Els reptes de la intel·ligència artificial des d'una perspectiva crítica i sociotècnica

ROSER PUJADAS
University College London
r.pujadas@ucl.ac.uk

1. De la cosificació de la intel·ligència artificial als reptes de les tecnologies anomenades *intel·ligència artificial*

El terme *intel·ligència artificial* (IA) denota una ambició que alguns situen ja a la Grècia clàssica: la de desenvolupar una tecnologia capaç d'imitar o superar les capacitats cognitives humanes. Es tracta d'un enfocament qüestionable, donat que les intel·ligències humanes són múltiples i socials, però que ha esperonat tot un seguit d'innovacions al llarg dels anys. Tanmateix, la consciència col·lectiva sobre les tecnologies anomenades *IA* sembla que s'ha despertat arran del llançament amb accés obert del bot de conversa (*chatbot*) ChatGPT, d'OpenAI, a finals de 2022, que va anar seguit d'expressions alarmistes sobre els perills de la IA. En concret, experts en IA i líders del sector tecnològic —inclòs el mateix Elon Musk, un dels fundadors d'OpenAI— demanaven una moratòria per tal de preparar-se per afrontar els riscos que generarà la IA. Irònicament, els mateixos responsables del desenvolupament de la IA parlaven de riscos existencials, en un futur, això sí, en què la IA s'assumeix com a inevitable.

Clarament, la IA presenta oportunitats i reptes, però tal com suggereixen els estudis crítics de la IA (Crawford, 2021; Suchman, 2023) per entendre'ls val la pena reflexionar sobre la funció que fan certes concepcions i discursos sobre la IA, qui els produeix i amb quins efectes, i quines qüestions queden relegades. Per començar, el concepte mateix d'IA sembla donar coherència i estabilitat a un fenomen que, de fet, és força difús i distribuït. No és una tecnologia monolítica sinó que requereix diverses tecnologies i pràctiques socials, i es manifesta de manera diversa. Hi ha tecnologies anomenades *IA* que fa temps que s'usen en molts àmbits i, de manera més o menys visible, hi interactuem quotidianament —per exemple, en la detecció de malalties, la previsió del temps, la identificació de persones, la selecció de personal, en la producció i també en la detecció de notícies falses, o en la personalització i classificació de continguts i ofertes en

portals d'internet. Aquestes tecnologies són diverses, com ho són els seus àmbits d'aplicació, així que qualsevol generalització sobre els efectes, bondats o riscos de la IA amaga més que no pas ajuda a discernir, i tendeix a naturalitzar models dominants d'IA.

Aquesta reïficació tendeix a reduir la IA, els seus potencials i riscos, a qüestions estrictament tecnològiques, i a imbuir la IA d'agència pròpia (Suchman, 2023). A més, tant les visions utòpiques de la IA com les distopies de futur plantejades pels líders del sector tecnològic —predominantment, homes, blancs, adinerats i situats a Silicon Valley— poden distreure'ns dels problemes reals que la IA ja causa. Periodistes, activistes i acadèmics han aportat evidències empíriques, entre d'altres, de la implicació de la IA en l'erosió de la privacitat, el biaix i discriminació racial i de gènere, la perpetuació de desigualtats, la desinformació i la producció de notícies i imatges falses, la manipulació ideològica i de comportament, la concentració de poder o els efectes nocius per al medi ambient (Broussard, 2023; Crawford, 2021; Eubanks, 2018). Aquests problemes no són només estrictament tecnològics, sinó estructurals i ambientals i afecten de manera desproporcionada certs col·lectius, infrarepresentats en el sector tecnològic.

2. L'eticoontoepistemologia de la IA: quan tecnologia i coneixement configuren realitats i valors

Els estudis de ciència i tecnologia (coneguts com a STS, de l'anglès *science, technology and society*) palesen que la tecnologia per si mateixa no té agència ni té una força intrínseca inevitable. Hi ha moltes versions possibles de les tecnologies i les implementacions que se'n fan. El disseny i l'ús s'han d'emmarcar en contextos socials específics. El desenvolupament tecnològic és el resultat de negociacions i relacions de poder, però també de condicions materials específiques i requereix tasques sovint invisibilitzades. Alhora, la tecnologia és un actor social i polític important, que incorpora i ajuda a reproduir certs valors i influeix de manera important en les pràctiques i estructures socials. Així que és important que com a investigadors i com a societat n'escrutem el desenvolupament i les aplicacions.

El cas de la IA és particularment punyent, donada la capacitat performativa de la IA, és a dir, de (re)configurar la realitat, és significativa. D'una banda, tot i que la IA es pot fer servir per assessorar humans en tasques complexes, sovint es relaciona amb l'automatització de decisions. Particularment en aquest cas es produeix una delegació d'agència cap a una tecnologia cada cop més difícil d'escrutar. D'altra banda, malgrat que la IA es presenta sovint com una tècnica objectiva, cada versió de la IA incorpora una visió específica de *la* intel·ligència i de com generar prediccions. I és que, tal com suggereixen els estudis de STS i el feminisme, tot coneixement és situat i parcial. A més, les diverses pràctiques de coneixement formen assemblatges sociotècnics (Latour, 2005) i són constitutives de certes configuracions de la realitat i sistemes de valors; per tant, tenen conseqüències ètiques i polítiques. Així, cal preguntar-se: quines són les eticoontoepistemologies (Barad, 2007) de la IA? És a dir, quines són les epistemologies i

pràctiques de coneixement associades a la IA i quines realitats i valors ajuden a configurar? I amb quines conseqüències ètiques i materials?

