ISSN 1826-3534

federalismi.it
RIVISTA DI DIRITTO PUBBLICO ITALIANO, COMPARATO, EUROPEO

LA DISINFORMAZIONE ONLINE
24 APRILE 2020

Profilazione e decisione algoritmica: dal mercato alla sfera pubblica

di Francesca Lagioia
Senior Research Fellow
Istituto Universitario Europeo

e Giovanni Sartor
Professore Ordinario di Filosofia del diritto
Alma Mater Studiorum – Università di Bologna

Profilazione e decisione algoritmica: dal mercato alla sfera pubblica *

di Francesca Lagioia
Senior Research Fellow
Istituto Universitario Europeo

e Giovanni Sartor
Professore Ordinario di Filosofia del diritto
Alma Mater Studiorum – Università di Bologna

Abstract [It]: Il contributo esamina il ruolo e l'impatto della profilazione degli individui, e in particolare degli utenti delle piattaforme digitali, sui meccanismi sociali, economici e politici che regolano il funzionamento della Rete, del mercato e più in generale della sfera pubblica, con attenzione al fenomeno delle c.d. fake news. La profilazione e i meccanismi di sorveglianza sempre più persistenti e pervasivi, messi in campo grazie alle più recenti tecniche dell'intelligenza artificiale, sono analizzati con particolare rifermento alle capacità di classificazione, valutazione, influenza e manipolazione degli individui. Il contributo propone quindi un'analisi delle nuove frontiere dell'ecosistema della sorveglianza nella sfera pubblica, con particolare riferimento all'opinione pubblica e alle preferenze di voto, fino a ricomprendere fenomeni quali ad esempio il c.d. Sistema di credito Sociale Cinese. Infine si esamina il nesso tra la regolazione dell'intelligenza artificiale e le principali sfide tecnologiche dei prossimi anni.

Abstract [En]: The paper aims at analysing the role and the impact of user profiling (particularly on the Internet platforms) on the social, economical and political mechanisms of functioning of the Internet, the market, and, more broadly, the public sphere. In this paper, particular attention to the issue of the so-called fake news is paid.
In this sense, the increasingly persistent and pervasive user profiling and surveillance mechanisms developed due to the more recent artificial intelligence techniques are analysed by looking at the issues of the scope of classification, evaluation, influence, and manipulation of individuals. As a consequence, the paper explores the new frontiers of the surveillance ecosystem in the public sphere, with particular reference to public opinion and voting preferences, including phenomena such as the so-called Chinese Social Credit System.
Finally, the link between the regulation of artificial intelligence and the main technological challenges of the upcoming years is analysed.

1. Introduzione

La convergenza tra Internet e l'intelligenza artificiale (IA) genera uno scenario affascinante e problematico a un tempo: in un ecosistema circolare di reciproco potenziamento, la rete alimenta l'intelligenza artificiale rendendo disponibili enormi quantità di dati (c.d. Big Data), mentre l'intelligenza artificiale permette alla rete un crescente sfruttamento di quegli stessi dati.

Grazie a tale convergenza, l'IA sta già trasformando gli assetti economici, politici e sociali, le interazioni tra individui, e la stessa vita privata. Se da un lato, il connubio tra Internet e IA offre nuove promesse di benessere e progresso - con grandi opportunità per lo sviluppo economico, sociale e culturale, la

* Articolo sottoposto a referaggio.

sostenibilità energetica, la salute e la diffusione della conoscenza - dall'altro, esso aumenta le opportunità di controllo e manipolazione degli individui, prospettando nuovi pericoli per i valori sociali e i diritti fondamentali.

Si è osservato come la sostituzione di sistemi intelligenti al lavoro umano possa svalutare il lavoro di chi può essere sostituito dalle macchine: molti rischiano di perdere la «corsa contro le macchine»[1], e quindi di essere esclusi dalle attività produttive, il che può comportare povertà ed emarginazione sociale. Si pensi all'impatto delle auto autonome su tassisti e camionisti, o all'impatto dei *chatbot* (robot software per la conversazione) intelligenti sui lavoratori nei call center. Ciò comporta, da un lato, l'esigenza di garantire una vita dignitosa anche a chi abbia perso il proprio lavoro in seguito all'automazione, ma anche l'esigenza di offrire nuove opportunità di attività e iniziativa.

Inoltre, l'intelligenza artificiale, consentendo alle grandi imprese digitali di ottenere enormi profitti con forza lavoro limitata, concentra la ricchezza in chi investe in tali imprese e in chi sa meglio progettare e utilizzare le tecnologie da esse realizzate. Ciò favorisce modelli economici in cui «il vincitore prende tutto» (*winner takes all*) sia nei rapporti tra imprese (dove prevalgono posizioni monopolistiche determinate dall'accesso privilegiato o esclusivo a dati e tecnologie) sia nei rapporti tra lavoratori (dove prevale chi è in grado di svolgere funzioni di alto livello, non automatizzabili). Non si tratta di conseguenze inevitabili. La formazione dei lavoratori, modelli d'interazione uomo macchina che enfatizzano la creatività umana, misure di redistribuzione e di accesso ai dati e alle tecnologie, possono consentire a tutti di beneficiare dei frutti dell'intelligenza artificiale, ma ciò richiede adeguate scelte politiche e sociali.

I sistemi di intelligenza artificiale e i dati da essi utilizzati possono offrire nuove occasioni per attività chiaramente illegali: essi possono essere oggetto di attacchi o possono essere strumenti per commettere atti criminali (per esempio, veicoli autonomi possono essere utilizzati per assassini o atti terroristici, algoritmi intelligenti per frodi o reati finanziari)[2].

Anche al di fuori delle attività chiaramente illegali, attori motivati dal profitto possano usare l'intelligenza artificiale per perseguire legittimi interessi economici in modi dannosi per gli individui e la società. Le imprese commerciali e i governi, combinando intelligenza artificiale e dati, possono sottoporre cittadini, utenti, consumatori e lavoratori a una sorveglianza pervasiva, limitare le informazioni e le opportunità cui gli stessi hanno accesso, e manipolarne le scelte in direzioni contrastanti con i loro interessi. Gli abusi sono incentivati dal fatto che molte imprese di Internet - come ad esempio le maggiori piattaforme che ospitano contenuti generati dagli utenti - operano in mercati a due o più lati: i loro servizi principali (per

[1] E. BRYNJOLFSSON - A. MCAFEE, *Race Against the Machine*, Digital Frontier Press, Archivio online, 2011.
[2] N. BHUTA - S. BECK - R. GEISS - C. KRESS - H.Y. LIU, *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge, 2015.

esempio, ricerca, gestione di reti sociali e accesso a contenuti) vengono offerti a singoli utenti, ma i ricavi provengono dagli inserzionisti, o da chi sia comunque interessato a influenzare gli utenti, come nel caso della pubblicità e della propaganda politica personalizzata. Pertanto, le piattaforme non si limitano a raccogliere le informazioni sugli utenti utili a meglio indirizzare pubblicità personalizzate, ma usano ogni mezzo disponibile per trattenere gli utenti, in modo che essi siano esposti a messaggi pubblicitari o ad altri tentativi di persuasione. Ciò conduce non solo ad una massiva raccolta di dati sugli individui, a danno della privacy, ma anche ad un'influenza pervasiva sul comportamento degli stessi individui, a danno non solo dell'autonomia dei singoli ma anche di interessi collettivi. Per esempio, la manipolazione del comportamento dei consumatori può condurre all'esclusione di individui e gruppi da scambi e opportunità, e di conseguenza a un cattivo funzionamento dei mercati. Inoltre, algoritmi guidati dalla ricerca del profitto possono convergere su strategie anticoncorrenziali, a danno dei concorrenti ma anche dei consumatori[3]. Ulteriori problemi derivano dalla possibilità che l'IA amplifichi i pregiudizi sociali come conseguenza di determinazioni algoritmiche inique o discriminatorie.

Problematiche simili possono emergere nel settore pubblico e nella dialettica politica. I governi possono servirsi dell'intelligenza artificiale non solo per scopi politici e amministrativi legittimi (es. efficienza, risparmi sui costi, servizi migliorati), ma anche per anticipare e controllare il comportamento dei cittadini in modi che limitano le libertà individuali e interferiscono con il processo democratico. La manipolazione dell'opinione politica, basata sulla profilazione e l'invio di messaggi mirati può condurre alla polarizzazione e all'estremizzazione delle opinioni, e quindi al deterioramento del dialogo pubblico.

La formazione dell'opinione pubblica può essere compressa anche in assenza di un'intenzione antidemocratica, grazie ai modelli economici della rete e alle dinamiche socio-informative che ne emergono. Infatti, la ricerca dei profitti pubblicitari induce le piattaforme a catturare l'attenzione degli utenti, e trattenerli proponendo loro contenuti che ne assecondino i gusti e concordino con le loro opinioni, profittando della propensione alla conferma (*confirmation bias*) che caratterizza la psicologia umana[4]. Tale pratica può condurre alla polarizzazione e frammentazione della sfera pubblica[5], oltre che alla proliferazione di notizie sensazionalistiche, non verificate o false (*fake news*). L'intelligenza artificiale e i big data contribuiscono a questo fenomeno, mediante la generazione di messaggi persuasivi, e mediante la direzione di messaggi personalizzati verso chi possa essere maggiormente influenzato.

