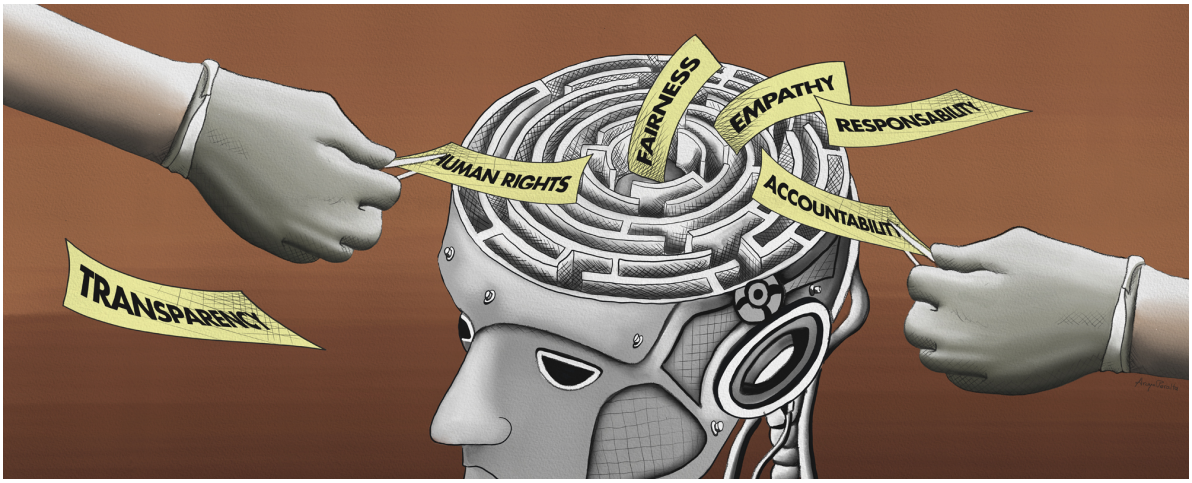
REPTES ÈTICS

Reflexió sobre l'«ètica» en l'ètica de la IA

Pak-Hang Wong, Judith Simon



Araya Peralta

A principis del 2019 una gran consultoria internacional va identificar l'«ètic de la IA» com un càrrec fonamental perquè les empreses apliquin correctament la intel·ligència artificial (IA). Sosté que els ètics de la IA són necessaris per ajudar les empreses a explorar les qüestions ètiques i socials que planteja l'ús de la IA [1]. L'opinió que la IA és positiva i, tot i així, potencialment perillosa per als individus i la societat és molt compartida pel sector, els cercles acadèmics, els governs i les organitzacions de la societat civil. En conseqüència, i a fi de materialitzar-ne els avantatges i alhora esquivar els obstacles ètics i els efectes nefastos, s'han posat en marxa nombroses iniciatives per a) examinar les dimensions ètica, social, jurídica i política de la IA i b) elaborar unes directrius i recomanacions ètiques per al disseny i l'aplicació de la IA [2]

.

Tanmateix, de vegades les qüestions terminològiques dificulten l'anàlisi profunda dels problemes ètics de la IA. Les definicions d'intel·ligència i intel·ligència artificial sovint són imprecises i les diferents interpretacions d'aquests termes destaquen diferents inquietuds. Per evitar la confusió i el risc que tot plegat acabi sent un diàleg de sords, qualsevol debat constructiu sobre l'ètica de la IA exigeix que s'expliqui la definició d'IA que es fa servir i també que s'especifiqui el tipus d'IA sobre el qual es debat. Pel que fa a la definició, consultem el Grup Europeu d'Ètica de la Ciència i de les Noves Tecnologies de la Comissió Europea, que defineix la IA com els «sistemes de programari (i possiblement també de maquinari) dissenyats per éssers humans que, donat un objectiu complex, actuen en la dimensió física o digital percebent el seu entorn mitjançant l'adquisició de dades,

interpretant les dades […] recopilades, raonant sobre el coneixement o processant la informació derivada d'aquestes dades i decidint quines són les millors mesures que cal prendre per assolir un objectiu determinat. Els sistemes d'IA poden utilitzar regles simbòliques o aprendre un model numèric i també poden adaptar la seva conducta analitzant l'impacte de les seves accions anteriors sobre l'entorn» [3].

