Traccia dell'intervento di **Fosca Giannotti**[1],

Messaggio 1: **Politica, ricerca ed innovazione allineati per governare la rivoluzione digitale della società in una direzione che sia "utile alla persona ed alla società"**

L'insieme dei documenti ben riportati nella commissione parlamentare (iter e dossier) vanno di pari passo con i documenti di preparazione del nuovo programma di finanziamento della ricerca ed innovazione. Quello che abbiamo di fronte da parte dell'Europa è una strategia di attacco **per governare la rivoluzione digitale della società in una direzione che sia "utile alla persona ed alla società**" in riferimento ai valori europei. Quindi riflettere nella tecnologia i valori dell'Europa, e cioè quali? Il "welfare", la dignità umana, l'autonomia (il non essere manipolati e sfruttati), la diversità. Insomma **disegnare una proiezione digitale della società che potenzi i diritti degli individui e della collettività, che non li comprima**: Non sorveglianza e Non liberismo sfrenato

Occasione unica per politica e ricerca ed innovazione di avere un unico obiettivo: è una cosa molto difficile da fare.

Messaggio 2: **Fare dell'Europa il posto migliore dove i dati possano essere condivisi e riusati nel rispetto della privacy e della sicurezza**;

- l'obiettivo è ambizioso, "costruire un vero e proprio mercato unico dei dati - dove i dati personali e non personali, compresi i dati aziendali sensibili, sono sicuri e le aziende possono accedere ad una quantità quasi infinita di dati industriali di alta qualità, stimolando la crescita e creando valore, riducendo al minimo l'impronta di carbonio umano e ambientale. **Dove sono rispettati i diritti degli individui, dei consumatori e della concorrenza."**
  Ha l'aspetto positivo di mettere a lavoro le comunità di riferimento su 9 spazi, per definire obiettivi e standard, io vedo alcune criticità:
  1. Manca uno spazio importantissimo: i dati socio-economici, di solito questo spazio è la sfera della statistica ufficiale, ma anche questo dovrebbe essere uno spazio aperto.
  2. Rischio di fare di questo mercato gli "open-data" Europei, e non i big data. C'è il rischio che i dati che saranno resi disponibili siano di fatto degli aggregati relativi a dati prodotti e consumati in quei domini,…., ma spesso le scoperte interessanti sono dovute ai "microdati" ed al "riuso":
  3. es. non si cita da nessuna parte i dati degli operatori telefonici che sono un dato importantissimi per osservare fenomeni sociali: mobilità umana, fenomeni migratori, indicatori del benessere e dello sviluppo economico, epidemie
  4. dati sanitari. Durante Covid abbiamo visto quanto sia stata importante la condivisione dei dati fatta dagli scienziati. Quanto fosse strategico per combatter al malattia la condivisione tra i medici, ebbene, **abbiamo chiuso un paese ma non siamo stati in grado di aprire i dati sanitari per studiare la malattia**. Es. UK https://opensafely.org/,

RACCOMANDAZIONI:

- **condivisione responsabile di dati micro**: in questi spazi dei dati si creino delle esperienze interspazio e si condividano anche dati micro. SoBigData (si veda flyer SoBigData) è esempio di cosa significhi costruire un ecosistema dove si permette la sperimentazione di usi diversi dei dati per il bene comune in modo responsabile.
- **federazione di infrastrutture dati** (EOSC) e CLOUD che promuovano accesso controllato, responsabile e decentralizzato, e modelli analitici che portano gli algoritmi dove sono i dati e non viceversa.

[1] Dirigente di Ricerca CNR, Istituto di Scienza e Tecnologia dell'informazione "A. Faedo", Pisa; Coordinatrice della RI europea SoBigData.eu, **Social Mining e Big Data Analytics**, 16 Milioni di finanziamento, dal 2015 al 2024, da 11 a 32 partners; PI ERC grant Explainable AI, PI del nodo CNR di Humane-AI-net e TAILOR (ICT-48) ed AI4Eu; Membro dello Shadow Committee per HE nel cluster4-Digital, Industry and Space e del gruppo di lavoro nell'area HPC e BigData per il PNR.

Messaggio 3: **AI umana e sociale: la spiegazione come strumento abilitante**

- Abbiamo fatto notevoli progressi negli algoritmi di apprendimento da dati: riconoscimento di immagini, da testo etc, ma è ancora tutto da esplorare il tema della interazione **persona-macchina nei processi decisionali critici**:
    - Stiamo rendendo la persona più intelligente? Il risultato congiunto persona macchina è un risultato migliore? Il risultato collettivo di tante persone/macchine è un risultato migliore? Es. le echo-chambers. (si vedano i due documenti pubblicati su Gnosys)
    - Abbiamo bisogno di costruire processi interattivi basati sulla fiducia del funzionamento, ma anche di metodi di interazione che sfidino l'approccio cognitivo, non lo deprimano. Che aiutino ad immaginare il "*perché*", il "*perché no*" ed anche il "*cosa succederebbe se*".
    - Abbiamo bisogno di inserire nelle piattaforme di raccomandazione principi che alimentino la diversità e la democrazia. (documento AI&Reccomendation)

RACCOMANDAZIONI:

Dal punto di vista tematico, è fondamentale affrontare le seguenti sfide,

- un'intelligenza artificiale affidabile e spiegabile per combattere nuove forme di discriminazione e manipolazione e dare potere ai cittadini;
- un'intelligenza artificiale consapevole della società per combattere la polarizzazione e la disuguaglianza e per promuovere la diversità e l'apertura.

Dal punto di vista strumentale, è importante rendersi conto che il panorama scientifico e tecnologico non è ancora pienamente maturo per affrontare tutte le sfide aperte qui discusse. È quindi necessario un mix di politiche che affrontino il problema a tre livelli:

- un coraggioso investimento nella ricerca fondamentale e applicata nell'IA centrata sulla persona e la società;
- un coraggioso investimento nella creazione di competenze di AI e Bigdata a tutti i livelli (compreso il long-life learning)
- un coraggioso investimento dell'UE e dei paesi nella creazione di nuove piattaforme online e servizi digitali che incorporino meccanismi di IA antropocentrici (e/o che sostengano la scala delle iniziative coerenti esistenti);
- un insieme coerente di regolamenti UE riguardanti l'IA, i big data e i servizi digitali, progettati non solo per cogliere le opportunità e mitigare i rischi, ma anche per ispirare la ricerca e lo sviluppo nell'IA, i big data e le piattaforme digitali verso una società inclusiva, uguale e diversificata.

CONSORTIUM

Consiglio Nazionale delle Ricerche

Università di Pisa

The University Of Sheffield.