[3] Si vedano A. EZRACHI - M.E. STUCKE, *Virtual competition*, in *Journal of European Competition Law & Practice*, nn. 7-9/2016, pp. 585–586; e S.F. JANKA - S. UHSLER, *Antitrust 4.0-the rise of artificial intelligence and emerging challenges to antitrust law*, in *European Competition Law Review,* 39, n. 3/2018, pp. 112–123.
[4] E. PARISER, *The filter bubble: What the Internet is hiding from you,* New York, 2011.
[5] C.R. SUNSTEIN, *Algorithms, correcting biases*, in *Social Research*, 86, n. 2/2019, pp. 499–511.

2. L'intelligenza artificiale e i dati personali

I sistemi di intelligenza artificiale hanno fatto un vero salto di qualità - dando luogo a numerose applicazioni di successo, in diversi settori, dalla traduzione automatica, all'ottimizzazione dei processi industriali, al marketing, ecc. - quando l'attenzione si è spostata dalla rappresentazione logica della conoscenza alla possibilità di applicare metodi di apprendimento automatico (*machine learning*) a grandi masse di dati. Nel modello dell'apprendimento automatico l'uomo fornisce alla macchina un metodo di apprendimento, da applicare ai dati cui essa ha accesso, per estrarre automaticamente da quei dati le indicazioni su come svolgere il compito affidatole.

Nel campo dell'apprendimento automatico si è soliti distinguere tre approcci principali: l'apprendimento supervisionato, l'apprendimento per rinforzo e l'apprendimento non supervisionato.

L'apprendimento supervisionato è oggi il modello più frequentemente impiegato. Al sistema viene fornito un vasto insieme di esempi di comportamento corretto, in modo che il sistema impari ad agire in modo analogo. Per esempio, un sistema che raccomanda prodotti potrà suggerire certi acquisti a determinati individui, sulla base degli acquisti effettuati da altri clienti in passato che condividono le caratteristiche di quegli individui. Il sistema impara a classificare in modo corretto nuovi casi sulla base di un insieme di precedenti (ciascuno etichettato come corretto o sbagliato), senza che esseri umani abbiano rappresentato la conoscenza rilevante per la classificazione nella forma di regole o concetti generali.

All'addestramento supervisionato, si affianca l'addestramento per rinforzo, nel quale il sistema apprende dai risultati delle azioni proprie o altrui: è in grado cioè di distinguere successi e fallimenti (a seconda di come le azioni incidano sul raggiungimento della utilità o valori perseguiti dal sistema, per esempio, i guadagni nella finanza). Si immagini un sistema che migliori le proprie scelte di investimento sulla base dei risultati finanziari cui conducono tali scelte.

Nell'addestramento non supervisionato, infine, il sistema apprende senza indicazioni dall'esterno. Tecniche per l'addestramento non supervisionato sono utilizzate soprattutto per il *clustering*, cioè per raggruppare all'interno di un insieme (per esempio, di documenti, immagini, consumatori, elettori) soggetti o oggetti che presentano similarità.

In numerose applicazioni per l'apprendimento automatico, l'insieme di addestramento consiste di dati che vertono su caratteristiche e comportamenti individuali e sociali, e quindi, di dati personali. Per esempio, l'uso dell'intelligenza artificiale per l'invio di pubblicità personalizzate presuppone la raccolta di informazioni sugli individui e le loro scelte, così da poter connettere caratteristiche individuali (genere, età, estrazione sociale, acquisti precedenti, navigazione in rete, ecc.) alla propensione a rispondere a determinati messaggi pubblicitari con nuovi acquisti.

L'utilità di apprendere propensioni e attitudini degli individui fa sì che l'intelligenza artificiale sia sempre più desiderosa di dati personali, e questo desiderio stimola la raccolta continua di nuovi dati, che a loro volta rendono più efficaci le applicazioni di intelligenza artificiale, in una spirale di feedback che si rafforza[6].

Di conseguenza, lo sviluppo di sistemi di intelligenza artificiale basati sulla raccolta di dati presuppone e stimola al tempo stesso la creazione di ampi insiemi di dati, i cosiddetti *Big Data*[7]. La creazione dei Big Data è facilitata dal fatto che l'uso di ogni sistema informatico dà luogo alla raccolta di dati digitali concernenti le interazioni con lo stesso sistema[8]. La massiva digitalizzazione dei dati ha infatti preceduto lo sviluppo della maggior parte delle applicazioni d'intelligenza artificiale. Per esempio, innumerevoli dati vengono raccolti ogni secondo dai computer che mediano transazioni economiche (in particolare nel commercio elettronico), dai sensori che controllano oggetti fisici (per esempio, veicoli e dispositivi domestici), dalla gestione dei flussi di attività amministrative (per esempio, banche, trasporti, e gestione delle imposte), da dispositivi di sorveglianza (per esempio, telecamere su strada) e infine dai sistemi usati per attività non commerciali (per esempio, accesso a Internet, ricerca di dati, reti sociali).

Questi flussi di dati sono oggi integrati in un'infrastruttura globale universale per la comunicazione, l'accesso alle informazioni, la fornitura di servizi pubblici e privati. Tale struttura, che si incentra su Internet ma non si limita ad essa, opera mediante algoritmi che indirizzano e trasmettono i dati, e mediano l'accesso a contenuti e servizi, selezionando per noi informazioni e opportunità. Tale infrastruttura connette oggi più di 30 miliardi di dispositivi fra loro interconnessi - computer, telefoni, mezzi di trasporto, macchine industriali, telecamere, ecc. -, che generano un'enorme quantità di dati elettronici, decine di volte superiore a tutti i dati registrati in forma analogica nella storia dell'umanità. I flussi di dati tra macchine sono già oggi molto superiori alle comunicazioni umane.

3. La psicometria e le macchine: dalla profilazione all'influenza e alla manipolazione

L'intelligenza artificiale e i Big Data, in combinazione con la disponibilità di ampie risorse informatiche, hanno molto aumentato le opportunità di profilazione degli individui, vale a dire la possibilità di compiere inferenze – classificazioni, previsioni o decisioni – sulla base di dati riguardanti gli stessi. Ciò è in parte dovuto al fatto che l'automazione riduce i costi per la raccolta, l'archiviazione e l'elaborazione di informazioni, aprendo la strada a meccanismi di sorveglianza molto più persistenti e pervasivi.

[6] N. CRISTIANINI, *The road to artificial intelligence: A case of data over theory*, in *New Scientist*, 26 ottobre 2016, consultabile al sito < https://www.newscientist.com/article/mg23230971-200-the-irresistible-rise-of-artificial-intelligence/>.
[7] V. MAYER-SCHÖNBERGER - K. CUKIER, *Big Data*, Harcourt, 2013.
[8] H.R. VARIAN, *Computer Mediated Transactions*, in *100 American Economic Review*, 2, n. 1/2010.

Grazie all'intelligenza artificiale, tutti i tipi di dati personali possono essere utilizzati per analizzare, prevedere e influenzare il comportamento umano, un'opportunità che trasforma tali dati personali in merci dotate di valore. Informazioni che un tempo non erano raccolte o erano scartate (i cosiddetti "dati di scarto" (exhaust data)), sono oggi diventate una risorsa preziosa. Tutte le tracce del comportamento online possono essere utilizzate per addestrare un sistema di IA ad inferire informazioni sulle abitudini, gli interessi, la personalità, le condizioni di salute (come depressione e dipendenza), e i tratti psicologici degli individui[9], inclusi gli stati emotivi, i valori, le opinioni politiche e morali, gli orientamenti sessuali[10].
Grazie alla panoplia di sensori che tracciano in modo crescente ogni attività umana, la raccolta e l'analisi dei dati su dispositivi, sistemi, applicazioni e piattaforme online hanno assunto connotati di assoluta pervasività e ubiquità. I dati raccolti vengono elaborati attraverso le tecnologie dei Big Data e dell'IA, cosicché gli individui siano soggetti a sorveglianza, influenza e manipolazione in molti più casi e in molti più contesti, sulla base di un insieme più ampio di caratteristiche personali (che spaziano dalle condizioni economiche e finanziarie allo stato di salute, al luogo di residenza, fino ad includere le scelte di vita personali di ciascuno, il comportamento online e offline, ecc.). Questa dinamica contribuisce sempre più a plasmare l'economia globale, il flusso di idee e l'accesso alle informazioni.
I dati personali possono essere usati per la profilazione degli individui, cioè per classificarli, valutarli e prevederne i comportamenti. Il termine "profilazione" deriva da "profilare", che in origine significava tracciare una linea, e più in particolare i contorni di un oggetto. Questa è precisamente l'idea alla base della profilazione mediante l'elaborazione di dati: espandere le informazioni e i dati disponibili di individui e gruppi, in modo da disegnarne – descriverne o anticiparne – i tratti e le propensioni.
Un sistema di profilazione stabilisce (prevede) che gli individui con determinate caratteristiche C1, hanno anche una certa probabilità di possedere alcune caratteristiche aggiuntive C2. Si consideri, per esempio, il caso di un sistema che stabilisca (predica) che coloro che presentano certe caratteristiche genetiche hanno anche la tendenza a sviluppare il cancro con una probabilità superiore alla media, o che gli individui con una certa istruzione e un certo storico lavorativo o che appartengono a una certa etnia, hanno anche una probabilità superiore alla media di inadempienza dei propri debiti. In questi casi, è possibile affermare che il sistema in esame ha profilato il gruppo di individui che possiedono le caratteristiche C1, aggiungendo un nuovo segmento di informazioni alla descrizione (il profilo) di tale gruppo, vale a dire la probabilità di possedere le caratteristiche aggiuntive C2. Successivamente, indicando al sistema che un soggetto possiede le caratteristiche C1, il sistema sarà in grado di inferire che, con una certa probabilità,