Per oferir orientació i recomanacions específiques, l'anàlisi ètica de la IA també ha d'especificar la *tecnologia* (per exemple, vehicles autònoms, sistemes de recomanació, etc.), els *mètodes* (per exemple, aprenentatge profund, aprenentatge de reforç, etc.) i els *sectors d'aplicació* (per exemple, atenció sanitària, finances, notícies, etc.). En aquest article ens centrarem en les qüestions ètiques relacionades amb la *IA autònoma*, és a dir, els agents artificials que poden decidir i actuar independentment de la intervenció humana, i il·lustrarem les qüestions ètiques de la IA autònoma amb un gran nombre d'exemples.

Fixem-nos primer en el cas dels vehicles autònoms (VA). La possibilitat que hi hagi escenaris d'accidents amb VA implicats, en què inevitablement passatgers o vianants resultarien ferits, ha obligat els investigadors i desenvolupadors a reflexionar sobre qüestions relatives a l'acceptabilitat ètica de les decisions preses pels VA, com ara quines decisions han de prendre els VA en aquests casos, com es poden justificar aquestes decisions, quins valors reflecteixen els VA i les seves decisions, etc [4].

> Els algoritmes de contractació acostumen a funcionar aplicant els criteris que han après d'un conjunt de dades d'entrenament. Lamentablement, aquestes dades d'entrenament poden ser esbiaixades i donar lloc a models potencialment discriminatoris

Encara millor, fixem-nos en el cas dels algoritmes de contractació, que s'han introduït per automatitzar el procés de recomanació, preselecció i potser fins i tot selecció de candidats a una feina. Els algoritmes de contractació acostumen a funcionar aplicant els criteris que han après d'un conjunt de dades d'entrenament. Lamentablement, aquestes dades d'entrenament poden ser esbiaixades i donar lloc a models potencialment discriminatoris [5].

A fi de garantir la protecció contra la discriminació, que no només és un dret humà sinó que està inclosa en les constitucions de molts països, ens hem d'assegurar que aquests algoritmes són, com a mínim, no discriminatoris i idealment també *equitatius*. Tanmateix, l'equitat té diferents interpretacions: les persones discrepen sobre el significat de l'equitat, però a més el concepte d'equitat també pot dependre del context. Encara més, s'ha demostrat que no es poden obtenir diferents paràmetres d'equitat de manera simultània [6]. Això planteja la qüestió de com s'han de concebre valors com l'equitat en quin context i com es poden aplicar.

Per tant, una de les qüestions fonamentals de l'ètica de la IA es pot formular com un

problema d'alineació de valors: com podem crear una IA autònoma alineada amb els valors arrelats en la societat [7]. Virginia Dignum ha distingit tres dimensions de l'ètica de la IA: l'«ètica per disseny», l'«ètica en el disseny» i l'«ètica per al disseny» [8], que serveixen per identificar dues respostes diferents al problema de l'alineació de valors. Estructurarem l'anàlisi següent d'acord amb les tres dimensions esmentades i explorarem més detalladament les dues vies diferents per respondre al problema de l'alineació de valors.

