures*, 8 (2): 104–121. https://doi.org/10.1016/0016-3287(76)90061-6

Mori, Masahiro (1989). *The Buddha in the Robot.* Kosei Publishing Company.

Mou, Bo (2004). A reexamination of the structure and content of Confucius' version of The Golden Rule. In *Philosophy East and West*, 54 (2): 218-248. https://doi.org/10.1353/pew.2004.0007

Narvaez, Darcia (2010). The emotional foundations of high moral intelligence. *New Directions for Child and Adolescent Development*, 129: 77-94. https://doi.org/10.1002/cd.276

Navarrete, C. David, McDonald, Melissa M., Mott, Michael L., & Asher, Benjamin (2012). Virtual morality: emotion and action in a simulated three-dimensional "trolley problem". *Emotion*, *12* (2): 364–370. https://doi.org/10.1037/a0025561

Nisbett, Richard E. (2003). *The geography of thought: how Asians and westerners think differently... and why*. New York: The Free Press.

Norenzayan, Ara, & Nisbett, Richard E. (2000). Culture and causal cognition. *Current Directions in Psychological Science*, *9* (4): 132–135. https://doi.org/10.1111/1467-8721.00077

Nowak, Martin A., Page, Karen M., & Sigmund, Karl (2000). Fairness versus reason in the Ultimatum Game. *Science*, 289 (5485): 1773-1775. https://doi.org/10.1126/science.289.5485.1773

Pagallo, Ugo (2018). Vital, Sophia, and Co.—The Quest for the Legal Personhood of Robots. *Information*, *9* (9): 230. https://doi.org/10.3390/info9090230

Pauly, Mark V. (1968). The economics of moral hazard: comment. *The American Economic Review*, 58 (3): 531-537. https://doi.org/10.1017/cbo9780511528248.009

Rembert, Ron B. (1983). The Golden Rule: Two Versions and two Views. *Journal of Moral Education*. https://doi.org/10.1080/0305724830120205

Riehl, Christina, & Frederickson, Megan E. (2016). Cheating and punishment in cooperative animal societies. In *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371 (1687). https://doi.org/10.1098/rstb.2015.0090

Rowell, David, & Connelly, Luke B. (2012). A History of the Term "Moral Hazard." *Journal of Risk and Insurance* 77 (4): 1051-1075. https://doi.org/10.1111/j.1539-6975.2011.01448.x

Sanfey, Alan G., Rilling, James K., Aronson, Jessica A., Nystrom, Leigh E., & Cohen, Jonathan D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, 300 (5626): 1755-1758. https://doi.org/10.1126/science.1082976

Slote, Michael (2007). *The Ethics of Care and Empathy*. Routledge. https://doi.org/10.4324/9780203945735

Stevens, Douglas E., & Thevaranjan, Alex (2010). A moral solution to the moral hazard problem. *Accounting, Organizations and Society*, 35 (1): 125-139. https://doi.org/10.1016/j.aos.2009.01.008

Thaler, Richard H. (1988). Anomalies: The Ultimatum Game. *Journal of Economic Perspectives* 2(4): 195-206. https://doi.org/10.1257/jep.2.4.195

Thaler, Richard H., & Sunstein, Cass R. (2009). *Nudge: Improving Decisions about Health, Wealth, and Happiness: Amazon.es: Thaler, Richard H., Sunstein, Cass R.: Libros en idiomas extranjeros*. Penguin LCC US.

Ushimura, Kei (2003). *Beyond the "Judgement of Civilization". The Intellectual Legacy of the Japanese War Crimes Trials, 1946-1949*. (1st ed.). The International House of Japan.

Vallverdú, Jordi (2007a). Las raíces de lo ético: tras Rabossi. *Cuadernos de Ética*, *22*(35): 89-118.

Vallverdú, Jordi (2007b). *Una ética de las emociones*. Barcelona: Anthropos.

Vallverdú, Jordi (2009). *Bioética computacional. e-Biotecnología: simbiosis de valores*. Fondo de Cultura Económica.

Vallverdú, Jordi (2019). ¿Nazis kantianos? El homo politicus desde la racionalidad limitada o La banalidad de la Ética. In A. Estany & M. Gensollen (Eds.), *Democracia y conocimiento* (1st ed., pp. 245–260). Barcelona: Servei de Publicacions de la Universitat Autònoma de Barcelona.

Vallverdú, Jordi, & Müller, Vincent C. (2019). *Blended cognition: the robotic challenge*. Springer. https://doi.org/10.1007/978-3-030-03104-6

Wagner, Cosima (2009). "The Japanese way of robotics": Interacting "naturally" with robots as a national character? *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 510-515. https://doi.org/10.1109/ROMAN.2009.5326221

Wong, Pak-Hang. (2019). Rituals and Machines: A Confucian Response to Technology-Driven Moral Deskilling. *Philosophies*, 4 (4): 59. https://doi.org/10.3390/philosophies4040059

Woodruff, Allison, Augustin, Sally, & Foucault, Brooke (2007). Sabbath day home automation: "it's like mixing technology and religion." *Conference on Human Factors in Computing Systems - Proceedings*, 527–536. https://doi.org/10.1145/1240624.1240710