[9] M. KOSINSKI - D. STILLWELL - T. GRAEPEL, *Private traits and attributes are predictable from digital records of human behavior*, in *Proceedings of the National Academy of sciences*, 110, n. 15/2013, pp. 5802-5805.
[10] C. BURR - N. CRISTIANINI, *Can machines read our minds?*, in *Minds and Machines*, 29, n. 3/2019, pp. 461-494.

quel determinato soggetto possiederà anche le caratteristiche aggiuntive C2. Ciò può comportare effetti sia positivi che negativi e far sì che l'individuo sia trattato in modo favorevole rispetto ai propri interessi o viceversa in modo dannoso. Per esempio, nel caso in cui la caratteristica inferita sia la maggiore probabilità di contrarre il cancro, la previsione del sistema potrà fornire la base per effettuare esami e terapie preventive, o viceversa per un aumento del premio assicurativo.

Una correlazione appresa può anche riguardare la propensione di un individuo a rispondere in determinati modi a certi stimoli. Ciò consente il passaggio dalla previsione all'influenzamento del comportamento, che può consistere in un legittimo indirizzamento verso scelte vantaggiose per l'individuo stesso, o invece in forme di manipolazione illegale o immorale. Per esempio, le informazioni inferite dal sistema possono riguardare la propensione a rispondere positivamente a un certo trattamento terapeutico, con un conseguente miglioramento delle condizioni cliniche, o la propensione all'acquisto di un prodotto rispetto a un determinato annuncio pubblicitario o ad una certa variazione di prezzo, o la propensione a rispondere a un certo tipo di messaggio con un mutamento di preferenze o stati d'animo (per esempio, relativamente alle scelte politiche e alle preferenze di voto). In tali circostanze, la profilazione comporta quindi la possibilità di influenzare e manipolare gli individui, innescando il comportamento desiderato.

Supponiamo, per esempio, che il sistema colleghi certe caratteristiche (ad esempio, una certa età, sesso, stato sociale, tipo di personalità, ecc.) alla propensione degli individui a reagire a un determinato messaggio (ad esempio, un annuncio pubblicitario mirato) con un certo comportamento (per esempio l'acquisto di un prodotto). Supponiamo anche che al sistema venga data l'informazione che un certo individuo possiede tali caratteristiche (è un giovane maschio, appartiene alla classe operaia, ha un carattere estroverso, ecc.). Il sistema sarà in grado di inferire che, con una certa probabilità, inviando quel determinato messaggio a quello specifico individuo, egli risponderà adottando il comportamento previsto.

La nozione di profilazione appena presentata trova riscontro nella seguente più elaborata definizione, che introduce il concetto generale usato in ambito informatico e sociologico:

«La profilazione è una tecnica di trattamento (parzialmente) automatizzato di dati personali e/o non personali, finalizzata alla creazione di conoscenza predittiva mediante la scoperta di correlazioni tra i dati e la costruzione di profili, che possono essere poi utilizzati per assumere decisioni. Un profilo è un insieme di dati correlati che rappresentano un soggetto (individuale o collettivo). La costruzione di profili è il processo di scoperta di schemi ricorrenti e sconosciuti tra i dati, all'interno di grandi insiemi di dati, che possono essere utilizzati per creare profili. L'applicazione di profili consiste nell'identificazione e

rappresentazione di uno specifico individuo o gruppo come corrispondente a un determinato profilo, e nel processo decisionale basato su tale identificazione e rappresentazione»[11].

La più specifica nozione delineata dall'articolo 4 del Regolamento per la Protezione dei dati (GDPR) collega la profilazione alle valutazioni delle decisioni relative agli individui, sulla base di dati personali, escludendo da tale nozione la costruzione di profili di gruppi di individui:

«la "profilazione" (…) consiste in qualsiasi forma di trattamento automatizzato di dati personali che valuti gli aspetti personali relativi a una persona fisica, in particolare per analizzare o prevedere aspetti quali le prestazioni lavorative dell'interessato, la situazione economica, la salute, le preferenze o gli interessi personali, l'affidabilità o il comportamento, la posizione geografica o gli spostamenti, laddove ciò produca effetti giuridici che lo o la riguardano o che influiscono in modo significativo su di lui o su di lei».

Anche qualora un sistema automatizzato di valutazione e decisione - basato sulla profilazione - sia imparziale e in linea di principio volto a servire scopi benefici, esso può influire negativamente sugli interessati. Coloro che sono soggetti a sorveglianza pervasiva, e a forme persistenti di valutazione e influenza sono infatti sottoposti a forti pressioni psicologiche, capaci di incidere sulla loro autonomia personale, rendendoli suscettibili di essere fuorviati, manipolati e sfruttati.

4. Le decisioni algoritmiche tra equità e discriminazione

Oltre ai rischi legati alle violazioni della privacy e della protezione dei dati, la profilazione crea nuovi rischi di stereotipizzazione, diseguaglianza e discriminazione (sulla base di età, genere, razza, religione, orientamento sessuale, ecc.), a causa delle classificazioni e delle categorizzazioni su cui essa si basa. Essa può condurre a scelte che compromettono l'interesse dei singoli a un trattamento algoritmico equo e corretto, vale a dire, l'interesse a non essere soggetti a pregiudizi ingiustificati in seguito a elaborazioni automatiche.

La combinazione di Big Data e IA consente infatti di automatizzare i processi di decisione, anche in ambiti che richiedono scelte complesse, basate su numerosi fattori, in base a criteri non esattamente

[11] F. BOSCO - N. CREEMERS - V. FERRARIS - D. GUAGNIN - B.J. KOOPS, *Profiling technologies and fundamental rights and values: regulatory challenges and perspectives from European Data Protection Authorities*, in *Reforming European data protection law*, Dordrecht, 2015, p. 8 (traduzione degli autori); si veda anche M. HILDEBRANDT, *Profiling and AML*, in K. RANNENBERG – D. ROYER – A. DEUKER (a cura di), *The Future of Identity in the Information Society. Challenges and Opportunities*, Dordrecht, 2009.

predeterminati. Ciò può migliorare la qualità delle decisioni pubbliche e private, ma comporta nuovi rischi.

Si è aperto infatti negli ultimi anni un ampio dibattito su prospettive e rischi delle decisioni algoritmiche[12]. Alcuni studiosi hanno rilevato come in molti settori le previsioni e le decisioni algoritmiche, anche relative alla valutazione degli individui, siano più precise ed efficaci di quelle umane. I sistemi automatici possono evitare le propensioni all'errore proprie dell'uomo, in particolare nel valutare dati statistici[13], così come i pregiudizi - etnici, sociali, di genere, ecc. - da cui spesso siamo affetti. Si è osservato che in molti ambiti - dagli investimenti, al reclutamento di personale, alla concessione di libertà vigilata - le determinazioni algoritmiche risultano migliori, con riferimento ai criteri usuali, di quelle adottate anche da individui esperti[14].

Altri hanno posto l'accento sulle possibilità di errore e discriminazione delle decisioni algoritmiche. Solo in rari casi gli algoritmi assumeranno decisioni esplicitamente discriminatorie, la cosiddetta discriminazione diretta (*disparate treatment*), basando le proprie previsioni su caratteristiche vietate, come razza, etnia o genere. Più spesso il risultato di una determinazione algoritmica comporterà una discriminazione indiretta (*disparate impact*), cioè avrà un impatto sfavorevole sproporzionato su individui appartenenti a certi gruppi, senza una giustificazione accettabile.