Crear una IA ètica: perspectives i limitacions

L'ètica per disseny és «la integració tècnica/algorítmica de la capacitat de raonament com a part de la conducta de [la IA autònoma]» [9]. Aquesta línia de recerca també s'anomena «ètica de les màquines» i aspira a crear agents morals artificials, que són agents artificials amb capacitat ètica i poden prendre decisions ètiques sense la intervenció humana [10]. Així doncs, l'ètica de les màquines respon al problema de l'alineació de valors amb la creació d'una IA autònoma que *per si sola* s'alinea amb els valors humans. Il·lustrem aquesta perspectiva amb els exemples dels VA i els algoritmes de contractació: els investigadors i desenvolupadors s'esforcen per crear VA que puguin raonar sobre la decisió moralment correcta i actuar en conseqüència en escenaris en què es provoqui un dany inevitable. Així mateix, se suposa que els algoritmes de contractació prenen decisions no discriminatòries sense la intervenció humana.

Wendell Wallach i Colin Allen van definir tres tipus d'aproximacions a l'ètica de les màquines en el seu llibre seminal *Moral machines* [11]. Els tres tipus d'aproximacions són, respectivament, (i) aproximacions descendents, (ii) aproximacions ascendents i (iii) aproximacions híbrides, que combinen les aproximacions descendents i ascendents. En la seva forma més senzilla, l'aproximació descendent intenta formalitzar i aplicar una teoria ètica concreta a la IA autònoma, mentre que l'aproximació ascendent pretén crear una IA autònoma que pugui aprendre de l'entorn o a partir d'una sèrie d'exemples què és moralment correcte o no. En darrer lloc, l'aproximació híbrida combina tècniques i estratègies tant de l'aproximació descendent com de l'ascendent [12].

Aquestes aproximacions, però, estan subjectes a algunes limitacions *teòriques* i *tècniques*. Per exemple, les aproximacions descendents han de superar el repte de trobar i defensar una teoria ètica incontrovertible entre tradicions filosòfiques *contradictòries*. Si no, la IA ètica corre el risc de construir-se sobre una base *inadequada* o fins i tot *falsa*. D'altra banda, les aproximacions ascendents dedueixen què és ètic a partir d'allò que és *popular* o que es considera *universalment* ètic en el context o entre exemples. Amb tot, aquestes deduccions no garanteixen que la IA autònoma adquireixi regles o principis ètics *genuïns*, perquè ni la popularitat ni el fet que quelcom es consideri ètic no ofereixen una *justificació* ètica adequada [13]. A més, hi ha la dificultat tècnica de crear una IA ètica que pugui discernir amb eficàcia la informació èticament rellevant de l'èticament irrellevant entre la gran quantitat d'informació disponible en un context determinat. Aquesta capacitat seria necessària tant per a l'aplicació correcta dels principis ètics en les aproximacions descendents com per a l'adquisició correcta dels principis ètics en les aproximacions ascendents [14].

> La IA autònoma en general i la IA ètica en particular poden erosionar considerablement l'autonomia humana perquè les decisions que prenen per nosaltres o sobre nosaltres estaran fora del nostre control

A part dels reptes teòrics i tècnics, la creació d'IA autònoma amb capacitats ètiques ha rebut algunes crítiques ètiques. En primer lloc, la IA autònoma en general i la IA ètica en particular poden erosionar considerablement l'autonomia humana perquè les decisions que prenen per nosaltres o sobre nosaltres estaran fora del nostre control i, per tant, reduiran la nostra independència de les influències externes [15]. En segon lloc, no queda clar qui o què ha de ser responsable de les decisions errònies que la IA autònoma prengui, la qual cosa suscita inquietud en relació amb els seus efectes sobre les nostres pràctiques de responsabilitat moral [16]. En darrer lloc, els investigadors han argumentat que convertir la IA autònoma en agents morals o pacients morals complica innecessàriament el nostre món moral amb la introducció d'elements estranys que són aliens a la nostra consciència moral i, en conseqüència, amb la imposició d'una càrrega ètica innecessària sobre els éssers humans, ja que ens exigeix que parem una atenció moral excessiva a la IA autònoma [17].