Fraunhofer FHR

IMT Istituto Alti Studi Lucca

Leibniz Universität Hannover

King's College London

Scuola Normale Superiore

Aalto University

ETH Zürich

Tartu Ülikool

TU Delft Delft University of Technology

EGI

Paris School of Economics

Roma Tre Università degli Studi

OpenAIRE

cnrs

RIE

CEU Central European University

Universitat Rovira i Virgili

CSD Center for the Study of Democracy

KTH Vetenskap och Konst

Barcelona Supercomputing Center Centro Nacional de Supercomputación

upf. Universitat Pompeu Fabra Barcelona

Università degli Studi dell'Aquila

eliante

Carlo Ratti Associati

Nubisware

University of Amsterdam

STACC Turning data into value

Scuola Superiore Sant'Anna Pisa

10 ML EURO

31 PARTNERS

+4700 USERS

+205 RESOURCES

120 COMPANIES

+900 STUDENTS

+170 HIGH PROFILE PUBLICATIONS

120 PILOT PROJECTS

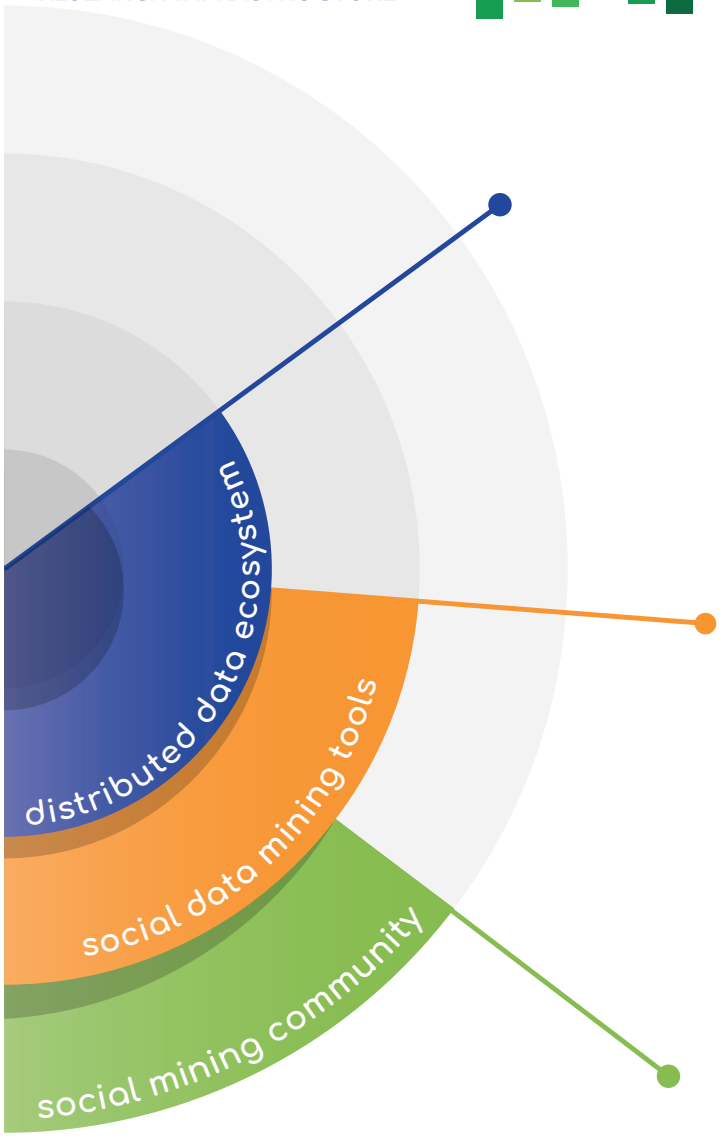
15 TRANSNATIONAL ACCESS SITES

Transnational access:
On-site access services provided by seven national infrastructures. The access is granted through project calls fundend by EU.

Social Mining & Big Data Analytics

SoBigData

RESEARCH INFRASTRUCTURE



SoBigData

SoBigData

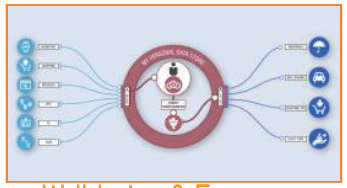
www.sobigdata.eu

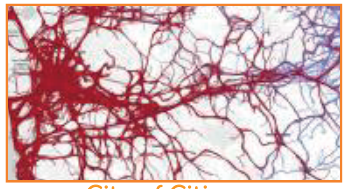
THE SOBIGDATA ECOSYSTEM

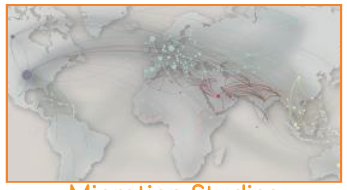
A distributed pan-European multi-disciplinary research infrastructure for big social data analytics, a cross-disciplinary European research community, a research infrastructure designed to promote large-scale, interdisciplinary social data mining experiments repeatable and open-science oriented.

6 EXPLORATORIES


Societal Debates

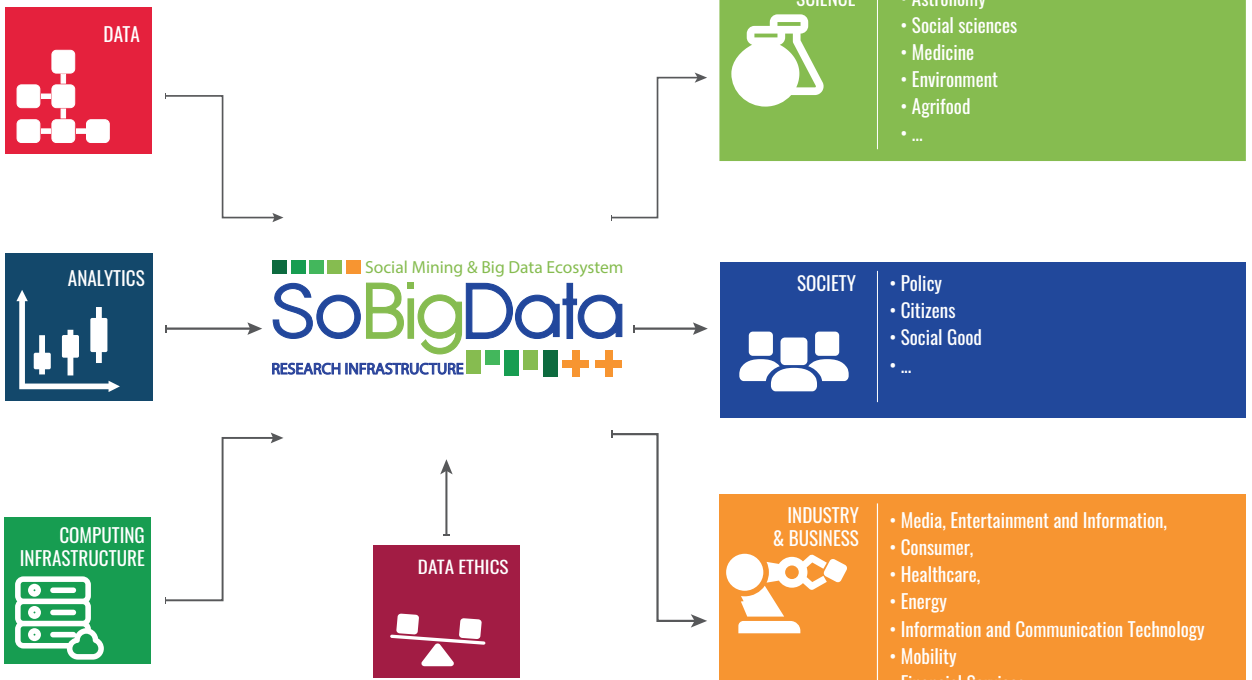

Well-being & Economy


City of Citizens


Migration Studies


Explainable Machine Learning


Sport Data Science



INFO:
Project No: 871042 | Program: H2020 | INFRAIA 2018/2019
Duration: 01/2020 - 12/2023

SoBigData++ receives funding from the European Union's Horizon 2020 research and innovation programme.

SoBigData covers different thematic research environments, called exploratories where researchers do analyses and use innovative methodologies.

5 ERC GRANTS

Five SoBigData scientists are PIs of ERC grants in the field of artificial intelligence.

IL BIAS DELL'ALGORITMO E LA POLARIZZAZIONE DELLE OPINIONI

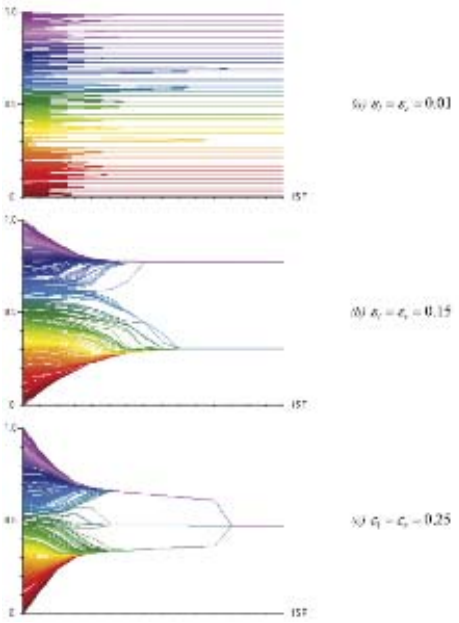
Dino Pedreschi – Fosca Giannotti

Per millenni l'uomo ha formato le proprie convinzioni grazie al confronto con coloro che lo circondano. Tali relazioni, sulla base del livello di fiducia reciproca, hanno un ruolo fondamentale nel processo che ci induce a rafforzare o cambiare le nostre opinioni. Con la diffusione dei social media assistiamo a un processo di sfruttamento su vasta scala, soprattutto per fini di business, del meccanismo del bias di conferma che ci espone soprattutto alle informazioni che più ci piacciono e ci induce a interagire con chi è più vicino alle nostre posizioni. In che modo tali meccanismi alterano le tradizionali modalità di formazione delle opinioni? Che impatto hanno sulla frammentazione, la polarizzazione e la radicalizzazione delle idee? L'articolo sintetizza gli esiti dell'attività svolta costruendo modelli che imitano la formazione di opinioni nella società e suggerisce alcune iniziative che, anche mediante il ricorso all'Intelligenza Artificiale, potrebbero favorire il ritorno a un contraddittorio aperto alle diversità.