I sistemi basati su metodi di apprendimento supervisionato imparano dagli esempi contenuti nel loro insieme di addestramento, e quindi tendono a riprodurne pregi e difetti, inclusa la propensione all'errore e al pregiudizio. Per esempio, un sistema per il reclutamento, addestrato su esempi di decisioni passate in cui donne e minoranze siano state discriminate, riprodurrà la stessa logica[15]. Il pregiudizio all'interno di un insieme di addestramento può essere presente anche se le informazioni che costituiscono i dati di ingresso del sistema non comprendono caratteristiche discriminatorie il cui uso sia giuridicamente vietato, come l'etnia o il genere. Ciò può verificarsi ogni qual volta esista una correlazione tra caratteristiche discriminatorie e alcuni dati di imput utilizzati dal sistema. Supponiamo, ad esempio, che un responsabile delle risorse umane non abbia mai assunto candidati di una certa etnia a causa di un suo pregiudizio e che gli individui appartenenti a tale etnia abitino per lo più in certi quartieri della città. Un insieme di addestramento basato sulle decisioni di tale dirigente insegnerebbe al sistema a non selezionare gli

[12] Per una discussione degli aspetti costituzionali delle decisioni algoritmiche, vedi A. SIMONCINI, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, in *Rivista di biodiritto*, n. 15/2019, pp. 63-89.
[13] D. KAHNEMAN, *Thinking: fast and slow*, New York, 2011.
[14] C.R. SUNSTEIN, *Algorithms, in Correcting Biases*, cit.
[15] J. KLEINBERG - J. LUDWIG - S. MULLAINATHAN - C.R. SUNSTEIN, *Discrimination in the age of algorithm*, in *Journal of Legal Analysis*, n. 10/2018, pp. 113–174.

individui residenti in quei quartieri, comportando il mancato accoglimento di tutte le domande di assunzione provenienti da coloro che appartengano all'etnia discriminata.

In altri casi, un insieme di addestramento può essere parziale e contenere pregiudizi rispetto a un certo gruppo, poiché il raggiungimento del risultato desiderato (ad esempio, certe prestazioni lavorative) è approssimato mediante un *proxy* (un elemento sostitutivo) che comporta un impatto discriminatorio indiretto su quel gruppo. Supponiamo, ad esempio, che le future prestazioni lavorative dei dipendenti (obiettivo di interesse nei processi di assunzione) siano valutate considerando unicamente una predizione concernente il numero di ore spese in ufficio. Tale criterio potrebbe comportare una valutazione peggiore delle donne rispetto agli uomini, qualora in passato le donne abbiamo speso meno ore in ufficio, a causa di impegni familiari più pesanti. Il sistema assocerebbe correttamente al genere femminile una minore propensione al lavoro straordinario, ma risulterebbe viziato dal collegamento di tale propensione ad una minore capacità lavorativa (senza considerare altri fattori, molto più significativi).

In altri casi, errori e discriminazioni possono derivare da pregiudizi presenti nei dati usati nell'apprendimento automatico. Determinazioni inique e ingiuste possono derivare dall'uso di dati vantaggiosi applicabili ai soli membri di un certo gruppo (ad esempio, il fatto di aver frequentato certe scuole o Università, socialmente selettive).

L'ingiustizia può anche derivare dall'uso di dati basati su giudizi e valutazioni umane distorte, faziose, o parziali (ad esempio, le lettere di raccomandazione).

Infine, errori e discriminazioni possono risultare dal fatto che gli esempi nell'insieme di addestramento non riflettono la composizione statistica della popolazione e quindi possono condurre a determinazioni scorrette e discriminatorie rispetto a certi gruppi, che in passato non hanno avuto modo di rivestire certe posizioni (in particolare posizioni di responsabilità). Supponiamo ancora, per esempio, che nella valutazione di domande per ottenere la libertà condizionale, la presenza di precedenti penali a carico del reo abbia un peso sfavorevole, e che i membri di certi gruppi (ad esempio coloro che appartengono a una certa etnia o che professano un certo credo religioso) siano soggetti a controlli più rigorosi e stringenti, tali che la loro attività criminale è più spesso scoperta e conseguentemente condannata. Per i membri di tale gruppo, ciò comporterebbe generalmente una valutazione meno favorevole rispetto ai membri di altri gruppi, a parità di condizioni.

I membri di un determinato gruppo possono anche subire discriminazioni quando quel gruppo è rappresentato solo da un sottoinsieme molto piccolo dell'insieme di addestramento, poiché ciò ridurrà l'accuratezza delle previsioni per quel determinato gruppo (si consideri, ad esempio, il caso di un'azienda che ha abbia assunto poche donne in passato e che utilizzi i propri registri storici delle assunzioni come insieme di addestramento).

La possibilità di usare sistemi automatici per valutazioni e decisioni sugli individui - diminuendo il costo di tali valutazioni e decisioni, e accrescendo i fattori su cui esse si possono basare - ne aumenta la quantità e la pervasività. In generale, grazie all'intelligenza artificiale, qualsiasi tipo di dato personale può essere usato per analizzare, prevedere e influenzare il comportamento, un'opportunità che trasforma i dati in merce preziosa.

Attraverso le tecnologie basate sull'intelligenza artificiale e i Big Data - in combinazione con tutti i dispositivi e i sensori che tracciano quotidianamente ogni attività umana - gli individui possono essere costantemente sorvegliati e influenzati, in casi e contesti molto più numerosi, sulla base di un insieme più ampio di caratteristiche personali (che spaziano dalle condizioni economiche e di salute, al luogo di residenza, alle scelte di vita personali, fino al comportamento online e offline, ecc.).

Grazie ai sistemi informatici - e in particolare a intelligenza artificiale e big data - determinazioni errate, discriminatorie, o comunque inique (in tema di reclutamento, progressione di carriera, prestiti, premi assicurativi, ecc.) possono essere adottate in molti più casi e contesti, in base a un più ampio insieme di caratteristiche personali (dalla situazione economica, alle condizioni di salute, alla residenza, alle scelte e vicende della vita, ai comportamenti online e offline, ecc.)[16]. Essere soggetti a un persistente osservazione e valutazione genera una pesante pressione psicologica e incide sull'autonomia personale, anche quando esista una correlazione statistica tra le caratteristiche della persona interessata, e la valutazione/previsione che giustifica una certa determinazione. Inoltre, la disponibilità di valutazioni automatiche - efficienti, economiche, e apparentemente oggettive, anche quando prive di una motivazione comprensibile all'uomo - può indurre da un lato a trattare le persone in modo iniquo, senza tener conto delle speciali circostanze che le riguardano, dall'altro lato a respingere o ostacolare le contestazioni da parte degli interessati, poiché tali contestazioni, anche quando giustificate, interferiscono con l'operatività del sistema, causando costi e incertezze. Questo problema è stato efficacemente trattato da Cathy O'Neill che identifica negli algoritmi delle possibili «armi di distruzione matematica».

«Un algoritmo elabora un sacco di statistiche e ne estrae una probabilità che una certa persona passa essere una cattiva assunzione, un mutuatario rischioso, un terrorista, o un insegnante scadente. Questa probabilità è distillata in un punteggio, che può sconvolgere la vita dell'interessato. E tuttavia quando la persona reagisce, prove in contrario, meramente "evocatrici", semplicemente non funzionano. La

[16] C. O'NEIL, *Weapons of math destruction: how big data increases inequality and threatens democracy. Crown Business*, 2016.

contestazione deve essere inoppugnabile. Le vittime umane delle armi di distruzione matematica (…) sono tenute a uno standard di prova molto più alto rispetto agli stessi algoritmi»[17].

A queste critiche si è risposto osservando che se è vero che i sistemi algoritmici, e in particolare quelli basati sull'apprendimento automatico, possono riprodurre o anche esacerbare le iniquità esistenti, è anche vero che i processi algoritmici sono più controllabili di quelli sottesi alle decisioni umane, e possono essere migliorati e «ingegnerizzati» in modo da prevenire risultati iniqui.

«Mettendo in atto requisiti appropriati, l'uso degli algoritmi renderà possibile esaminare e valutare più facilmente l'intero processo di decisione, così da rendere molto più facile conoscere se si è verificata una discriminazione. Imponendo un nuovo livello di specificità, l'uso degli algoritmi mette in luce, e rende trasparenti, i compromessi tra valori in competizione. Gli algoritmi non sono solo una minaccia da regolare; accompagnati da corrette garanzie, essi hanno il potenziale per essere una forza positiva per l'equità»[18].

Non possiamo qui approfondire il tema della valutazione comparativa di decisioni umane e decisioni algoritmiche. Quello che è certo è che l'idea che solo le decisioni di routine possano essere affidate agli algoritmi, e non quelle che coinvolgono condizioni di incertezza, discrezionalità e valutazioni è oggi superata. I sistemi di intelligenza artificiale hanno dimostrato di poter operare con successo anche in ambiti in cui mancano criteri precisi e univoci, e quindi anche in settori tradizionalmente affidati all'intuizione umana, allenata con l'esercizio (per esempio, nella diagnosi medica, negli investimenti finanziari, e nella concessione del credito).
L'alternativa al processo decisionale automatizzato non sono le decisioni perfette ma le decisioni umane con tutti i loro difetti e le loro imperfezioni: anche un sistema algoritmico imperfetto potrebbe essere più giusto ed equo di un essere umano incline all'errore o al pregiudizio. Ciò non significa che in qualsiasi settore e ambito di attività la macchina prevalga sull'uomo. Quando è necessario affrontare questioni nuove, sviluppare soluzioni creative, cogliere attitudini, preferenze e interessi umani, la mente dell'uomo è insostituibile. La sfida per il futuro è trovare le migliori combinazioni tra intelligenza umana e artificiale, valutazioni umane e valutazioni automatizzate, integrandole fra loro, e tenendo conto delle potenzialità di ciascuna. Inoltre, l'intelligenza artificiale può essere utilizzata per controllare le sue stesse applicazioni,

[17] C. O'NEIL, W*eapons of math destruction: how big data increases inequality and threatens democracy*, cit., p. 16. (traduzione degli autori).
[18] J. KLEINBERG et al, *Discrimination in the age of algorithm*, in *Journal of Legal Analysis*, 10, n. 14/2018, p. 113 (traduzione degli autori).

così da individuare eventuali difetti nei meccanismi della decisione automatica, e aiutare nella predisposizione di contestazioni.