L'ètica de les màquines, l'ètica truncada

La nostra anàlisi de les dificultats teòriques, tècniques i ètiques de l'ètica de les màquines no pretén ser exhaustiva ni conclusiva i, en realitat, aquestes dificultats es podrien superar en la recerca i el desenvolupament futurs de la IA autònoma. Això no obstant, creiem que aquestes dificultats es mereixen que ens aturem i reconsiderem les possibilitats de crear una IA ètica. De fet, volem proposar una crítica més fonamental de l'ètica de les màquines abans d'explorar una altra via per respondre al problema d'alineació de valors.

Recordem que l'objectiu de l'ètica de les màquines és crear una IA autònoma que pugui prendre decisions ètiques i actuar amb ètica sense la intervenció humana. Se centra a infondre a la IA autònoma la capacitat de prendre decisions ètiques i fer accions ètiques, cosa que reflecteix una interpretació peculiar de l'«ètica» que escollim qüestionar. Més concretament, si se centra només en la capacitat de decidir i actuar amb ètica, l'ètica de les màquines s'exposa a una visió truncada de l'ètica que considera que les decisions i accions ètiques es poden separar dels seus contextos socials i relacionals. La filòsofa i novel·lista Iris Murdoch, per exemple, fa molt de temps argumentava que la moralitat no té res a veure amb «una sèrie d'eleccions explícites que es produeixen en una sèrie de situacions especificables» [18], sinó amb l'«autorreflexió o actituds complexes davant la vida que es manifesten i s'elaboren contínuament en el discurs públic i privat, però no es poden separar temporalment en situacions» [19]. Per a Murdoch, allò que és ètic està inherentment vinculat a un rerefons de valors. Per això, a l'hora de reflexionar sobre l'ètica, és fonamental mirar més enllà de la capacitat de decidir i actuar amb ètica i els moments de la decisió i l'acció ètiques, i fixar-se en el rerefons de valors i els relats que hi ha darrere de la decisió i l'acció. S'han esgrimit arguments semblants per reafirmar la funció dels contextos

socials i relacionals a l'hora de limitar les decisions ètiques i configurar les conseqüències morals i, per tant, la importància de tenir-los en compte en la nostra reflexió ètica [20].

En aquesta mateixa línia crítica, l'èmfasi per infondre les capacitats ètiques de la IA a l'ètica de les màquines es pot considerar una equivocació si aquest èmfasi eclipsa el fet que les conseqüències ètiques de la IA autònoma estan determinades per diversos factors interconnectats aliens a la seva capacitat de raonament ètic, i que hi ha un procés dilatat de negociació social i política sobre els criteris de correcció i incorrecció que assenyalen les possibles decisions i accions ètiques de la IA autònoma. L'experiment de la Màquina Moral, dut a terme per investigadors del MIT Media Lab, és un bon exemple [21]. En l'experiment, els investigadors del MIT proven d'externalitzar les decisions ètiques en diferents escenaris d'accidents amb VA implicats i els resultats han de servir per orientar el disseny ètic dels VA. Tanmateix, falten el rerefons social, cultural i polític i les històries personals d'accidents reals que els escenaris d'accidents de l'experiment no descriuen —i sovint no poden descriure— adequadament [22]. En aquest sentit, l'experiment de la Màquina Moral també es basa en una visió truncada de l'ètica, en què *només* es té en compte la decisió que s'ha de prendre en situacions concretes i oblida el rerefons de valors i les dades contextuals crucials per prendre decisions ètiques.