Il processo di formazione e diffusione delle opinioni in una società è un fenomeno complesso, che dipende da molteplici fattori, come le preferenze personali, la cultura o l'istruzione. L'interazione con i nostri pari gioca un ruolo rilevante: discutiamo questioni importanti con gli amici e cambiamo o rafforziamo le nostre convinzioni quotidianamente. Un altro effetto decisivo è, ovviamente, quello dei media: le informazioni esterne ci raggiungono e ci influenzano costantemente. La scelta degli individui con cui ci rapportiamo e delle notizie che consumiamo è dunque cruciale nella formazione dei nostri convincimenti. Nell'ultimo decennio, i modelli di interazione sono cambiati radicalmente. Fino ai primi anni del nuovo millennio tendenzialmente si leggeva il quotidiano locale, si guardava il telegiornale e si discuteva con amici e conoscenti, mentre oggi ci si rapporta a grandi distanze e si leggono notizie in tutto il mondo attraverso i giornali online e le reti sociali.

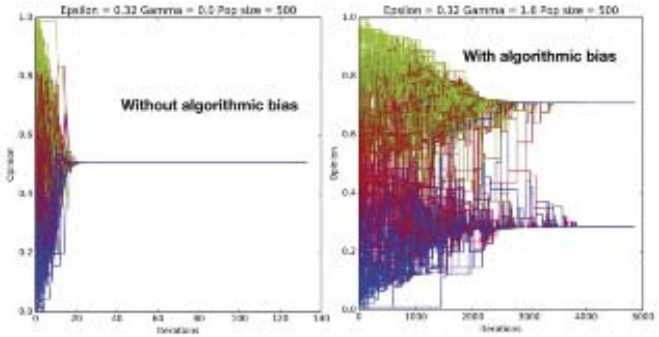
I social media, in particolare, sono sempre più utilizzati per condividere idee, ma anche, come mostra il rapporto Reuters 2018 sulle D*igital* N*ews*, per leggere e scambiare notizie. Ciò significa che gli interlocutori e i contenuti con cui interagiamo, insieme agli effetti esterni nella dinamica delle opinioni, diventano suscettibili alle influenze derivanti dai meccanismi d'interazione delle piattaforme di social media in uso, costruite con un obiettivo di marketing: massimizzare il numero di utenti e il tempo che questi trascorrono sulla piattaforma, ovvero catturarne l'attenzione. Per conseguire questo risultato, le informazioni che li raggiungono non vengono selezionate casualmente. Esiste un preciso *bias algoritmico*: siamo sistematicamente esposti alle notizie relative agli argomenti che più ci piacciono e alle opinioni di coloro che ne hanno di più vicine alle nostre, così da essere spinti a interagire coi contenuti proposti e a tornare sulla piattaforma. In fin dei conti, il marketing sfrutta da sempre un meccanismo cognitivo che è connaturato con l'essere umano, il *bias di conferma*, così definito da Wikipedia: «È un processo mentale che consiste nel ricercare, selezionare e interpretare informazioni in modo da porre maggiore attenzione, e quindi attribuire maggiore credibilità a quelle che confermano le proprie convinzioni o ipotesi, e viceversa, ignorare o sminuire informazioni che le contraddicono. Il fenomeno è più marcato nel contesto di argomenti che suscitano forti emozioni o che vanno a toccare credenze profondamente radicate». I meccanismi del marketing, commerciale, informativo o politico che sia, fanno leva sul bias di conferma, perché ciascuno di noi è gratificato dal sentirsi dare ragione, dall'incontrare soggetti o contenuti che confermano le proprie convinzioni. Proporre di interagire con persone e contenuti vicini alle convinzioni degli utenti e alle preferenze dei consumatori, assicura quindi un maggiore *click rate*, una maggiore probabilità di attrarre acquisti, attenzione, *like*. In ultima analisi, maggiore profitto. Una domanda sorge naturale: come interferisce questo bias con la formazione delle nostre opinioni? Cambiamo mai idea, o continuiamo perennemente a rinforzare la nostra? Rimaniamo chiusi nelle nostre piccole bolle informative? Insieme ad alcuni colleghi del nostro Knowledge Discovery and Data Mining Laboratory del Cnr e dell'Università di Pisa e del Dipartimento di Network and Data Science della Central European University di Budapest, stiamo cercando di rispondere a questi interrogativi costruendo modelli di dinamica delle opinioni che imitano la formazione di opinioni nella società. Sulla base della loro evoluzione, un gruppo di persone può raggiungere il consenso – una opinione condivisa da tutti – su un determinato argomento, ovvero possono emergere frammentazione e polarizzazione delle opinioni in gruppi segregati e radicalizzati, impermeabili al dialogo. Questo processo può essere modellato in modo idealizzato rappresentando le opinioni su un tema controverso, ad esempio i vaccini, con valori numerici associati a ciascun individuo in una folla, in genere un valore compreso fra 0 (totale contrarietà) e 1 (totale favore). Nel tempo, gli individui nella folla interagiscono fra di loro, modificando i propri convincimenti. La popolazione, di solito, parte da una configurazione in cui le opinioni sono distribuite in modo random ed evolve nel tempo per raggiungere il consenso – una opinione condivisa da tutti – oppure una situazione frammentata in «partiti» contrapposti, ciascuno con la propria convinzione.

In un modello molto studiato per queste simulazioni sociali, noto come modello di Deffuant dal nome del suo ideatore, a ogni istante due individui scelti a caso nella folla interagiscono fra di loro, scambiandosi le proprie opinioni e avvicinando le proprie posizioni; ma questo avviene solo se la loro opinione non è troppo lontana, ovvero al di sotto di una soglia che ne rappresenta l'«apertura mentale», ovvero il tasso di fiducia. Intuitivamente, la soglia rappresenta la distanza massima fra la mia opinione e quella di un altro, oltre la quale non sono disposto ad ascoltarlo perché «troppo diverso» da me. Se lasciamo il modello di Deffuant libero di evolvere attraverso la ripetizione delle interazioni, fino a raggiungere una situazione stabile, osserveremo che si formano gruppi polarizzati su opinioni diverse se il tasso di fiducia è basso; le diverse «fazioni» saranno tanto più numerose quanto più è basso il tasso di fiducia. Viceversa, se la fiducia è sufficientemente ampia, emerge un consenso generalizzato (vedi immagine in basso). Un'allegoria piuttosto efficace del senso compiuto della democrazia, intesa non come dittatura della maggioranza ma come ricerca costruttiva del consenso: l'apertura all'ascolto delle ragioni dell'altro facilita il raggiungimento di un compromesso in cui tutti rinunciano a qualcosa per ottenere ciò che per tutti è irrinunciabile.



L'evoluzione delle opinioni in tre simulazioni con diverse soglie di fiducia ε: bassa (ognuno rimane della propria opinione), media (si formano due fazioni contrapposte) e ampia (si raggiunge il consenso).

In una ricerca recentemente pubblicata sulla rivista «PlosOne»[1], abbiamo apportato una modifica, apparentemente piccola, per introdurre in questo modello il meccanismo del bias algoritmico delle piattaforme. Invece di selezionare le coppie di individui che interagiscono in modo del tutto casuale, queste vengono selezionate favorendo la vicinanza delle opinioni: una persona ha maggiori probabilità di trovarsi a interagire con chi ha un'opinione vicina alla propria, invece che con chi ne ha di lontane. Più vicina è l'opinione di due individui, più è probabile che la piattaforma suggerisca ai due di interagire (figura in basso).