5. L'ecosistema della sorveglianza

L'uso dell'intelligenza artificiale, in combinazione con i grandi dati, ha trovato ampi sbocchi tanto nell'economia privata quanto nell'amministrazione pubblica. Alcuni autori hanno visto in modo positivo lo sviluppo di applicazioni basate sulla raccolta massiva di informazioni, anche sui comportamenti degli individui, osservando che l'integrazione di intelligenza artificiale e big data consente maggiore efficienza, e fornisce nuovi mezzi di direzione e controllo sul comportamento sociale[19].

Così Al Varian, che ha svolto un ruolo chiave nella definizione dei metodi economici usati da Google, osserva che quando gli scambi economici - e più in generale, le interazioni sociali e le attività degli individui - sono mediati da computer, essi danno luogo a continue e dettagliate registrazioni di dati: il computer può osservare e verificare ogni aspetto dell'attività in cui è coinvolto. I dati raccolti possono essere usati per analisi, valutazioni e feedback che consentono di personalizzare la relazione con gli interessati (come nella pubblicità personalizzata), di effettuare continue sperimentazioni (per esempio, valutando le risposte a cambiamenti nei prezzi o nei messaggi), di guidare e controllare i comportamenti. Diventano così possibili nuovi, più ampi ed efficienti, modelli di interazione economica e sociale, basati non più sull'incerta aspettativa del rispetto di norme sociale, morali e giuridiche, o in vacillanti atteggiamenti di reciprocità e collaborazione, ma piuttosto nella possibilità di osservare ogni comportamento e di collegare ad esso sanzioni e ricompense. Si pensi a come tutti noi ci affidiamo a venditori di beni o fornitori di servizi con cui non abbiamo avuto alcun contatto personale, confidando nella piattaforma attraverso cui beni e servizi vengono forniti, e nei meccanismi di valutazione e votazione (*rating* e *scoring*), selezione ed esclusione attuati dalla piattaforma. Si pensi altresì a come sistemi basati sulla catena dei blocchi (*blockchain*) consentano di creare monete digitali, meccanismi contrattuali che si auto-eseguono (*smart contract*), e organizzazioni economiche digitali, sulla base di un archivio immodificabile, replicato in tutti i nodi del sistema, in cui vengono registrate tutte le transazioni.

Alex Pentland, che dirige lo *Human Dynamics Lab* presso il celebre *MIT Media Lab*, ha affermato che intelligenza artificiale e big data offrono la prospettiva di una «fisica sociale», quale scienza in grado di capire e guidare la società. La disponibilità di grandi masse di dati e di metodi e risorse computazionali per elaborarli consentirebbe di realizzare finalmente il sogno di August Comte, cioè di ottenere una scienza sociale dotata di fondamenti teorico-matematici così come di capacità operative.

[19] H.R. VARIAN, *Computer Mediated Transactions*, cit.

«Attraverso una migliore conoscenza di noi stessi, potremmo potenzialmente costruire un mondo senza guerra o crisi finanziarie, in cui le malattie infettive siano rapidamente individuate e fermate, in cui l'energia, l'acqua e le altre risorse non siano più sprecate, e in cui i governi siano parte della soluzione invece che parte del problema»[20].

Alle prospettive di crescita economica e sociale offerte dall'integrazione di intelligenza artificiale e big data si accompagno i rischi associati sia al «capitalismo di sorveglianza»[21], sia allo «Stato di sorveglianza»[22]. Shoshana Zuboff identifica infatti nel «capitalismo di sorveglianza» (*surveillance capitalism*) il modello economico tipico della nostra epoca. Zuboff riprende il classico lavoro di Karl Polanyi[23], il quale aveva osservato che l'avvento del capitalismo, la «grande trasformazione», è consistito nell'occupazione prima dell'economia e poi di ambiti crescenti dello spazio sociale da parte del mercato. In particolare, Polanyi osservava che il capitalismo trasforma in merci (prodotti da vendere sul mercato) anche entità che non sono state prodotte per il mercato: la terra (l'ambiente), il lavoro, e il denaro. Ne risultano tensioni distruttive per l'intera società, se le dinamiche del mercato non sono soggette a limiti e controlli, da parte del diritto, dalla politica e dei movimenti sociali (come quelli di lavoratori e consumatori). Nel capitalismo della sorveglianza, afferma Zuboff, il dominio del mercato si estende all'esperienza umana: il comportamento degli individui viene registrato ed analizzato e i dati, le previsioni e le conseguenti capacità di influenza, diventano una nuova merce.

«Il capitalismo della sorveglianza annette l'esperienza umana alla dinamica del mercato, cosicché essa rinasce come comportamento: la quarta «merce fittizia». Le prime tre merci fittizie - la terra, il lavoro e il denaro - sono state assoggettate alla legge. Nonostante queste leggi siano state imperfette, le istituzioni del diritto del lavoro, del diritto dell'ambiente e del diritto bancario sono quadri di regolazione volti a difendere la società (e la natura, la vita e lo scambio) dai peggiori eccessi del potere distruttivo del capitalismo allo stato grezzo. L'espropriazione dell'esperienza umana da parte del capitalismo della sorveglianza non ha incontrato impedimenti siffatti»[24].

[20] A. PENTLAND, *Social Physics: How Social Networks Can Make Us Smarter*, New York, 2015, p. 28.
[21] S. ZUBOFF, *The Age of Surveillance Capitalism*, Londra, 2019.
[22] J.M. BALKIN, *The constitution in the national surveillance state*, in *Minnesota Law Review*, 93, 2008, p 3 (traduzione degli autori).
[23] K. POLANYI, *The Great Transfomation*, Boston, 2001 [1944].
[24] S. ZUBOFF, *The Age of Surveillance Capitalism*, cit., p. 514 (traduzione degli autori).

Zuboff osserva che anche nel caso del capitalismo della sorveglianza le dinamiche del mercato, se lasciate a sé stesse, conducono a esiti distruttivi. Le persone sono soggette a manipolazione, sono private del controllo sul proprio futuro e di uno spazio in cui sviluppare la propria personalità. Le reti sociali di collaborazione vengono sostituite da meccanismi di incentivi e disincentivi che danno luogo a nuove forme di controllo e sfruttamento. Si pensi alle dinamiche generate dalle piattaforme di Internet, che monitorando il comportamento degli utenti e analizzandolo mediante tecniche di intelligenza artificiale sono in grado di influenzare le scelte degli utenti stessi. Si pensi altresì alle piattaforme per la fornitura di servizi - come Uber o Lyft nel campo dei trasporti - che registrano ogni attività svolta dai prestatori del servizio, le modalità in cui tali attività sono svolte, le reciproche valutazioni di prestatori e clienti, e associano ricompense e penalità ad ogni aspetto di tali attività, così da guidare il comportamento di ciascuno nel senso desiderato. Si tratta di un nuovo sistema di governo del comportamento umano (che sostituisce i tradizionali meccanismi del diritto e del contratto) con esiti economicamente efficienti, ma potenzialmente negativi per il benessere mentale e l'autonomia degli interessati[25]. Mentre il capitalismo classico trovò correttivi sufficienti a limitarne gli esiti più distruttivi, secondo Zuboff mancano a tutt'oggi risposte adeguate ai nuovi rischi rappresentati dal capitalismo di sorveglianza.

Il capitalismo di sorveglianza trova un corrispettivo nella dimensione pubblica. Qui è emerso negli ultimi decenni il modello dello "Stato di Sorveglianza", caratterizzato:

«dalla raccolta e dal confronto di informazioni sui cittadini, e dall'analisi di tali informazioni per identificare problemi, prevenire minacce potenziali, governare la popolazione, e fornire utili servizi sociali. Lo Stato di sorveglianza è un caso particolare dello Stato dell'informazione, uno stato che cerca di identificare e risolvere i problemi di *governance* attraverso la raccolta, il confronto, l'analisi e la produzione di informazioni»[26].

Anche in questo caso, ai possibili vantaggi relativi all'efficienza nella gestione delle attività pubbliche, al coordinamento dei comportamenti dei cittadini, alla prevenzione dei rischi, si accompagnano prospettive inquietanti, nuove forme di influenza e controllo volte ad asservire i cittadini rispetto agli scopi e ai valori di chi controlla l'infrastruttura di sorveglianza.

