bid textos universitaris de biblioteconomia i documentació
ISSN 1575-5886
DL B-19.675-1998

Número 41 (desembre 2018)

Intel·ligència artificial i transparència algorítmica: "It's complicated"

[Versión castellana][English version]

RAMON SANGÜESA
DESISLAB, Laboratori d'innovació social
Escola Elisava de Disseny i Enginyeria
rsanguesa@elisava.net

Opcions
<meta /> Metadades

DOI: https://dx.doi.org/10.1344/BiD2018.41.11

Citació recomanada

Sangüesa, Ramon (2018). "Intel·ligència artificial i transparència algorítmica : "It's complicated"". *BiD: textos universitaris de biblioteconomia i documentació*, núm. 41 (desembre) . <http://bid.ub.edu/41/sanguesa.htm>. DOI: http://dx.doi.org/10.1344/BiD2018.41.11 [Consulta: 20-12-2018].

Paraules clau: Privadesa, Algorismes, Accés a la informació, Processament de la informació, Intel·ligència artificial
Keywords: Privacy, Algorithms, Information access, Information processing, Artificial intelligence

Rebut: 08/09/2018. Acceptat: 13/09/2018.

Una part important de la societat en què ens movem es troba en l'economia d'informació, una economia de dades i algoritmes. Perquè les dades si no es tracten no són res. Aquesta economia és també una política basada en la implementació algorítmica i en xarxa de la societat de vigilància (Zuboff, 2016). Al seu torn, aquesta noció de *vigilància automatitzada* té una genealogia cientificotècnica que es pot rastrejar en la lògica del pensament cibernètic, una visió que cercava desenvolupar la ciència del control de màquines i organismes (Wiener, 1961).

Des de l'expansió del pensament cibernètic i les seves successives transformacions (Tiqqun, 2018), de la digitalització de quasi tot i de l'acceleració dels fluxos globals d'informació i del poder que se'n deriva, hem arribat a un moment de poder usar una capacitat de tractament de la informació més gran i més complexa. En aquesta acceleració hem passat per la interconnexió (internet), la fusió de continguts digitals globals (web), l'acumulació i l'encreuament de tota mena de dades (*big data*), una nova connexió de mitjans de captació de dades (internet de les coses) i ara som a l'autonomització dels processos d'anàlisi de les dades i de presa automàtica de decisions a partir dels models extrets d'aquesta anàlisi de dades. En altres paraules, a la interconnexió d'un gran nombre de sistemes d'intel·ligència artificial.

Si no fa ni seixanta anys la competició entre les nacions es plantejava en termes de capacitat de creació de coneixement científic i tecnològic, ara ens trobem que les potències globals amb aspiracions hegemòniques basteixen plans per excel·lir i avançar en termes d'intel·ligència artificial o, si voleu, en termes de codificació i automatització del coneixement a través de tecnologies de la cognició, la més representativa de les quals és la intel·ligència artificial. No és estrany que els que es pretenen grans poders hagin establert els seus plans nacionals estratègics d'intel·ligència artificial (Xina, 2018; NSTC, 2016; Villani, 2018; Hogarth, 2018).

Els incentius per recollir quantitats immenses de dades de tota mena són dobles: econòmics i polítics. El poder desfermat a través de les dades i el seu tractament han fet que dades i algoritmes, amb l'ajut de la intel·ligència artificial, i més en concret, de l'aprenentatge automàtic (*machine learning*), hagin passat d'apropar-se a les dades amb una primera intenció analítica (usar algoritmes per entendre què diuen les dades) a una de predictiva (anticipar) i, finalment, a una acció clarament prescriptiva (orientar la conducta de milions de persones a través del que s'ha esbrinat d'elles i del seu context amb models predictius i classificatoris). De fet, ara mateix conviuen totes aquests perspectives i es necessiten mútuament. Plegades, fan reals una forma d'economia i política basades en el control de la demanda i de la conformació de conductes individuals a escala planetària (Turkle, 2006).