> A l'hora de reflexionar sobre l'ètica, és fonamental mirar més enllà de la capacitat de decidir i actuar amb ètica i els moments de la decisió i l'acció ètiques, i fixar-se en el rerefons de valors i els relats que hi ha darrere de la decisió i l'acció

De fet, els contextos socials i relacionals són importants per a l'anàlisi ètica de la IA autònoma *abans* i *després* de la seva aplicació. Per exemple, es pot dissenyar un algoritme de contractació imparcial, que valori els candidats a un lloc de treball basant-se *només* en les qualitats requerides en una oferta. Tot i així, aquest algoritme de contractació imparcial podria continuar sent discriminatori, i per tant èticament discutible, si les qualitats específiques requerides en una oferta estan lligades involuntàriament a la raça, el gènere i la classe social. En aquest cas, cal anar amb compte per no reproduir el *biaix social preexistent* en l'algoritme de contractació. A més, fins i tot les tecnologies millor intencionades poden afectar molt negativament els (no) usuaris, ja que el biaix i el perjudici podrien sorgir de la interacció entre la tecnologia i els usuaris i la societat [23]. Imagineu una aplicació que els ciutadans poden fer servir per denunciar incidents, com ara infraestructures deteriorades, a l'ajuntament de la ciutat, que al seu torn utilitza un algoritme per ordenar i classificar els problemes locals segons aquestes denúncies. Si donem per fet que l'accés als telèfons intel·ligents i, per tant, a l'aplicació té una distribució desigual, això pot significar una falta de denúncies a les zones amb ciutadans més pobres. Si aquest biaix de les dades d'entrada no es té en compte en l'ordenació i la classificació algorítmiques, les desigualtats entre les zones més riques i més pobres de la ciutat podrien créixer encara més [24].

La lliçó principal d'aquests dos exemples és que incloure unes quantes normes i principis ètics en la IA autònoma és insuficient per resoldre el problema d'alineació de valors, perquè els rerefons i els contextos *sí* que contribueixen a la nostra posició global sobre què és ètic. Ens hem de recordar que la IA autònoma *sempre* se situa en uns contextos socials i relacionals més amplis i que, per això, no ens podem centrar *només* en la seva *capacitat* de prendre decisions i d'actuar amb ètica. Hem de tenir en compte no només *quines* decisions i accions ha de fer la IA autònoma, sinó també (i) *per què* nosaltres (o, millor dit, la societat) creiem que aquestes decisions i accions són ètiques, (ii) *com* arribem a aquestes conclusions i (iii) *si* tenim justificació per pensar així. Així doncs, l'experiment de la Màquina Moral és qüestionable perquè suposa injustificablement que la resposta més *intuïtiva* o *popular* als escenaris d'accidents és la resposta *ètica*. En realitat, la reformulació de les preguntes ens ofereix dos avantatges. Primer, ara podem incloure fàcilment altres parts i factors *a més de* la IA autònoma en la nostra reflexió ètica. Segon, també explicita la possibilitat de (re)negociar quines regles o quins principis ètics s'han d'incloure en la IA autònoma (o fins i tot qüestionar l'ús de la IA autònoma en un context concret).

Una ètica distribuïda de la IA

Siguem clars: no neguem la necessitat d'examinar els valors incorporats a la tecnologia ni la importància de dissenyar i crear tecnologia amb uns valors alineats amb els interessos humans [25]. Tal com demostren els exemples d'aquest article, la IA autònoma pot tenir un paper en la presa de decisions ètiques i ocasionar uns efectes èticament rellevants. Per això, cal examinar els valors que incorpora i utilitzar els valors socials compartits per orientar-ne el disseny i el desenvolupament. Això no obstant, volem qüestionar el desig de delegar el raonament ètic i la decisió a les màquines i, així, arrancar el raonament i la decisió dels contextos socials i relacionals. Una bona anàlisi de l'ètica de la IA ha d'ampliar el seu àmbit de reflexió i incloure altres parts i factors que siguin rellevants per a la presa de decisions ètiques i hagin contribuït als efectes ètics de la IA autònoma. Per a això, és fonamental que l'ètica de la IA inclogui diferents parts interessades (com ara responsables polítics, directius d'empreses, dissenyadors, enginyers, usuaris, no usuaris i el públic en general) en la reflexió ètica sobre la IA autònoma. De fet, només si ho fem així podrem donar una resposta adequada a les preguntes següents: (i) *per què* creiem que les decisions i els efectes de la IA són ètics, (ii) *com* arribem a aquestes conclusions i (iii) *si* tenim justificació per opinar així.