6. Il business model della pubblicità online e la formazione dell'opinione pubblica.

Le tecniche per la sorveglianza, la profilazione, la persuasione e la decisione algoritmica sono state elaborate soprattutto in ambiti commerciali, in particolare nel contesto del commercio elettronico. La raccolta di dati sugli utenti, l'anticipazione delle loro attitudini e preferenze, la conseguente possibilità di

[25] N. CRISTIANINI - T. SCANTAMBURLO, *On social machines for algorithmic regulation*, in *AI and Society*, 2019.
[26] J.M. BALKIN, *The constitution in the national surveillance state*, 93, n. 1/2008, p. 3 (traduzione degli autori).

influenzare il comportamento individuale, hanno consentito di realizzare forme particolarmente efficaci di pubblicità mirata o micromirata (*microtargeting*) rispetto alle caratteristiche dei singoli utenti. La possibilità di inviare pubblicità mirata, a propria volta, ha conferito un vantaggio competitivo alle piattaforme per il commercio elettronico, che in pochi anni hanno acquisito la fetta più importante del mercato pubblicitario, traendone enormi profitti.

Il modello di affari (*business model*) basato sull'invio di pubblicità mirata, non solo ha determinato la sorveglianza e l'influenzamento algoritmico a fini pubblicitari ma, ha altresì stimolato l'adozione generalizzata di metodi per l'invio di messaggi e notizie personalizzati. Tali metodi si basano sull'esigenza di trattenere gli utenti all'interno delle piattaforme, o dei siti, affinché gli stessi utenti possano essere oggetto di messaggi pubblicitari[27].

All'interno delle piattaforme, la stessa struttura di comunicazione basata sull'invio di messaggi che confermano preferenze e attitudini del lettore, e sull'attivazione di corrispondenti relazioni interpersonali, ha un impatto sulla formazione dell'opinione pubblica. Il funzionamento delle piattaforme tende infatti a favorire la polarizzazione delle opinioni[28], e la suddivisione dei cittadini in gruppi ideologicamente omogenei e non comunicanti.

Inoltre, gli incentivi della pubblicità online hanno favorito la proliferazione dei contenuti che, come le cosiddette *fake news*[29], maggiormente attirano l'attenzione degli utenti, esponendoli ai messaggi pubblicitari. Il termine *fake news* identifica il fenomeno relativo alla generazione di contenuti distorti, fuorvianti, e/o falsi, generalmente "micro-mirati" (*micro-targeted*) e distribuiti online al fine di influenzare le opinioni di singoli individui e gruppi, profittando delle loro debolezze e insicurezze. Inoltre, le piattaforme e i produttori di *fake news* ottengono ingenti profitti dalla loro pubblicazione e diffusione, il che disincentiva i controlli da parte delle piattaforme stesse.

Si consideri, per esempio, il modello di business alla base della proliferazione di *fake new*s su Facebook[30]. Tale modello comprende i seguenti momenti:

1. Gli individui interessati alla diffusione di *fake news* pubblicano le stesse nelle pagine di siti web dedicati a tale uso. Tali pagine contengono inserzioni pubblicitarie. Il titolare del sito può essere anche il produttore delle notizie, o un terzo interessato ad avvalersi del sito.

[27] E. PARISER, *The filter bubble,* cit.
[28] C.R. SUNSTEIN. *Algorithms, correcting biases,* cit.
[29] Su cui si veda in questo numero speciale M. CAVINO, *Il triceratopo di Spielberg. Fake news, diritto e politica*, in *questo fascicolo*. V. anche G. RUFFO – M. TAMBUSCIO, *Capire la diffusione della disinformazione e come contrastarla*, in *questo fascicolo*.
[30] K. HAENSCHEN - P. ELLENBOGEN, *Disrupting The Business Model of the Fake News Industry*, in *Freedom to thinker: research and expert commentary on digital technologies in public life*, February 29, 2020, disponibile al seguente link: <https://freedom-to-tinker.com/2016/12/14/disrupting-the-business-model-of-the-fake-news-industry/>.

2. Gli stessi individui acquistano spazi pubblicitari da Facebook, per distribuire agli utenti della rete sociale i link alle pagine contenenti le *fake news*. Facebook propone i link nelle *newsfeed* (aggionamenti alle notizie) degli utenti di Facebook (selezionati in base a vari criteri).
3. Facebook percepisce introiti pubblicitari in base al numero di click sui link proposti ai propri utenti.
4. Gli utenti che cliccano sul link di una *fake news* vengono indirizzati alla pagina contenente tale notizia nel sito di *fake news*, generando una registrazione per ogni pubblicità presentata loro durante la visita del sito.
5. Il sito di *fake news* viene retribuito dagli inserzionisti sulla base delle registrazioni (di accessi alla pubblicità) generate dalle visite degli utenti, e i relativi profitti possono essere condivisi con chi ha generato le notizie.

A questo ciclo principale, basato su siti specializzati, si può aggiungere un ciclo secondario.

6. I produttori di *fake news* oltre a pubblicarle su siti dedicati a *fake news* (punto 1 sopra), possono pubblicare le stesse notizie sulla loro pagina Facebook, e sponsorizzare la propria pagina, per ottenere delle adesioni da parte di altri utenti, i cosiddetti seguaci (*followers*)
7. Le *fake news* pubblicate nella pagina Facebook del produttore sono visualizzate automaticamente nelle *newsfeed* dei seguaci, che a loro volta condividono il link alla notizia con i propri contatti. In questo modo la notizia diventa virale, grazie alla sua ripetuta replicazione.
8. Se la pagina Facebook del produttore di *fake news* contiene inserzioni pubblicitarie, il produttore, a propria volta, ne trae un profitto.

Sulla base dei meccanismi appena indicati, chi pubblica le *fake news*, e gestisce i siti relativi trae proventi pubblicitari, cosicché un'attività dannosa per la società diventa vantaggiosa per chi la pone in essere.

Le *fake news* tendono a diventare virali specialmente nei momenti di crisi sociale ed economica, quando gli individui cercano spiegazioni e possibilmente responsabili per i propri disagi

«ogni volta che incombe una minaccia o si è verificato un evento spaventoso, è inevitabile che si diffondano voci incontrollate (…). Nel periodo che segue una crisi vengono fatte molte congetture. Alcuni le riterranno plausibili, forse perché offrono al risentimento e al desiderio di trovare un colpevole. Gli eventi tragici generano risentimento, e in un simile stato d'animo le persone accettano molto più facilmente le dicerie che giustificano le loro condizioni emotive, e sono inoltre più propense ad attribuire gli eventi ad azioni intenzionali»[31].

[31] C.R. SUNSTEIN, *On rumors: How falsehoods spread, why we believe them, and what can be done*, Princeton, 2014, pp. 24-25 (traduzione degli autori).

7. Profilazione e intelligenza artificiale nella comunicazione politica

Come abbiamo visto nella sezione precedente, i meccanismi dominanti nell'economia della rete possono condurre ad usi dell'intelligenza artificiale che hanno un impatto negativo sulla formazione dell'opinione pubblica e quindi sull'assetto democratico della società, anche in assenza di influenzamento politico intenzionale. Infatti, la formazione di un'opinione pubblica consapevole presuppone il confronto delle diverse opinioni e la conoscenza dei fatti rilevanti. Invece, il meccanismo per l'invio di informazioni mirate e gradite al destinatario (la c.d. bolla del filtro) fa sì che ciascun individuo non abbia la possibilità di accedere a informazioni nuove e a opinioni diverse dalla propria. Inoltre, il meccanismo delle *fake news* genera disinformazione, sfiducia nelle istituzioni e nei media tradizionali, e credenze incompatibili con la realtà dei fatti.

Le tecniche dell'intelligenza artificiale possono però essere consapevolmente utilizzate anche nella comunicazione politica, influenzando le modalità del dibattito, la propaganda, e lo stesso esito delle campagne elettorali.

Da un lato, l'intelligenza artificiale può contribuire a fornire ai cittadini maggiori informazioni, meglio corrispondenti agli interessi, ai bisogni, e alle modalità cognitive di ciascuno, e può facilitare altresì l'istaurazione di interazioni, l'aggregazione delle opinioni, e lo sviluppo di iniziative collettive. Dall'altro lato però, l'intelligenza artificiale può essere usata intenzionalmente a fini di disinformazione e manipolazione delle opinioni politiche e del comportamento elettorale. L'effetto delle tecnologie per l'influenzamento è accresciuto dalla sinergia con la polarizzazione e la disinformazione che risultano dai meccanismi economici della rete.

I pericoli derivanti dall'uso scientifico delle tecniche per la disinformazione, l'influenza e la manipolazione a fini politici sono emersi con chiarezza nel caso di Cambridge Analytica, in merito ai tentativi di influenzare le preferenze di voto dei cittadini durante le elezioni presidenziali americane del 2016 e probabilmente anche durante il referendum sulla Brexit[32].

Il coinvolgimento di Cambridge Analytica nelle elezioni presidenziali americane, può essere analizzato in quattro fasi principali[33]. di seguito descritte.

[32] Sulla relazione tra populismo, democrazia e Internet, vedi M. BARBERIS, *Come Internet sta uccidendo la democrazia*, Milano, 2020.

[33] A. HERN, *Cambridge Analytica: how did it turn clicks into votes*, in *The Guardian*, 6 maggio 2018, disponibile al seguente link <https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie>.