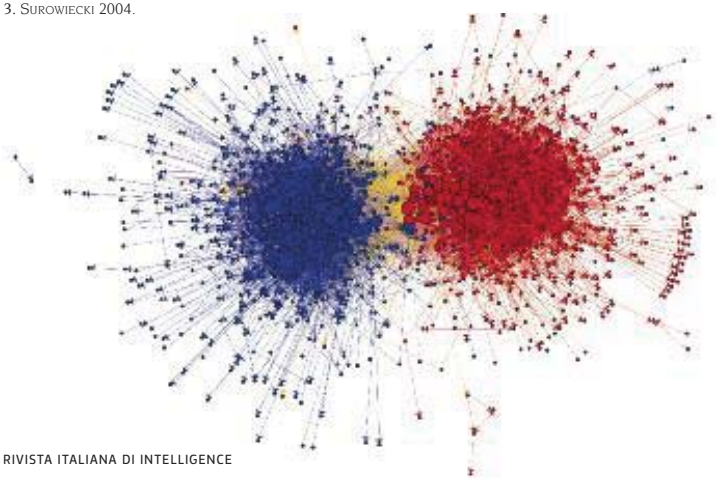
Simulazione della formazione delle opinioni senza e con bias algoritmico, a parità di soglia di fiducia. **Nella pagina successiva,** La polarizzazione dei blog politici statunitensi durante la campagna elettorale presidenziale del 2004 illustrata attraverso la visualizzazione della rete dei link fra i diversi blog liberali (blu) e conservatori (rossi).

Le simulazioni del modello esteso con il bias algoritmico suggeriscono in modo evidente che le piattaforme online hanno un effetto decisivo sulla formazione dell'opinione e del consenso. Il risultato più scioccante è che il numero di gruppi con opinioni diverse aumenta, a parità di «apertura mentale» della popolazione, con l'aumentare dell'intensità del bias algoritmico. In altre parole, le piattaforme online possono favorire la frammentazione delle opinioni provocando la nascita di «partiti» contrapposti anche quando, in assenza di bias, la popolazione arriverebbe al consenso – una divisione quindi totalmente artificiale, indotta dal meccanismo della piattaforma che inibisce il confronto fra posizioni diverse. In pratica, il bias algoritmico ha portato a un aumento della polarizzazione e della radicalizzazione delle opinioni, amplificando il già potente meccanismo cognitivo che abbiamo incorporato, il bias di conferma.

1. Sîrbu et al. 2019.

Il secondo effetto del bias algoritmico è sulla velocità della formazione delle opinioni: i cambiamenti di parere degli individui sono molto più lenti quando esso agisce, e quindi la popolazione, nel suo complesso, impiega un tempo molto più lungo per stabilizzarsi in una situazione delineata. Anche quando si arriva a ottenere il consenso, il tempo per raggiungerlo diventa assai lungo. In pratica, ciò significa che potrebbero volerci anni prima che le persone concordino su un problema, rimanendo a lungo in uno stato di confusione e frammentazione. Pur nella semplicità del modello che abbiamo descritto, questi risultati evidenziano che il bias algoritmico delle piattaforme può influenzare profondamente i risultati dei dibattiti pubblici e della formazione del consenso nella società. Un meccanismo apparentemente innocuo, da sempre utilizzato nel marketing, se utilizzato su vasta scala ha un «effetto di rete» potenzialmente distruttivo, capace di creare e rafforzare le «camere dell'eco», le bolle informative, radicalizzare il dibattito pubblico e destabilizzare la democrazia. Un fenomeno a cui stiamo assistendo con crescente apprensione, registrato da studiosi ed esperti delle dinamiche sociali di tutto il mondo[2] (figura in basso). Probabilmente questo effetto di polarizzazione è ancora più pericoloso per la nostra democrazia delle fake news e del fenomeno degli odiatori (*hate speech*), in quanto questi ultimi sono, nella maggior parte dei casi, un effetto delle bolle artificialmente rinforzate dal bias. Il nostro risultato evidenzia come le piattaforme per il consumo di notizie e la conversazione pubblica dovrebbero, al contrario di ciò che accade, mettere in atto meccanismi algoritmici che cerchino di mitigare il bias di conferma degli utenti anziché rafforzarlo, e trovare modi per esporre gli utenti a contenuti e individui con opinioni diverse, aiutando a farci un'idea più consapevole sulle questioni controverse o complesse, osservandole da più punti di vista. Un diverso ecosistema dell'informazione mirato alla diversità sarebbe anche essenziale per preservare e nutrire la nostra intelligenza collettiva, o la saggezza della folla, a qualsiasi scala, dalle comunità alla cittadinanza globale. Come dimostrato da una lunga storia di studi di ricerca e di evidenze empiriche[3], una moltitudine dotata di una sana diversità e

2. Schmidt et al. 2018.
3. Surowiecki 2004.



indipendenza di pareri può rispondere intelligentemente a domande difficili (ad esempio prevedere un risultato dall'esito incerto), ma la diversità e, quindi, l'intelligenza collettiva viene rapidamente minata dall'influenza eccessiva e dalla polarizzazione. Il bias delle piattaforme è, insomma, un pericolo, e sono necessarie misure per arrestarne o almeno mitigarne gli effetti, o addirittura invertirli, come ricercatori stanno cercando di fare[4]. Nel frattempo, gli utenti potrebbero essere informati del modo in cui notizie sono veicolate sui *feed* delle piattaforme e del fatto che ciò potrebbe influenzarne le opinioni, e forse i meccanismi implementati dalle piattaforme potrebbero essere lentamente ritirati. La conseguenza dei risultati scientifici discussi è che le attuali piattaforme per l'accesso all'informazione, sia tradizionali che i social media, basate sul servizio 'gratuito' pagato dalla pubblicità e dal marketing personalizzato, non sono (e non possono essere) funzionali a contrastare la polarizzazione e la radicalizzazione, e a promuovere la diversità e il pluralismo. Lo scopo ultimo del modello di business mirato a catturare l'attenzione degli utenti è, per definizione, in conflitto con quello di promuovere la diversità e il pluralismo. Pertanto, un diverso *ecosistema dei media*, volto ad aiutare i cittadini a confrontarsi con narrazioni e contenuti diversi da quelli preferiti o consueti, dovrebbe implementare strategie intelligenti per connettere visioni opposte e alternative. Dovrebbe basarsi in modo trasparente su bias algoritmico con finalità opposta a quella attuale, strumenti di I*ntelligenza Artificiale* (AI) che spingono verso la diversità e le visioni pluralistiche. Dovrebbe essere una piattaforma pubblica aperta, un bene comune, indipendente dai media, dalle aziende e dai governi, che persegue la finalità di preservare la socio-diversità, essenziale per la democrazia. Questo progetto è alla portata dello stato attuale di avanzamento della tecnologia digitale. La AI e la scienza dei dati non sono tecnologie neutre, possono essere utilizzate per scopi buoni o cattivi; sta a noi, al nostro dibattito democratico, all'Europa in particolare, decidere quali valori instillare *by-design* in queste tecnologie per creare una vera ecologia dell'informazione[5].

4. Aslay et al. 2018; Garimella et al. 2017.
5. De Biase 2016.

BIBLIOGRAFIA

L.A. Adamic – N. Glance, *The political blogosphere and the* 2004 US *election: divided they blog*, Proceedings of the 3rd International Workshop on Link discovery 2005, pp. 36-43.
C. Aslay et al., *Maximizing the diversity of exposure in a social network*, IEEE International Conference on Data Mining 2018, pp. 863-868.
L. De Biase, *Homo pluralis. Esseri umani nell'era tecnologica*, Codice, Torino 2016.
K. Garimella et al., *Reducing controversy by connecting opposing views*, Proceedings of the Tenth ACM International Conference on Web Search and Data Mining 2017, pp. 81-90.
R. Hegselmann – U. Krause, *Opinion dynamics and bounded confidence: models, analysis and simulation*, «Journal of Artificial Societies and Social Simulation» V (2002) 3.
A.L. Schmidt et al., *Polarization of the vaccination debate on Facebook*, National Center for Biotechnology Information 36 (2018) 25, pp. 3606-3612.
A. Sîrbu et al., *Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model*, «PLoS ONE» 2019.
J. Surowiecki, *The wisdom of crowds*, Anchor Books, New York 2004.