> En el disseny i l'aplicació de la IA s'han de tenir en compte les desigualtats i les injustícies socials existents, considerar-les i, en el millor dels casos, intentar atenuar-les mitjançant les decisions de disseny

Aquesta ètica ampliada de la IA l'anomenarem l'*ètica distribuïda de la IA*. L'adjectiu «distribuïda» pretén reflectir el fet que hi ha múltiples parts i factors que són rellevants i

han contribuït als efectes ètics de la IA autònoma i que, per tant, la responsabilitat es «distribueix» entre les parts i els factors rellevants i coadjuvants [26]. Tornem als exemples dels VA i els algoritmes de contractació: una mala ordenació urbana i unes carreteres deficients han de ser preocupacions legítimes en l'ètica dels VA, de la mateixa manera que els biaixos socials i culturals existents són consideracions vàlides per als algoritmes de contractació ètics. Per tant, en el disseny i l'aplicació de la IA s'han de tenir en compte les desigualtats i les injustícies socials *existents*, considerar-les i, en el millor dels casos, intentar atenuar-les mitjançant les decisions de disseny.

L'ètica distribuïda de la IA necessita el que Dignum va denominar «ètica *en* el disseny», és a dir, «els mètodes legislatius i d'enginyeria que ajuden a analitzar i avaluar les implicacions ètiques dels sistemes d'IA, ja que aquests integren o substitueixen les estructures socials tradicionals», així com l'«ètica *per al* disseny», és a dir, «els codis de conducta, les normes i els processos de certificació que garanteixen la integritat dels desenvolupadors i els usuaris quan investiguen, dissenyen, construeixen, utilitzen i gestionen sistemes d'intel·ligència artificial» [27]. Les qüestions ètiques de la IA autònoma no es poden resoldre amb una capacitat ètica *individual(ista)* «millor», sinó només amb l'esforç col·lectiu. Per orientar aquest esforç col·lectiu, les directrius ètiques ofereixen recursos pràctics per suscitar una reflexió basada en valors i principis en relació amb la IA autònoma i coordinar amb eficàcia l'esforç entre les diferents parts rellevants i coadjuvants [28].

Conclusions: sobre la IA fiable de la UE

L'abril del 2019, el grup d'experts d'alt nivell va publicar les *Directrius ètiques per a una IA fiable*, en què es concreta la visió europea sobre la IA. D'acord amb aquestes directrius, Europa ha d'investigar i desenvolupar una IA fiable, és a dir, *lícita, ètica i robusta*.

Hi ha dos punts de les directrius que mereixen una menció especial en el tema de debat que ens ocupa. En primer lloc, és interessant assenyalar que la preocupació per la fiabilitat a les directrius concerneix «no només les propietats inherents a aquesta tecnologia, sinó també les qualitats dels sistemes sociotècnics en què s'aplica la IA. […] Per tant, els esforços adreçats a garantir la fiabilitat de la IA no concerneixen només la confiança que suscita el mateix sistema d'IA, sinó que requereixen un enfocament integral i sistèmic que abraci la fiabilitat de tots els agents i processos que formen part del context sociotècnic en què s'emmarca el sistema al llarg de tot el seu cicle de vida». En aquest sentit, la idea d'una IA fiable coincideix clarament amb l'ètica distribuïda de la IA que hem descrit abans. En segon lloc, també és interessant assenyalar que els quatre principis ètics identificats en les directrius són *principis de nivell mitjà*, és a dir:

1. El principi de respecte de l'autonomia humana.
2. El principi de prevenció dels danys.
3. El principi d'equitat.
4. El principi d'explicabilitat.