- Fase 1: in prima istanza, i cittadini registrati come elettori negli Stati Uniti furono invitati a effettuare un test della personalità per valutare il proprio profilo psicologico, sulla base dei seguenti fattori: apertura verso nuove esperienze, coscienza, estroversione, gradevolezza ed emotività. Il modello algoritmico utilizzato durante il test era stato addestrato per distinguere gli individui, suddividendoli in gruppi, sulla base di tratti comuni della personalità. Quindi, per esempio, coloro che si definivano "chiassosi" venivano classificati anche come verosimilmente "socievoli". I partecipanti che non si ritrovavano in questa prima descrizione venivano classificati come verosimilmente molto diversi da coloro che vi si riconoscevano. Al fine di incentivare la partecipazione al test, disponibile online e basato su circa 120 domande, ai volontari fu promesso un premio consistente in una piccola somma di denaro (da due a cinque dollari). Ai partecipanti fu anche detto che i dati raccolti sarebbero stati utilizzati unicamente a fini di ricerca in ambito accademico. Circa 320.000 individui parteciparono al test. Per ricevere la ricompensa, a ciascun partecipante fu richiesto l'accesso al rispettivo profilo Facebook.
- Fase 2: L'accesso ai profili Facebook consentì di correlare le risposte di ciascun partecipante con le informazioni e le attività presenti sul profilo (per esempio *like*, condivisioni e altri indicatori). Dopo aver ottenuto l'accesso ai profili, Cambridge Analytica raccolse non solo i dati ivi contenuti ma anche le informazioni presenti sui profili dei contatti di coloro che avevano partecipato al test, fino a coinvolgere complessivamente tra i 30 e i 50 milioni di individui.
- Fase 3: una volta terminata la raccolta dei dati ,Cambridge Analytica ebbe a disposizione due tipologie di dati personali da elaborare. Da un alto, i dati dei partecipanti al test, vale a dire le informazioni raccolte dai rispettivi profili, combinate con le risposte fornite durante il questionario, e dall'altro, le informazioni sui loro contatti, relative al solo contenuto dei profili Facebook. I dati dei partecipanti al questionario furono utilizzati da Cambridge Analytica come insieme di addestramento, per la costruzione di un modello per la profilazione di coloro che non avevano partecipato al questionario, ma di cui erano state raccolte le informazioni e di altri individui con caratteristiche simili. In particolare, l'insieme di addestramento era costituito dalle informazioni estratte dai profili Facebook dei partecipanti (*like*, *post*, condivisioni, etc), utilizzate come dati di input (caratteristiche), e dalle risposte al questionario (e dalle attitudini psicologiche e politiche correlate a ciascuna risposta), utilizzate come variabili da prevedere (*target variables*). Grazie ai metodi di apprendimento automatico Cambridge Analytica costruì il modello capace di correlare

le informazioni contenute nei profili Facebook di ciascun individuo con le previsioni relative ai profili psicologici e alle preferenze di voto. Ciò permise a Cambridge Analytica di compiere una profilazione massiva, ampliando i dati disponibili di coloro che non avevano partecipato al test (le informazioni sui profili Facebook e altri dati disponibili online), con le previsioni fornite dal modello addestrato. Se, per esempio, i partecipanti con certe caratteristiche comuni (il tipo di *like* e condivisioni) erano stati classificati come dotati di una personalità nevrotica, la stessa valutazione veniva estesa a coloro che non avevano partecipato al questionario e che tuttavia presentavano caratteristiche simili ai primi, relativamente alle loro informazioni personali presenti sui profili Facebook.

- Fase 4: sulla base di tale profilazione psicologica e politica, Cambridge Analytica identificò gli elettori indecisi, vale a dire coloro che avrebbero potuto cambiare le proprie preferenze di voto se debitamente sollecitati, e li sottopose a comunicazioni politiche mirate e all'invio di messaggi capaci di innescare il cambiamento desiderato, basandosi sull'emotività e i pregiudizi da cui erano affetti, e senza che tali individui fossero consapevoli dello scopo di tali messaggi.

Il caso di Cambridge Analytica mostra come la combinazione e l'analisi dei dati possano rivelare dettagli profondamente personali e granulari su ciascun individuo. Tali dettagli posso essere utilizzati per creare e inviare messaggi personalizzati e influenzare emotivamente scelte che idealmente dovrebbero essere deliberative, private e ponderate. La profilazione psicometrica può essere utilizza per manipolare i comportamenti degli individui, limitandone così il libero arbitrio e la capacità di autodeterminazione e compromettendo alcuni diritti fondamentali come il diritto alla privacy e alla protezione dei dati, vale a dire all'uso legittimo e proporzionato dei dati personali. Ciò è difficilmente compatibile con un ambiente online nel quale ogni comportamento sia registrato, e le relative informazioni siano utilizzate per estrarre ulteriori conoscenze sugli individui, al di fuori del loro controllo, e per elaborare tali conoscenze in modi potenzialmente contrari agli interessi degli stessi individui, per scopi non condivisi, possibilmente violando le aspettative fiduciarie riposte su chi controlla i sistemi di IA in questione[34].

L'uso delle tecnologie di Internet e dell'intelligenza artificiale all'interno della comunicazione politica trova ulteriore riscontro nella diffusione sempre maggiore dei cosiddetti bot politici all'interno delle piattaforme online. Si tratta di software progettati per generare contenuti, condividere notizie e informazioni e interagire con gli delle piattaforme utenti in modo automatico, all'interno della rete. I bot

[34] J.M. BALKIN, *The Three Laws of Robotics in the Age of Big Data*, in *Ohio State Journal Law Journal*, 2017, pp. 1217-1241.

generano una quantità crescente di informazioni e traffico online e rappresentano una parte significativa dei profili attivi sulle piattaforme, per esempio, governando solo su Twitter circa 30 milioni di account[35]. I bot politici possono essere utilizzati per simulare una maggiore popolarità di personaggi politici, contribuendo ad aumentare in modo fittizio il numero di seguaci, e per diffondere notizie all'interno delle piattaforme, aumentandone la replicazione, o ancora per disturbare la comunicazione politica degli avversari e dar vita a campagne diffamatorie. Per esempio, essi possono generare etichette (*hastag)* prive di rispondenza alle notizie e alle immagini cui sono associate, o essi possono cercare notizie e dichiarazioni politiche al fine di negarle, generando così disinformazione, e manipolando l'opinione pubblica[36]. I bot politici sono generalmente caratterizzati da una forte polarizzazione ideologica e programmati per promuovere una particolare opinione all'interno del dialogo politico.

L'uso di bot, in combinazione con le tecnologie dell'IA e dei big data può intensificare gli effetti negativi della cosiddetta propaganda computazionale, derivante dalla combinazione di tecniche di profilazione per la raccolta massiva di informazioni, decisioni algoritmiche, e pubblicità politica micro-mirata.

8. Nuove dimensioni della sorveglianza: il caso del Sistema di credito Sociale Cinese

Nelle sezioni precedenti si è esaminato come l'intelligenza artificiale possa essere usata in modo da disinformare e manipolare i cittadini. A questi rischi si affiancano quelli connessi all'uso delle stesse tecnologie, per sorvegliare, valutare, e indirizzare i comportamenti dei cittadini. Mentre nel primo caso, si manipolano le opinioni in modo da indurre comportamenti desiderati, nel secondo si influenza direttamente il comportamento mediante (micro) ricompense e sanzioni.

Un esempio paradigmatico è rappresentato da un'iniziativa del governo cinese, il cosiddetto Sistema di Credito Sociale (SCS), che assegna ad ogni cittadino un punteggio che ne quantifica il valore sociale e ne oggettiva la reputazione. Tale sistema si basa sull'aggregazione e l'analisi di informazioni relative alla sfera pubblica e privata della vita dei singoli cittadini, come i loro comportamenti finanziari (per esempio, la capacità di ottemperare puntualmente alle proprie obbligazioni e onorare i contratti stipulati), politici (per esempio, la partecipazione a movimenti e manifestazioni), giudiziari (per esempio, procedimenti passati e presenti), e sociali (per esempio, partecipazione a reti sociali, rapporti interpersonali, etc.). Ad ogni comportamento rilevante - e l'ambito dei comportamenti rilevanti è in linea di principio illimitato - il

[35] J. MOTTI, *Twitter acknowledges 23 million active users are actually bots*, in *Tech Times*, 12 Agosto 2014, disponibile al seguente link: <https://www.techtimes.com/articles/12840/20140812/twitter-acknowledges-14-percent-users-bots-5-percent-spam-bots.htm>.

[36] Si vedano, S. WOOLLEY - P.N. HOWARD, *Social media, revolution, and the rise of the political bot*, in P. ROBINSON - P. SEIB - R. FROHLICH (a cura di), *Routledge handbook of media, conflict, and security*, New York, 2016; P.N. HOWARD - S. WOOLLEY - R. CALO, *Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration*, in *Journal of information technology & politics*, 15, n. 2/2018, pp. 81-93.

sistema assegna un punteggio positivo o negativo. Il totale dei punteggi costituisce il credito sociale di ciascun individuo, e ne determina l'accesso a servizi (come università, locazioni, e trasporti), posizioni lavorative, finanziamenti, ecc.