3. El model d'IA basat en dades massives i en l'extractivisme

Actualment, el terme *IA* sol referir-se a diverses tècniques d'aprenentatge automàtic o *machine learning* (ML). De manera simplificada, el ML es desenvolupa usant un conjunt de dades d'entrenament per produir algoritmes i models matemàtics que representin algun aspecte de la realitat. L'ús del ML permet aplicar aquest model per analitzar de manera automatitzada noves dades i fer prediccions. La idea que una quantitat més gran de dades d'entrenament permet el desenvolupament de models cada cop més acurats, sumada a l'ambició de produir IA d'aplicació general, com ChatGPT, ha acabat provocant l'ús indiscriminat i massiu de dades. Això és problemàtic per diversos motius.

Malgrat les pretensions d'objectivitat i aplicació universal d'alguns models, és una fal·làcia que més quantitat de dades garanteixi una millor representació de la realitat. Qualsevol model d'IA és parcial, té limitacions i contextos específics d'aplicació que poden ser útils. Per exemple, el model de llenguatge extens de ChatGPT prediu estadísticament quina és la combinació de paraules més adient per respondre a una pregunta. Pot ser útil per millorar l'estil d'un text, però a l'hora de respondre una pregunta produeix text versemblant tot i que no necessàriament verídic.

A més, té implicacions ètiques i de justícia social. D'una banda, certes pràctiques d'extracció o compra de dades massives, per exemple del web o les plataformes, violen drets bàsics com la privacitat, o els drets d'autor, i deriven en l'apropiació de coneixement. D'altra banda, les dades recollides no són sempre representatives, es descontextualitzen i per tant perden qualitat o sentit, i alhora contenen els biaixos socials presents en els contextos d'extracció. Aquests biaixos i incorreccions es reprodueixen en els models i prediccions de la IA. Si la IA s'usa en la presa de decisions, pot donar com a resultat decisions injustes i l'amplificació de desigualats. Hi ha molts exemples d'aplicacions de tecnologies d'IA que reprodueixen biaixos, amb implicacions perjudicials per a col·lectius vulnerables (Eubanks, 2018). Això es veu exacerbat en models de ML que requereixen l'etiquetatge de les dades d'entrenament (Crawford, 2021), com ara en la identificació d'imatges o vídeos, ja que suposen la imposició de categories no neutres i la classificació subjectiva per part d'una mà d'obra sovint precaritzada, i de vegades exposada a material pertorbador, com ara escenes de violència (Gray i Suri, 2019).

Aquest model d'IA no només reprodueix estructures de classe, sinó que també amplifica estructures de discriminació i distribució desigual de capital a escala global. La capacitat d'acumular i processar dades massives per generar aquests models de ML requereix molts recursos naturals, tecnològics, econòmics i capital intel·lectual a l'abast de poques empreses, amb tendències

monopolístiques. Aquestes companyies estan concentrades als Estats Units i a la Xina, on són afavorides per polítiques proteccionistes dirigides a liderar la innovació en IA (Rikap i Lundvall, 2021). Finalment, les implicacions mediambientals són significatives, ja que la gran potència de càlcul necessària per processar dades i generar models consumeix molta energia i cal molta aigua per refrigerar els centres de dades. A més, els components tecnològics contenen minerals que sovint s'extreuen de països del Sud Global.

La IA s'ha convertit en una altra manera de fer política tot i que gens democràtica, particularment perquè, en un model econòmic en què preval la llei del més fort, la inèrcia de la versió extractiva —de dades, treball i recursos naturals— de la IA sembla difícil d'aturar. Caldran polítiques valentes i un activisme social determinat per afrontar els complexos reptes econòmics, epistèmics, ètics, socials, polítics i mediambientals que planteja la IA i repensar quines versions d'IA volem fomentar.

Referències

Barad, Karen M. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Durham, NC: Duke University Press.

Broussard, Meredith (2023). *More than a glitch: Confronting race, gender, and ability bias in tech*. Cambridge, Massachusetts: The MIT Press.

Crawford, Kate (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. New Haven: Yale University Press.

Eubanks, Virginia (2018). *Automating inequality*. Nova York: St. Martin's Press.

Gray, Mary L.; Suri, Siddharth (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Boston: Houghton Mifflin Harcourt.

Latour, Bruno (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford; Nova York: Oxford University Press.

Rikap, Cecilia; Lundvall, Bengt-Åke (2021). *The digital innovation race: Conceptualizing the emerging new world order*. Cham, Suïssa: Palgrave Macmillan.

Suchman, Lucy (2023). «The uncontroversial "thingness" of AI». *Big Data & Society*, 10 (2), p. 1-5.