La formulació de principis ètics basats en *principis de nivell mitjà* és especialment esclaridora, ja que els principis de nivell mitjà *exigeixen* una interpretació humana i organitzar-ne l'aplicació i no estan pensats per (i de fet no poden) aplicar-se a la IA autònoma. La necessitat d'interpretació i organització també apunta als contextos socials i relacionals, on es troben els recursos per a la interpretació i l'organització.

Tot i que la idea europea d'una IA fiable i les directrius tenen uns fonaments sòlids des de la perspectiva conceptual, plantegen alguns problemes. Per exemple, l'ús de principis de nivell mitjà en les directrius dona força espai per a la interpretació, cosa que els agents malintencionats poden aprofitar per seleccionar les interpretacions i espolsar-se la responsabilitat. Aquest problema s'agreuja per l'èmfasi en l'autorregulació que es fa en les directrius. Els polítics i les empreses poden fer falses promeses amb mesures *barates* i *superficials*, com ara la propaganda i la creació de consells assessors simbòlics, sense abordar *substancialment* els efectes negatius de la IA. Per tant, hi ha problemes importants relacionats amb el marc regulador i institucional *real* pel que fa a l'ètica de la IA i la materialització d'aquesta idea europea. En concret, hi ha la necessitat de crear un marc clar per fer una distribució *equitativa* dels beneficis i els riscos de la IA i la necessitat d'introduir lleis i normatives «dures» contra la violació dels valors ètics bàsics i els drets humans.

Malgrat aquests problemes, l'atenció de les directrius *en els humans* i *més enllà de la tecnologia* s'ha de considerar una posició *normativa* adequada per a l'ètica de la IA i la idea europea. Per acabar aquest article, volem recordar que les qüestions ètiques relatives a la IA autònoma es distribueixen per naturalesa i que nosaltres (o, millor dit, la societat) hem de tenir veu en el seu disseny i la seva aplicació.

REFERÈNCIES

1 — KPMG (2019) Top 5 AI hires companies need to succeed in 2019.

2 — AlgorithmWatch ha elaborat una llista de marcs i directrius ètics disponible a: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/.

3 — European Commission High-Level Expert Group on Artificial Intelligence [AI HLEG] (2019) Ethics guidelines for trustworthy AI. European Commission.

4 — El tipus d'escenaris d'accidents es coneix com «el dilema de la vagoneta». És tan sols un dels temes que es debaten en l'ètica dels vehicles autònoms i el posem d'exemple només per il·lustrar un dels nombrosos problemes ètics que la IA autònoma podria plantejar. Vegeu:
Lin, P. (2016) Why ethics matters for autonomous cars. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous Driving: Technical, Legal and Social Aspects* (pp. 69-85). Berlin: Springer.
Keeling, G. (2019) Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*.

5 — Bogen, M. (2019) All the ways hiring algorithms can introduce bias. *Harvard Business Review*, May 6, 2019.

6 — Friedler, S., Scheidegger, C., & Venkatasubramanian, S. (2016) On the (Im)possibility of fairness. arXiv:1609.07236.
Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* 5(2): 153-163.
Wong, P.-H. (2019) Democratizing algorithmic fairness. *Philosophy & Technology*.

7 — El problema de l'alineació de la IA el va formular explícitament per primera vegada Stuart Russell el 2014. Vegeu:
Peterson, M. (2019) The value alignment problem: a geometric approach. *Ethics and Information Technology* 21 (1): 19-28.

8 — Dignum, V. (2018) Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology* 20 (1): 1-3.

Íbid., p. 2

10 — Winfield, A., Michael, K., Pitt, J., & Evers, V. (2019) Machine ethics: the design and governance of ethical AI and autonomous systems. *Proceedings of the IEEE* 107 (3): 509-517.
Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.
Misselhorn, C. (2018) Artificial morality. concepts, issues and challenges. *Society* 55 (2): 161-169.