Això és evident, per exemple, en els sistemes de recomanació que tan aviat ens presenten un nou servei com ens recomanen certes propostes culturals o ens predisposen envers determinades opcions polítiques. Els cercadors indueixen una funció semblant, a partir de la personalització dels resultats de cerca. De l'accessibilitat ampliada a determinats continguts culturals es passa a configurar o estabilitzar identitats culturals hegemòniques (Pasquale, 2015*a*). El fet que més del 80 % de les cerques es duguin a terme usant portals americans ha de tenir i té un paper en pautes de consum i configuració culturals, per exemple. I qui parla de cultura pot parlar de molts altres àmbits d'activitat com ha demostrat el

recent escàndol de manipulació política desfermat al votant de Facebook i Cambridge Analytica (Grassegger; Krogerus, 2018). Igualment cada cop és més clar que la combinació de dades i algoritmes indueix i propaga mecanismes de biaix, discriminació i tracte injust i desigual (Eubanks, 2018).

Davant d'aquest estat de coses, hi ha un cert consens que cal obrir altres escenaris alternatius. Aquest consens presenta diverses variants però, d'una forma o d'una altra, totes volen recuperar l'agència de certs subjectes davant l'abús de poder i asimetria de capacitats de captació de dades, del seu tractament, interpretació i decisió. Un punt comú a les diverses variants d'aquest consens és l'exigència de transparència.

La transparència de dades i d'algoritmes(*transparència algorítmica* per abreviar) implica la capacitat de saber quines dades es fan servir, com es fan servir, qui les fa servir, per a què les fan servir i com s'arriba a partir d'elles a prendre les decisions que afecten l'esfera vital de qui reclama aquesta transparència. Si una persona ha estat rebutjada en algun procés (per exemple, no rep una beca o un crèdit) hauria de saber a partir de quines dades s'ha pres aquesta decisió i com s'ha decidit excloure-la, que és una cosa diferent. Igualment, si un sistema de reconeixement l'ha classificat com a sospitós terrorista. Avui en dia, una esfera pública informada hauria d'estar composta per agents capaços d'esbrinar el subtext de l'univers algorítmic en què es desenvolupen els ciutadans com a subjectes econòmics i polítics.

Ara bé, aquest primer nivell de transparència, recollit en iniciatives molt recents i meritòries com el reglament europeu sobre protecció de dades, és només un primer pas i força feble. Habitualment s'argumenta que la transparència també ha d'incloure no només l'accés a les dades sinó al codi dels algoritmes que les tracten. Però això, tot i ser un pas necessari, no és del tot suficient. Per començar, obvia que aquesta mena d'accés no sempre és garantia de comprensió. I un codi al qual s'accedeix d'aquesta manera pot ser incomprensible no només pels ciutadans no experts, sinó també pels mateixos experts en dades, algoritmes i intel·ligència artificial. Molts algoritmes un cop "oberts", un cop accedim al seu codi, continuen sent autèntiques "caixes negres" (Pasquale, 2015*b*). Només cal considerar que el 95 % del codi de programació habitual no s'executa mai. Una de les raons és que els programadors anteriors el van desenvolupar en el seu moment per motius no sempre documentats en el codi mateix i, com a mesura de precaució, és millor no arriscar-se a efectes desconeguts provant de modificar-lo. Les rutines professionals en un entorn de gran competitivitat i celeritat pesen en aquesta pràctica que va amplificant la incomprensió del codi. Evidentment, aquest fet no exclou una voluntat malèvola en configurar un codi algorítmic d'efectes clarament nocius. El que és ben real és que, tant si hi ha mala intenció com bona en la voluntat darrere l'algoritme, la seva comprensió continua sent problemàtica.