BIBLIOGRAFIA J. ADEBAYO ET AL., *Sanity Checks for Saliency Maps*, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal 2018. ALGORITHMWATCH, *Automating Society. Taking Stock of Automated Decision-Making in the EU*, Report 2019: <https://algorithmwatch.org/en/automating-society/> [20-5-2019]. A. CALISKAN-ISLAM ET AL., *Semantics derived automatically from language corpora necessarily contain human-like biases*, Cornell University, Ithaca 2016. C. CARTER ET AL., *The Credit Card Market and Regulation: In Need of Repair*, «North Carolina Banking Institute» XXIII (2006) 10, pp. 23-56: <http://scholarship.law.unc.edu/ncbi/vol10/iss1/4> [20-5-2019]. M. CRAGLIA ET AL., *Artificial Intelligence: A European Perspective*, Publications Office of the EU, Brussels 2018: <http://publications.jrc.ec.europa.eu/repository/bitstream/JRC113826/ai-flagship-report_on line.pdf> [20-5-2019]. EUROPEAN COMMISSION, *Artificial Intelligence for Europe*, Communication SWD (2018) 137 final, Brussels 2018: <http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51625> [20-5-2019]. EUROPEAN COMMISSION (High-Level Expert Group on Artificial Intelligence), *Draft Ethics Guidelines for Trustworthy AI*, Brussels 2018. R. GUIDOTTI ET AL., *A Survey of Methods for Explaining Black Box Models*, «ACM Computing Surveys» (CSUR) LI (2018) 5, pp. 2-45. L. HU ET AL., *Interpretable Recommendation via Attraction Modeling: Learning Multilevel Attractiveness over Multimodal Movie Contents*, Proceedings of the 27th International Joint Conference on AI – IJCAI 2018, «ACM Digital Library» 2018, pp. 3400-3406. Y. LECUN ET AL., *Deep learning*, «Nature» (2015) 521, pp. 436-444. S. LOWRY – G. MACPHERSON, *A blot on the profession*, «British medical journal» CCXCVI (1988) 6623, pp. 657-658. F. PASQUALE, *The black box society*, Harvard University Press, Cambridge 2015. D. PEDRESCHI ET AL., *Discrimination-aware Data Mining*, Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining, «ACM Digital Library» (2008), pp. 560-568. D. PEDRESCHI ET AL., *Open the Black Box: Data-Driven Explanation of Black Box Decision Systems*, Cornell University, Ithaca 2018: <https://arxiv.org/abs/1806.09936 > [20-5-2019]. M.T. RIBEIRO ET AL., *"Why should I trust you?" Explaining the predictions of any classifier*, Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining, «ACM Digital Library» (2016), pp. 1135-1144.

«EXPLAINABLE AI»

APRIRE LE SCATOLE NERE PER UNA INTELLIGENZA ARTIFICIALE UMANA

FOSCA GIANNOTTI – DINO PEDRESCHI

La crescita esponenziale delle capacità dei modelli di AI consentono di arrivare a livelli di valore e generalizzazione mai registrati prima. È però cresciuta anche l'opacità di tali esempi e la loro natura di black box *che rende difficile, anche per gli esperti, spiegarne le conclusioni. Questo può rappresentare una criticità in termini tecnologici e sociali, essendo reale il rischio, come dimostrato da recenti episodi, di addestrare sistemi compromessi dal pregiudizio e dalla discriminazione. È pertanto ancor più vero il principio in base al quale «apprendere dalle tracce digitali delle decisioni del passato può portare a incorporare in modo invisibile nei modelli risultanti i pregiudizi esistenti, perpetuandoli». Riusciremo ad allineare gli algoritmi con i valori e le aspettative umane? Partendo da una serie di esempi concreti e dalle loro esperienze di ricerca, gli autori stimolano una riflessione sulla necessità di porre l'uomo e i suoi valori al centro dello sviluppo dei sistemi di AI.*

All'inizio del 2017 abbiamo collaborato alla stesura del *position paper* dell'Accademia dei Lincei sulla Data Science in preparazione del G7 Academies Meeting, svoltosi nel mese di marzo dello stesso anno[1]. In quel documento si affermava:

> Data science emerged due to three concurring factors, unleashed by the digital transformation of society: i) the advent of Big Data; ii) the advances in data analysis and machine learning techniques; and iii) the advances in scalable high-performance computing infrastructures. The three factors together are an explosive mix: big data provide the critical mass of factual examples to learn from; analytics are able to produce predictive models and behavioral patterns

1. PEDRESCHI ET AL. 2018.

> from these data; scalable computing platforms make it possible to ingest data and perform analytics. As a matter of fact, the three factors have come to maturity together, in the last few years. The spectacular advances of artificial intelligence, such as automated language understanding and translation, image recognition... are essentially successes of data science, explained by the synergic effect of the three factors above: rich data, data-driven models, computing power.

La nuova primavera della I*ntelligenza Artificiale* (AI), seguita al lungo inverno originato dalla disillusione degli anni Ottanta, è il regalo della scienza dei dati che, finalmente, riesce ad affrontare con successo la comprensione e la traduzione dei testi, il riconoscimento delle immagini e ad assolvere altri compiti 'intelligenti'.
È accaduto grosso modo a partire da dieci anni fa, quando ci si è accorti che alcuni modelli di apprendimento noti da tempo, come le reti neurali artificiali, fino ad allora inefficaci, se equipaggiate da un enorme numero di variabili interne (neuroni) e parametri associati, e addestrate adeguatamente, compiono improvvisamente un salto di qualità e sono in grado di generalizzare, dai pixel delle immagini o dai termini dei testi di esempio, i 'concetti' generali che permettono di riconoscere, classificare, prevedere con accuratezza nuove immagini e nuovi testi. Una modalità di apprendimento che richiede nuovi paradigmi di calcolo ad alte prestazioni. Questo quadro, I*ntelligenza Artificiale* = B*ig Data* + M*achine Learning*, a distanza di due anni, è ampiamente confermato.
Alla recente conferenza Aaai 2019, l'80% dei 7000 (!) lavori presentati e dei 1000 (!) accettati hanno a che fare con M*achine Learning* (ML), D*eep Learning* (DL), B*ig Data* (BD) e applicazioni della Data Science. In questo quadro si evidenzia tutta la forza e insieme la debolezza della situazione attuale. Da un lato, cresce la capacità dei modelli di DL, sempre più complessi, di generalizzare da dati di allenamento sempre più grandi e di maggiore qualità, e ciò spiega gli straordinari progressi nella visione robotica, nella comprensione del testo e del parlato, nella traduzione automatica, nella diagnosi, nella valutazione del rischio. Dall'altro lato, si amplia l'opacità e la natura di *black box* dei modelli di AI, insieme con il rischio di creare sistemi che nemmeno gli esperti riescono a comprendere. Lo sviluppo della AI, infatti, ha enfatizzato la qualità in termini di accuratezza e generalizzazione, invece che di comprensibilità e validazione. Dunque, grandi opportunità di progresso ma anche nuove criticità.

Parafrasando Frank Pasquale, autore di T*he black box society* (2015), abbiamo visto diffondersi algoritmi sempre più opachi, sempre più spesso creati dal DL, per inferire tratti intimi delle persone, come il rischio creditizio o assicurativo, lo stato di salute, il profilo della personalità, la propensione al crimine. Black box che, osservando le caratteristiche degli utenti, ne pronosticano una classe, un giudizio, un voto e suggeriscono decisioni; senza però spiegarne la ragione.
Non è solo un problema di trasparenza. Il ML opera su esempi ricostruiti sulla base delle tracce digitali delle attività degli utenti: movimenti, acquisti, ricerche online, opinioni espresse sui social... E quindi i modelli risultanti ereditano i pregiudizi e i difetti – *i bias* – che sono celati nei dati di allenamento, nascondendoli a loro volta negli algoritmi di decisione che rischiano di suggerire scelte ingiuste, discriminatorie o semplicemente sbagliate, e all'insaputa del decisore e del destinatario della decisione. Se un *chat box*, una AI che conversa con gli utenti dei social, impara a conversare dagli esempi sbagliati, ad esempio razzisti, sarà razzista a sua volta – e i suoi creatori dovranno sbrigativamente tacitarla. Molti casi già accaduti, come questo del Twitter bot Tay, ci ammoniscono che delegare scelte ad algoritmi black box è una pessima idea.

THE DANGER OF BLACK BOXES

Delegare decisioni a scatole nere può essere un errore, come illustrano i seguenti casi.