11 — Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

12 — Íbid., p. 79-81

13 — Trobareu una anàlisi de la dificultat de l'ètica de les màquines a:
Cave, S., Nyrup, R., Vold, K., & Weller, A. (2019) Motivations and risks of machine ethics. *Proceedings of the IEEE* 107 (3): 562-74.

14 — Això també es coneix com el problema del marc moral. Vegeu:
Horgan, T., & Timmons, M. (2009) What does the frame problem tell us about moral normativity? *Ethical Theory and Moral Practice* 12 (1): 25-51.

15 — Danaher, J. (2018) Toward an ethics of AI assistants: an initial framework. *Philosophy & Technology* 31 (4): 629-653.

16 — Matthias, A. (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6 (3): 175-83.

17 — Bryson, J. J. (2018) Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology* 20 (1): 15-26.

18 — Murdoch, I. (1956) Vision and choice in morality. *Proceedings of the Aristotelian Society, Supplementary* 30: 32-58. p. 34

19 — Íbid., p. 40

20 — Walker, M. U. (2007) *Moral Understandings: A Feminist Study in Ethics*. Oxford: Oxford University Press.

21 — Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018) The Moral Machine experiment. *Nature* 563: 59-64.

Jaques, A. E. (2019) Why the moral machine is a monster. Paper presented to We Robot 2019, University of Miami, April 11-13, 2019.

23 — Friedman, B., & Nissenbaum, H. (1996) Bias in computer systems. *ACM Transactions on Information Systems* 14 (3): 330-347.

24 — Simon J (2012) E-Democracy and Values in Design. *Proceedings of the XXV World Congress of IVR 2012*.
Simon, J. (2017) Value-sensitive design and responsible research and innovation. In S. O. Hansson (Ed.), *The Ethics of Technology Methods and Approaches* (pp. 219-235). London: Rowman & Littlefield.
26 — Floridi, L. (2013) Distributed morality in an information society. *Science and Engineering Ethics* 19 (3): 727-743.
Simon, J. (2015) Distributed epistemic responsibility in a hyperconnected era. In L. Floridi (Ed.), *The Onlife Manifesto* (pp. 145-159). Cham, Springer.
27 — Dignum, V. (2018) Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology* 20 (1): p. 2.
28 — Floridi, L. (2019) Establishing the rules for building trustworthy. *Nature Machine Intelligence* 1: 261-262.

Pak-Hang Wong

Pak-Hang Wong és graduat en Filosofia i Història per la Universitat de Hong Kong, i doctor per la universitat de Twente. És investigador associat posdoctorat del Grup de Recerca sobre Ètica en Tecnologies de la Informació del Departament d'Informàtica de la Universitat d'Hamburg. El seu àmbit d'interès principal és la filosofia de la tecnologia, on examina els reptes de les tecnologies digitals emergents –Big Data, AI, polítiques d'algoritmes– per comprendre la responsabilitat moral i la pràctica del cultiu de la virtut a través del confucianisme. Va coeditar l'obra *Well-Being in Contemporary Society* (Springer, 2015).

Judith Simon

Judith Simon és llicenciada en Psicologia per la Freie Universität Berlin i Doctora en Filosofia per la Universitat de Viena. Actualment, és professora d'Ètica de la Informació i la Tecnologia a la Universitat d'Hamburg. També és part dels consells de redacció de revistes com *Philosophy and Technology* (Springer) i *Big Data & Society* (Sage), així com membre del Consell d'Ètica alemany. Ha estat becària visitant a la Universitat d'Stanford i investigadora convidada a Trento, a Ljubljana i a l'Institut de Recerca en Intel·ligència Artificial (CSIC-IIIA) de Barcelona. Els seus principals àmbits de recerca són els problemes epistemològics i ètics entorn de les tecnologies de la informació, la comunicació i la computació, com l'ètica informàtica, i les responsabilitats d'usuaris i dissenyadors de les TIC. El 2013 va rebre el premi Herbert A. Simon de la *International Association for Computing and Philosophy* (IACAP).