En efecte, fins i tot si arribem a entendre aquest codi encara pot resultar més difícil comprendre com ha arribat a comportar-se d'una manera determinada. Això és especialment punyent en el cas de certs algoritmes d'aprenentatge automàtic que utilitzen xarxes neuronals (base de tot el que es coneix com a *deep learning*). És a dir, que si accedim al codi d'un sistema d'aquesta mena pot donar-se el cas que ni tan sols els experts que l'han programat puguin explicar per què mostra certs comportaments. En cas que fos possible entendre-ho, encara quedarien diverses altres coses a fer. Una seria poder comunicar de forma clara i comprensible a qui ho requereixi el "com" ha passat, el què ha passat, per què el sistema s'ha comportat d'aquella manera. Aquí hi ha un problema de llenguatge i de traducció molt gran. Entre experts podem comunicar-nos sobre un comportament anòmal o correcte derivat d'un procés d'entrenament d'un sistema d'aprenentatge automàtic però, com en comuniquem l'aplicació en un cas concret a un públic general? Una solució per a aquest tipus de situacions seria poder disposar de sistemes que interpretessin el comportament dels algoritmes de manera general o concreta i que en poguessin fornir explicacions en un altre llenguatge més proper al de la persona o col·lectiu afectats. És un camp ampli de recerca ple de dificultats (DARPA, 2018) que sempre està vorejant la regressió infinita d'explicacions d'un llenguatge al següent.

En cas que arribéssim a superar aquesta dificultat, encara quedaria un altre obstacle. Una cosa és descriure el mecanisme causal que ha portat a una decisió i una altra trobar-hi una justificació. Són dos plans diferents. Acceptaríem la justificació que "l'algoritme no ens ha donat un crèdit" perquè si ho feia augmentava el risc del banc? O que "l'algoritme no ens ha donat una beca" perquè així s'afavoreixen minories que sistemàticament queden fora del repartiment? Les justificacions apel·len a altres esferes de raonament no tècnic sinó ètic, moral o legal. Alguns aspectes com l'equitat (*fairness*) entren en joc aquí. I és un treball, després, difícil de retraduir a la implementació tècnica. Iniciatives com ara DTL, FATML o DAT (DTL, 2018; FATML, 2018; DAT, 2016) intenten trobar traduccions tècniques i metodològiques a la construcció i entrenament d'algoritmes d'intel·ligència artificial que permetin justificar-ne la transparència, l'equitat, la justícia i la traçabilitat. I no sempre se'n surten.

Amb freqüència moltes d'aquestes iniciatives acaben rendint-se davant el fet que la solució no rau només en les dades i en els algoritmes, sinó en el context de pràctiques socials i polítiques que els envolten. Diu molt del pes de la metàfora algorítmica en aquesta època que els grans centres de desenvolupament d'aquestes tecnologies hagin trigat força a desenvolupar una sensibilitat fora del fet tècnic (Boyd; Crawford, 2011) i proposessin marcs més complexos i interdisciplinaris on la solució tècnica acabaria implementant marcs ètics i legals que voldrien mitigar el risc. L'expansió de comitès ètics sobre intel·ligència artificial replica, potser, propostes semblants davant d'altres tecnologies de risc (Beck, 2002), com va succeir en el camp de la biologia i el desenvolupament de la bioètica corresponent. El fet que les associacions professionals dins l'àmbit de la intel·ligència artificial relacionin noves pràctiques de disseny que incorporen conceptes ètics mostra també una certa progressió respecte a l'estat de coses en què ens havíem situat (IEEE, 2018).

El que sembla clar és que encara estem en una fase relativament inicial de la consideració de la intel·ligència artificial com una altra tecnologia de risc, que ho és. En efecte, si seguim la caracterització d'Ulrich Beck (2002), és un tipus de tecnologia que assoleix un impacte sistèmic i distribueix els seus riscos i el seu increment de forma desigual en la població. És ben clar que els col·lectius fins ara més afectats per decisions guiades per sistemes d'intel·ligència artificial són precisament els més febles (Eubanks, 2018).

L'exigència de transparència sobre aquesta mena de sistemes i tecnologies, certament, pot començar a articular els debats de l'esfera pública i ho està fent. Ara bé, com hem esmentat més amunt, queda molt per recórrer perquè pugui haver-hi un debat informat i d'ampli abast que involucri tots els públics afectats i superi els problemes de comprensió, traducció i comunicació actuals que són d'envergadura.