– Compas, proprietà di Northpointe Inc., è un modello predittivo del rischio di recidiva criminale, utilizzato fino a poco tempo fa da varie corti di giustizia statunitensi a supporto delle decisioni dei giudici su richieste di scarcerazione. I giornalisti di propublica.org hanno raccolto migliaia di casi d'uso del modello e dimostrato che esso ha un forte bias razzista: ai neri che in realtà non delinqueranno nuovamente è assegnato un rischio doppio rispetto ai bianchi nelle stesse condizioni[2]. Il modello, sviluppato con tecniche di ML, ha presumibilmente ereditato i bias presenti nello storico delle sentenze e risente del fatto che la popolazione carceraria statunitense sovrarappresenta i neri rispetto ai bianchi[3];

2. <www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [21-2-2019].
3. *Ibidem*

- le tre principali agenzie per il rischio creditizio degli Stati Uniti, Experian, TransUnion e Equifax sono spesso discordanti. In uno studio su 500.000 casi, il 29% dei richiedenti credito ha avuto una valutazione di rischio con differenze di oltre 50 punti da parte delle tre compagnie, che può significare decine di migliaia di dollari di differenza sugli interessi complessivi. Una variabilità così ampia suggerisce ipotesi di valutazione molto diverse, oltre che opache, oppure una forte arbitrarietà[4];
- negli anni 70 e 80 la St. George's Hospital Medical School di Londra ha usato un software per filtrare le domande di lavoro che è stato in seguito giudicato fortemente discriminatorio verso le donne e le minoranze etniche, inferite dal cognome e dal luogo di nascita. La discriminazione automatica non è un fenomeno nuovo e non è necessariamente dovuta al ML...[5];
- un classificatore basato sul DL può essere molto accurato rispetto ai dati di allenamento, e al tempo stesso del tutto inaffidabile, ad esempio perché ha imparato da dati di cattiva qualità. In un caso di riconoscimento di immagini rivolto a distinguere lupi da husky in un grande dataset di immagini, la black box risultante è stata vivisezionata dai ricercatori[6] per scoprire che la decisione di classificare un'immagine come 'lupo' era basata unicamente... sulla presenza di neve nello sfondo! La colpa, ovviamente, non è del DL, ma dalla scelta accidentale degli esempi di allenamento in cui, evidentemente, ogni lupo era stato fotografato sulla neve. Quindi un husky sulla neve è automaticamente classificato come lupo. Ora trasportiamo questo esempio sul sistema di visione della nostra auto a guida autonoma: siamo sicuri che sarà in grado di riconoscere correttamente ogni oggetto intorno a noi?
- vari studi, come quello richiamato in nota[7], mostrano che i testi sul web (ma anche sui media in generale) contengono bias e pregiudizi, come ad esempio il fatto che i nomi di bianchi sono più spesso associati a vocaboli con una carica emotiva positiva, mentre i nomi di persone di colore sono più spesso associati a vocaboli con una carica emotiva negativa. Quindi, i modelli allenati su testi per l'analisi del *sentiment* e delle opinioni hanno forte probabilità di ereditare gli stessi pregiudizi;

4. Carter et al. 2006.
5. Lowry – Macpherson 1988.
6. Ribeiro et al. 2016.
7. Caliskan-Islam et al. 2016.

- i *data journalist* di Bloomberg[8] hanno mostrato come il modello automatico usato da Amazon per selezionare i quartieri delle città americane a cui offrire gratuitamente la modalità di consegna *same-day delivery* sia razzista. Il software, all'insaputa della compagnia, escludeva dall'offerta in modo sistematico le zone abitate da minoranze etniche in molte città, mentre includeva quelle limitrofe. Amazon ha replicato all'inchiesta giornalistica di non essere al corrente della pratica, perché il modello di ML era totalmente autonomo e basava le sue scelte sull'attività pregressa dei clienti. Insomma, è colpa dell'algoritmo.

Come abbiamo riportato nel lavoro che nel 2008 lanciò lo studio del *Discrimination-aware data mining*: «apprendere dalle tracce digitali delle decisioni del passato può portare a incorporare in modo invisibile nei modelli risultanti i pregiudizi esistenti, perpetuandoli».

IL 'DIRITTO ALLA SPIEGAZIONE'

Attraverso il DL stiamo creando sistemi che non sappiamo bene come funzionano. Anche il legislatore europeo si è reso conto della trappola, e la parte forse più innovativa e lungimirante della *General Data Protection Regulation* (Gdpr), la nuova regolamentazione sulla privacy entrata in vigore in Europa il 25 maggio scorso, è proprio il *diritto alla spiegazione*, ovvero di ottenere informazioni comprensibili sulla logica adottata da un qualunque sistema di decisione automatica che abbia effetti legali, o 'analogamente rilevanti', per le persone coinvolte. Senza una tecnologia in grado di spiegare la logica delle black box, però, il diritto alla spiegazione è destinato a rimanere lettera morta, oppure a porre fuorilegge molte applicazioni del ML opaco. Non si tratta solo di etica digitale, di evitare discriminazioni e ingiustizie, ma anche di sicurezza e responsabilità delle imprese. Automobili a guida autonoma, assistenti robotici, sistemi IoT domotici e manifatturieri, medicina personalizzata di precisione: in questi ambiti e altri ancora le imprese lanciano servizi e prodotti con componenti di AI che potrebbero incorporare inavvertitamente decisioni sbagliate, cruciali per la sicurezza, apprese da errori o da correlazioni spurie nei dati di apprendimento. Come, ad esempio, riconoscere un oggetto in una foto dalle proprietà non dell'oggetto stesso ma dello sfondo, a causa di un bias sistematico nella raccolta degli esempi per l'apprendimento. Come fanno le imprese a fidarsi dei loro prodotti senza comprenderne e validarne il funzionamento? Una tecnologia della spiegazione è essen-

8. <www.bloomberg.com/graphics/2016-amazon-same-day/> [21-2-2019].

ziale per creare prodotti con AI affidabili, per proteggere la sicurezza dei consumatori e circoscrivere la responsabilità industriale. Conformemente, l'uso scientifico del ML, come in medicina, biologia, economia o scienze sociali richiede comprensibilità non solo per poterci fidare dei risultati, ma per il carattere di per sé aperto della ricerca scientifica, perché possa essere condivisa e progredire. La sfida è dura e stimolante: una spiegazione non deve essere solo corretta ed esauriente, ma anche comprensibile a una pluralità di soggetti con esigenze e competenze diverse, dall'utente oggetto di una decisione, agli sviluppatori di soluzioni di AI, ai ricercatori, ai data scientist, ai policy makers, alle autorità di controllo, alle associazioni per i diritti civili, ai giornalisti.

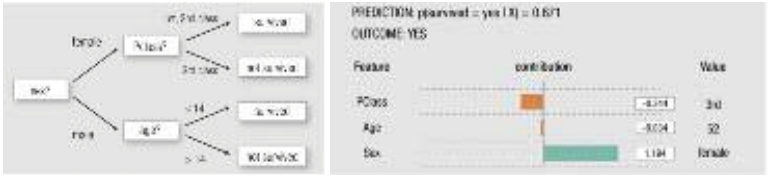
Che cosa sia una «spiegazione» se lo chiedeva già Aristotele in *Fisica*, trattato databile al IV secolo a.C. Oggi è urgente darne un significato operativo, di interfaccia fra le persone e gli algoritmi che suggeriscono decisioni, o che decidono direttamente, in modo che la AI serva a potenziare le capacità umane, non a soppiantarle. In generale, gli approcci di spiegazione differiscono per i diversi tipi di dato da cui si vuole apprendere un modello. Per dati tabellari, i metodi di spiegazione cercano di individuare quali siano le variabili che contribuiscono a una specifica decisione o predizione nella forma di insiemi di regole *if-then-else* o alberi di decisione. Si consideri, ad esempio, la tabella in basso relativa ai passeggeri del transatlantico britannico Titanic descritti dagli attributi: classe di viaggio, età, sesso, se sopravvissuti o meno al naufragio.

Passenger ID	Pclass	Age	Sex	Survived
1	3	22	M	N
2	1	38	F	Y
3	3	26	F	Y
4	1	35	F	Y

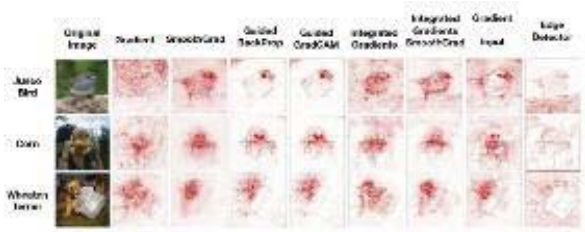
Tabella con passeggeri del Titanic.