Bibliografia

Beck, U. (2002). *La sociedad del riesgo: hacia una nueva modernidad*. Barcelona: Paidós Ibérica.

Boyd, D.; Crawford, K. (2011). "Six Provocations for Big Data". *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 21 Septembre 2011. Oxford Internet Institute.

DARPA (2018). *Explainable Artificial Intelligence*. <**https://www.darpa.mil/program/explainable-artificial-intelligence**>. [Consulta: 04/06/2018].

DAT (2016). *Workshop on Data and Algorithmic Transparency, 19 de Novembre de 2016*. Nova York, USA. <**http://datworkshop.org (http://datworkshop.org)** >. [Consulta: 12/05/2018].

DTL (2018). *Data Transparency Lab*. <**http://datatransparencylab.org (http://datatransparencylab.org)** >. [Consulta: 20/05/2018].

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

FATML (2018*). Fairness, Accountability and Transparency in Machine Learning*. <**www.fatml.org (http://www.fatml.org)** >. [Consulta: 20/05/2018].

Grassegger, H.; Krogerus, M. (2018). *Ich habe nur gezeigt, dass es die Bombe gibt*. Entrevista amb Michal Kosinski. <**https://www.dasmagazin.ch/2016/12/03/ich-habe-nur-gezeigt-dass-es-die-bombe-gibt/**>. [Consulta: 04/06/2018].

Hogarth, I. (2018). *AI Nationalism*. <**https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism**>. [Consulta: 20/06/2018].

IEEE (2018). *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, (EADv2).*

NSTC (2016). *Networking and Information Technology Research and Development Subcommittee. The National Artificial Intelligence Research and Development Strategic Plan.* National Science and Technology Council.

Pasquale, F. (2015*a*). "The Algorithmic Self". *The Hedgehog Review*. Vol. 17 No. 1 (Spring 2015).

Pasquale, F. (2015*b*). *The Black Box Society. The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Robert M. *et al*. (2012). "A 61-Million-Person Experiment in Social Influence and Political Mobilization". *Nature,* 489(7415) p. 295–298.

Tiqqun (2018). *La hipótesis cibernética*. <**https://tiqqunim.blogspot.com/2013/01/la-hipotesis-cibernetica.html**>. [Consulta: 07/06/2018].

Turkle, S. (2006). "Artificial Intelligence at Fifty: From Building Intelligence to Nurturing Sociabilities". Dartmouth Artificial Intelligence Conference, Hanover, NH, USA, 15 juliol, 2006, <**http://www.mit.edu/~sturkle/ai@50.html (http://www.mit.edu/~sturkle/ai@50.html)** >. [Consulta: 04/06/2018].

Villani, C. (2018). *For a Meaningful Artificial Intelligence. Towards a Fremch and European strategy.* <**https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf**>. [Consulta: 04/06/2018].

Wiener, N. (1961). *Cybernetics: Or Control and Communication in the Animal and the Machine.* 2nd ed. New York: MIT Press.

Xina. Consell d'Estat. (2017). *State Council Notice on the Issuance of the Next Generation Artificial Intelligence Development Plan.* Traducció a l'anglès per New America Foundation. <**https://www.newamerica.org/cybersecurity-initiative/blog/chinas-plan-lead-ai-purpose-prospects-and-problems**>. [Consulta: 03/03/2018].

Zuboff, S. (2016). "The Secrets of Surveillance Capitalism. Google as Fortune Teller." 3 de maig de 2016 *Frankfurter Allgemeine Zeitung*. <**http://www.faz.net/aktuell/feuilleton/debatten/the-digital-debate/shoshana-zuboff-secrets-of-surveillance-capitalism-14103616.html (http://www.faz.net/aktuell/feuilleton/debatten/the-digital-debate/shoshana-zuboff-secrets-of-surveillance-capitalism-14103616.html)** >. [Consulta: 08/04/2018].