Nella pagina successiva, in alto, a sinistra: Albero di decisione; a destra, Modello Lineare. Al centro, *Saliency Map*: spiegazione visuale della rilevanza della porzione di immagine per una specifica classificazione. Algoritmi diversi si concentrano su caratteristiche diverse dell'immagine da classificare.

In basso, spiegazione di un sistema di raccomandazione per film: membri del cast e frasi della trama sono enfatizzate in maniera differente a seconda della loro attrattività per quel particolare utente.



L'albero di decisione nella figura a sinistra informa che se era una donna che viaggiava in prima o seconda classe si sarebbe salvata, mentre la rappresentazione nella figura a destra riporta che se si fosse trattato di una cinquantaduenne in terza classe avrebbe avuto il 67% di probabilità di restare in vita grazie, soprattutto, all'essere donna, ma l'età e la classe di viaggio avrebbero giocato a suo sfavore. Nei metodi di classificazione di immagini si usano poi le *saliency maps*, cioè visualizzazioni che enfatizzano quella porzione che ha maggiormente contribuito alla classificazione. Nella figura sotto sono messi a confronto le saliency map generate da algoritmi che hanno classificato / riconosciuto le tre immagini: un fringuello, una rappa di granturco e un cane Terrier[9].



9. Adebayo et al. 2018.

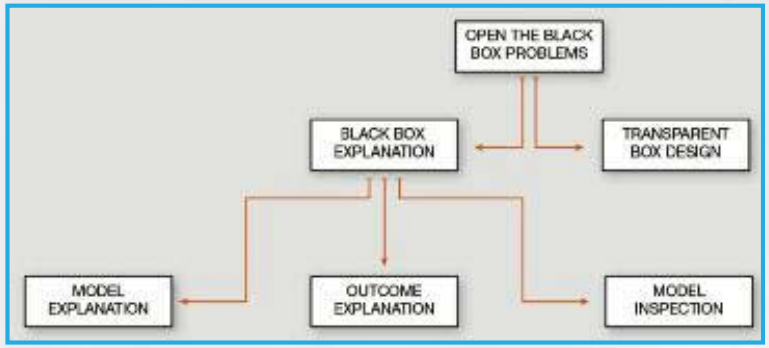
Un esempio di spiegazione su dati testuali sono i sistemi di raccomandazione che cercano di ricostruire una storia testuale che evidenzi le motivazioni del suggerimento, come nella figura sotto con due raccomandazioni per il medesimo film a due utenti. Aspetti diversi del testo sono enfatizzati per rispondere agli interessi dello specifico destinatario[10].

10. Hu et al. 2018.

Negli ultimi due anni si è registrato un impetuoso sforzo di ricerca sulla AI comprensibile, ma una tecnologia per la spiegazione che sia pratica e applicabile in modo sistematico non è ancora emersa. Nel lavoro pionieristico del 2008, *Discrimination-aware data mining*[11], si propongono metodi per fare auditing delle tracce dei sistemi di decisione e svelarne le logiche discriminatorie nascoste, aprendo la strada al filone di ricerca che, recentemente, ha guadagnato le luci della ribalta a causa dell'impetuoso successo della AI. Ma da qui alla soluzione del problema generale della «Explainable AI» c'è ancora molto lavoro da fare. Recentemente Fosca Giannotti ha ricevuto un ERC Advanced Grant proprio sul tema XAI: *science and technology for Science and technology for the eXplanation of* AI *decision making*. Un programma di ricerca di cinque anni che permetterà di affrontare le numerose sfide aperte. Semplificando la scena, ci sono due modi diversi di affrontare il problema:

– *spiegazione by-design*: XbD. Dato un insieme di dati di decisioni (ad esempio la tabella del Titanic), come costruire un «decisore automatico trasparente» che fornisca suggerimenti comprensibili;
– *spiegazione delle black-box*: Bbx. Dato un insieme di decisioni prodotte da un «decisore automatico opaco», come ricostruire una spiegazione.

Nel lavoro di rassegna[12] si discutono i lavori più recenti in letteratura organizzandoli secondo l'ontologia illustrata nella figura in basso. Oggi abbiamo risultati incoraggianti che permettono di ricostruire spiegazioni individuali, risposte a domande tipo «Perché non sono stato scelto per il posto a cui ho fatto domanda? Cosa dovrei cambiare per ribaltare la decisione?»[13].



Open Black Box Problem. La prima distinzione riguarda XbD e Bbx. Quest'ultimo può essere ulteriormente diviso tra *Model Explanation*, quando l'obiettivo della spiegazione è l'intera logica del modello oscuro, *Outcome Explanation*, quando l'obiettivo è la spiegazione delle decisioni su un particolare caso, e *Model Inspection*, quando l'obiettivo è capire il comportamento interno del modello oscuro su input diversi, come ad esempio le *saliency maps*.

11. Pedreschi et al. 2008.
12. Guidotti et al. 2018.
13. Pedreschi et al. 2018.

Stiamo evolvendo rapidamente da un tempo in cui erano le persone a codificare gli algoritmi, assumendosi la responsabilità della correttezza e della qualità del software prodotto e delle scelte in esso rappresentate, a un tempo in cui le macchine inferiscono autonomamente gli algoritmi sulla base di un numero sufficiente di esempi del comportamento di input/output atteso. In questo scenario dirompente, fare in modo che le black box della AI siano aperte e comprensibili non serve solo a verificarne correttezza e qualità, ma soprattutto ad allineare gli algoritmi con i valori e le aspettative umane e a preservare, anzi a espandere, l'autonomia e la consapevolezza delle nostre decisioni.

UOMO E VALORI AL CENTRO

Per questi motivi concordiamo con la posizione espressa in vari documenti della Commissione Europea e dalla migliore progettualità da parte degli scienziati europei, e crediamo che la nostra chance, appunto come europei, sia proprio lo sviluppo di forme più avanzate della AI attualmente *mainstream*, che pongano al centro di questo sviluppo la persona umana e i valori europei. Per essere davvero adottata su scala sociale e industriale, l'AI deve essere in grado di amplificare le potenzialità umane, soprattutto a livello cognitivo, non puntare a rimpiazzarle. I vincoli di tipo etico sono stimoli per creare innovazioni migliori, non ostacoli all'innovazione. La Gdpr, per dirla con Tim Cook, Ceo di Apple, non è un problema ma una grande occasione. I valori non sono orpelli del passato, sono princìpi di buon *design* per una tecnologia rivolta al benessere individuale e sociale. Non è una fisima 'buonista', ma una sfida scientifica e tecnologica entusiasmante, che richiede ricerca multidisciplinare fondamentale per lo sviluppo di nuovi modelli. Dall'apprendimento puramente statistico, per correlazioni, al ragionamento causale. Da modelli decisionali black box a modelli interpretabili, per disegnare sistemi di AI che conversano con gli umani per aiutarli a migliorare la qualità delle loro decisioni. Dall'ottimizzazione delle decisioni individuali all'ottimizzazione dell'effetto aggregato, per comprendere gli effetti di rete e armonizzare gli obiettivi individuali con quelli collettivi. Per misurare, prevedere e preservare la sostenibilità e la resilienza dei sistemi (tecno-)sociali complessi che abitiamo. Per disegnare nuove modalità di interazione fra persone e macchine in modo che le prime raggiungano livelli più alti di consapevolezza e le seconde livelli più alti di apprendimento e di comprensione del contesto e del ragionamento umano.
Questa è anche la visione di una iniziativa di ricerca europea che stiamo seguendo per conto del Cnr e dell'Università di Pisa, il progetto H2020 H*umane* AI, che ha il compito di definire l'agenda di ricerca per realizzare la scienza e la tecnologia della AI fondati su valori etici europei

SITOGRAFIA https://arxiv.org/abs/1608.07187> [20-5-2019]; <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=57112> [20-5-2019].

BRIEFING

Requested by the IMCO committee

Study in Focus

European Parliament

Artificial Intelligence (AI): New Developments and Innovations applied to E-Commerce

Challenges to the functioning of the Internal Market

The original full study[1] discusses the opportunities and challenges brought by the recent and the foreseeable developments of Artificial Intelligence into online platforms and marketplaces. The paper advocates the importance to support trustworthy, explainable AI (in order to fight discrimination and manipulation, and empower citizens), and societal-aware AI (in order to fight polarisation, monopolistic concentration and excessive inequality, and pursue diversity and openness).

Check out the original full study by scanning this QR code!

Background

Artificial Intelligence (AI) is becoming an essential component in all sectors of today's economy, with a significant impact on our private, social, and political lives. **AI** refers to an ecosystem of models and technologies for **perception**, **reasoning**, **interaction** and **learning**. The return of AI to the limelight is mainly related to *learning from data*, **Machine Learning (ML)**, that made a jump ahead with the emergence of **Big Data**. The mix is threefold: data reach the critical mass of examples to learn from, algorithms discover predictive patterns hidden in the data, the high-performance architectures succeed to provide the computing and storage resources needed. Based on this mix, AI managed to successfully tackle **long-standing open challenges**, such as understanding of texts and speech, recognition of the content of images and video, and other tasks believed to require intelligence. Approximately ten years ago, it was noticed that some long-known learning models, hitherto ineffective to the mentioned tasks, if generalised and trained on large sets of example data, can make a qualitative leap and learn, from the pixels of the example images or from the words of the example texts, the general "concepts" to recognise and classify accurately new images and new texts.

The current stage of development of AI exhibits strengths and weaknesses. On the one hand, the learning capacity of AI models is growing, bringing extraordinary progress in robotic vision and autonomous driving, in automated text and speech translation, in medical diagnosis, in risk assessment, in predictive maintenance. On the other hand, the gap with other aspects of AI is growing, in particular, with **reasoning** and **person-machine interaction**, central aspects for the development of a human, ethical, and anthropocentric AI, that is the focus of the European approach. The opacity and nature of black boxes of AI models are growing, together with the risk of creating systems exposed to biases in training data, systems that even experts fail to understand. **Tools are missing** to allow AI developers to certify the **reliability** of their models.

Policy Department for Economic, Scientific and Quality of Life Policies
Directorate-General for Internal Policies
Authors: Prof. Dr Dino PEDRESCHI and Dr Ioanna MILIOU
PE 652.720 - June 2020

EN

It is crucial to inject into AI technologies, ethical values of **fairness** (how to avoid unfair and discriminatory decisions), **accuracy** (how to provide reliable information), **confidentiality** (how to protect the privacy of the involved people) and **transparency** (how to make models and decisions comprehensible to all stakeholders). This **value-sensitive design** approach, yet to be fully operationalised, is strongly needed for boosting widespread social acceptance of AI, without inhibiting its power.

Furthermore, as increasingly complex sociotechnical systems emerge, consisting of many interacting people and intelligent and autonomous systems, AI acquires an important **societal dimension**. A key observation is that **a crowd of intelligent individuals (assisted by AI tools) is not necessarily an intelligent crowd**. On the contrary, it can be stupid in many cases, because of undesired, unintended **network effects** and **emergent aggregated behaviour**. Examples abound in contemporary society. For example, using a car navigation to avoid a traffic jam on the main route can cause additional jams in the local alternative routes. In the field of opinion formation and diffusion, a crowd of citizens using **social media** as a source of information is subject to the **algorithmic bias of the platform's recommendation mechanisms** suggesting personalised content. This bias will create echo chambers and filter bubbles, sometimes induced in an artificial way, in the sense that without the personalisation bias the crowd would reach a common shared opinion. Recommendations provided by AI systems may make sense at an individual level but they may result in undesired collective effects of information disorder and radicalisation.

The flow of information reaching us via the online media platforms and ecommerce marketplaces is optimised not by the information content or product quality but by **popularity** and **proximity to the target**. This is typically performed in order to maximise platform usage. The **recommendation algorithms** suggest the interlocutors, the products and the contents at which we get exposed, based on matching profiles, promoting the most popular choices for people similar to us. As a result, we observe the appearance of the "**rich get richer**" phenomenon: popular users, contents and products get more and more popular. In the online marketplaces this would mean that a few businesses get the biggest share of the market while many have to share the rest. These businesses become the hubs of the network, gathering most of users' purchases to the detriment of the vast majority. In the social media, that would mean that certain popular peers or contents gather all the user's attention, becoming the hubs of the social network. As a consequence of **network effects** of AI recommendation mechanisms for online marketplaces, search engines and social networks, the **emergence of extreme inequality** and **monopolistic hubs** is artificially amplified, while **diversity of offers** and easiness of **access to markets** are artificially impoverished.

Key findings

There is a wide consensus that AI will bring forth changes that will be much more profound than any other technological revolution in human history. Depending on the course that this revolution takes, AI will either empower our ability to make more informed choices or reduce human autonomy; expand the human experience or replace it; create new forms of human activity or make existing jobs redundant; help distribute well-being for many or increase the concentration of power and wealth in the hands of a few; expand democracy in our societies or put it in danger. Our generation carries the responsibility of shaping the AI revolution. The choices we face today are related to fundamental ethical issues about the impact of AI on society, in particular, how it affects labour, social interactions, healthcare, privacy, fairness, security and markets.

The current technological advancements and developments of AI that can occur in the near future, if driven along the path of a **human-centric AI**, could represent an important transformation factor for ecommerce/digital services and for the Internal Market. Novel AI platforms for ecommerce and digital services based on human-centric AI interaction mechanisms have the potential to **mitigate monopolistic concentration**, deliver more **open and resilient markets**, and better connect the diverse demands of European consumers to the diverse offer of European products and services, by fostering **diversity "by-design"**. A new generation of AI-based recommendation and interaction mechanisms may help departing from the current purely "advertisement-centric" model, focusing on the interests of platforms in maximising intermediation revenues, to a systemic approach where focus is on the interest of citizens in accessing and sharing of high quality contents, the interest of consumers to broaden their choices and opportunities, and the interest of advertisers in broadening their audience and customer base.

Within this landscape, the European strategy for the next-generation digital services and online platforms is of utmost importance, with impacts that will go far beyond consumer protection, towards shaping the digital society that will emerge.

Coherently with the recent strategic white paper of the European Commission "*On Artificial Intelligence – A European approach to excellence and trust*", we recommend to develop the European provisions on Artificial Intelligence in the area of ecommerce and digital services - in the context of reforming the E-commerce Directive and introducing the Digital Services Act - with the aim to address and exploit the transformative impact of upcoming human-centric AI developments, to the purpose of social progress. Accordingly, our proposed **recommendations to EU policy makers** follow a double line of reasoning and interventions: **topic-wise**, and **instrument-wise**.

Topic-wise, it is crucial to address and operationalise the following challenges,

- **trustworthy, explainable AI** in order to **fight novel forms of discrimination and manipulation** and **empower citizens**;
- **societal-aware AI** in order to **fight polarisation and inequality** and **pursue diversity and openness**.

Instrument-wise, it is important to realise that the scientific and technological landscape is not yet fully mature to address all the open challenges discussed here. Therefore a mix of policies is needed, that tackle the problem at three levels,

- a bold **EU investment in fundamental and applied research in human-centric AI**;
- a bold **EU investment in creating novel online platforms and digital services** embodying human-centric AI mechanisms (and/or supporting the scaling of existing coherent initiatives);
- a coherent set of **EU regulations concerning AI, big data and digital services**, designed not only to seize the opportunities and mitigate risks, but also to inspire research and development in AI, big data and digital platforms towards an inclusive, equal and diverse society.

[1] https://www.europarl.europa.eu/RegData/etudes/IDAN/2020/648791/IPOL_IDA(2020)648791_EN.pdf

IP/A/IMCO/2020-21; Manuscript completed: May 2020; Date of publication: June, 2020
Administrators responsible: Mariusz MACIEJEWSKI, Christina RATCLIFF; Editorial assistant: Irene VERNACOTOLA
Contact: Poldep-Economy-Science@ep.europa.eu
This document is available on the internet at: www.europarl.europa.eu/supporting-analyses

Print ISBN 978-92-846-6814-4 | doi: 10.2861/252506| QA-02-20-427-EN-C
PDF ISBN 978-92-846-6815-1 | doi: 10.2861/14351 | QA-02-20-427-EN-N