Anziché promuovere la virtù civica dei cittadini e la reciproca fiducia, come affermano le autorità cinesi, il Sistema di Credito Sociale rischia di promuovere comportamenti basati su opportunismo e conformismo, a discapito dell'autonomia dei singoli e delle loro genuine motivazioni morali e sociali.

Mentre secondo alcuni[37], non vi sarebbero differenze sostanziali tra il sistema di punteggio alla base dell'iniziativa del governo cinese e i sistemi di punteggio utilizzati nei paesi occidentali, nell'ambito di iniziative commerciali, altri ne hanno sottolineato le peculiarità[38].

In entrambi i casi, la sorveglianza pervasiva e la valutazione persistente possono innescare comportamenti di sottomissione, conformismo e adattamento alle regole, motivati unicamente dal desiderio di evitare sanzioni e ottenere ricompense. Ciò potrebbe sostituire ed eliminare l'autenticità dell'esperienza umana, del ragionamento e delle scelte morali e giuridiche, che richiedono a ciascun individuo di affrontare il conflitto tra interesse personale e imperativi morali o giuridici. Adottando una prospettiva Kantiana, infatti, l'esercizio della moralità richiede necessariamente la possibilità dell'immoralità, che sarebbe esclusa da un sistema di valutazione pervasivo, anche qualora i punteggi assegnati a ciascun individuo fossero associati in modo corretto a comportamenti pro-sociali (cosa che potrebbe rivelarsi errata o falsa a causa delle contingenti spinte economiche e politiche).

Si è sostenuto che i sistemi di classificazione e valutazione possono indurre conformità e sottomissione piuttosto che creatività e autonomia[39]. Tuttavia, un tale effetto dipende principalmente dai metodi utilizzati per l'assegnazione dei punteggi: se basati su fattori scalabili (per esempio, risultati scolastici, tempi di risposta, valutazioni degli utenti, ecc.), allora è possibile che essi incentivino l'eccellenza, sebbene secondo criteri prestabiliti. L'assegnazione di punteggi, tuttavia, produce inevitabilmente confronti, competizione, stress, problemi relazionali, con il rischio che gli individui investano in modo eccessivo nel raggiungimento di punteggi elevati al solo fine di primeggiare. Poiché non tutti possono eccellere, alcuni rimarranno necessariamente indietro, e ciò può comportare non solo risultati negativi circoscritti, come licenziamenti o mancanza di opportunità, ma anche perdita di autostima, emarginazione, ecc. Inoltre, la

[37] D. MAC SÍTHIGH - M. SIEMS, *The Chinese social credit system: A model for other countries?*, in *The Modern Law Review*, 82, n. 6/2019, pp. 1034-1071.

[38] C. FORD, *Seeing Like an Authoritarian State*, in L. ORGAD – W. REIJERS (a cura di), *A Dystopian Future? The Rise of Social Credit Systems*, Robert Schuman Centre for Advanced Studies Global Citizenship Governance, European University Institute Working Papers RSCAS, n. 94/2019, pp. 15-17.

[39] ORGAD - REIJERS, *A Dystopian Future? The Rise of Social Credit Systems*, cit.

difficoltà di migliorare il proprio punteggio (di riabilitazione o redenzione) può portare al disinvestimento e alla rinuncia.

Se si guarda invece agli scopi perseguiti dai sistemi di valutazione, nel contesto del capitalismo della sorveglianza, questi sono generalmente guidati da obiettivi economici, e talvolta anche politici, seppure indirettamente (come nel caso di operatori ingaggiati per influenzare i processi elettorali dietro pagamento di denaro). Alcuni sistemi di valutazione, possono essere basati sul cosiddetto *co-scoring*, dove gli individui si valutano reciprocamente (come nel caso di Airbnb ed eBay) o partecipano in modo collaborativo alla valutazione di prodotti ed esperienze (per esempio, attraverso il numero di *like*). Tali valutazioni possono essere utilizzate come input per l'elaborazione di punteggi ulteriori da parte delle società proprietarie dell'infrastruttura. Per esempio, i punteggi ottenuti dagli utenti di piattaforme di e-commerce possono essere utilizzati per premiare o punire (per esempio mediante esclusione dalla piattaforma) venditori, fornitori di servizi, acquirenti, ecc. I punteggi possono anche essere utilizzati per identificare le preferenze degli utenti e inviare loro pubblicità mirate o selezionare le informazioni da visualizzare.

Sebbene nei paesi occidentali, i sistemi di punteggio riguardino principalmente il settore privato, essi possono essere utilizzati anche dalle amministrazioni per misurare le prestazioni dei dipendenti pubblici così come nel settore della giustizia. Uno dei casi più controversi riguarda l'uso di sistemi predittivi per determinare il rischio di recidiva, che a propria volta influisce sul verdetto dei giudici[40].

Le considerazioni fatte fin qui si applicano ai sistemi automatizzati di valutazione e punteggio su larga scala, sia che questi riguardino i paesi occidentali sia il Sistema di Credito Sociale Cinese. Tuttavia, quest'ultimo presenta alcune peculiarità.

In primo luogo, il Sistema di Credito Sociale Cinese è basato su un unico punteggio potenzialmente relativo a molteplici aspetti della vita di un individuo e ottenuto mediante l'aggregazione dei diversi punteggi a questi relativi. Tale aspetto è particolarmente significativo rispetto alle società occidentali, dove i sistemi di valutazione guardano ai cittadini come "dividui" (cd. *dividuals*)[41]. Ogni sistema di valutazione e punteggio è infatti basato su singoli ambiti che riguardano la vita di ciascuno (per esempio, in quanto consumatore, debitore, giocatore, lavoratore, assicurato, paziente, medico, studioso, ecc.), classificato in relazione ai fattori pertinenti a tali ambiti. Da questa prospettiva, e abusando dell'approccio di Walzer,[42] si potrebbe affermare che ciascun sistema di punteggio applica i criteri di una diversa "sfera di giustizia" o meglio di una diversa "sfera di valutazione", poiché i criteri utilizzati non riflettono necessariamente criteri di giustizia accettabili. Una valutazione frammentata ha un impatto minore sull'autostima dei

[40] J. LARSON - L. KIRCHNER - J. ANGWIN, *How we analized the Compas recidivism algorithm*, in *ProPublica*, 2018.
[41] G. DELEUZE, *Postscriptum sur les sociétés de contrôle par gilles deleuze*, in *L'Autre Jurnal*, 1990.
[42] M. WALZER, *Spheres of Justice*, New York, 1983.

singoli così come sulla valutazione sociale, poiché ogni individuo può differenziare la propria identità sulla base di ciascun punteggio settoriale, il cui impatto è limitato al contesto corrispondente. Quando invece le diverse valutazioni vengono aggregate in un unico punteggio, come nel caso del Sistema di Credito Sociale Cinese, questo diviene l'unico indicatore del merito personale e qualsiasi comportamento è in grado di innescare effetti sui molteplici ambiti della vita di un individuo, a causa del suo impatto sul punteggio complessivo. Un punteggio globale, anziché settoriale, potrebbe condurre più facilmente all'eliminazione della riflessione e del pensiero morale, poiché tale punteggio sarebbe considerato l'unico indicatore della virtù individuale.

Una seconda caratteristica del Sistema di Credito Sociale Cinese riguarda il fatto che esso è gestito da un regime autoritario. Non vi è alcuna assicurazione che i meccanismi di classificazione e valutazione corrispondano a una nozione ragionevole e accettabile di virtù civica, determinata dalle autorità politiche cinesi (senza neppure i vincoli che caratterizzano la rappresentanza democratica e lo stato di diritto). É bene sottolineare, tuttavia, che anche nel caso in cui un sistema di valutazione e punteggio globale, basato sulla sorveglianza pervasiva degli individui, venisse utilizzato da pubblici poteri all'interno di una democrazia, questo produrrebbe i medesimi effetti e solleverebbe le medesime questioni descritte fin qui.

9. Conclusioni

L'intelligenza artificiale costituisce forse la principale sfida alla quale l'umanità dovrà far fronte nei prossimi decenni. Si aprono grandi opportunità di progresso individuale e sociale, e si prospettano, al tempo stesso, gravi rischi, anche nel campo della formazione dell'opinione pubblica e del funzionamento dei processi democratici.

Da un lato, l'intelligenza artificiale può contribuire all'informazione dei cittadini, allo sviluppo del dibattito pubblico, alla nascita di nuove forme di aggregazione, alla formazione di opinioni razionali, basate su dati di fatto e risultati scientifici. D'altro lato, essa può invece incidere negativamente sul dibattito democratico, sulle elezioni e sul funzionamento delle istituzioni.

Nel presente contributo ci siamo soffermati sulle possibili disfunzioni indotte dall'AI. In particolare, abbiamo esaminato i rischi della profilazione e delle decisioni automatiche, della polarizzazione, della disinformazione, della manipolazione e infine della sorveglianza. L'analisi realistica dei pericoli presenti e l'anticipazione di quelli futuri non deve condurre, tuttavia, a forme di rifiuto generalizzato delle tecnologie dell'intelligenza artificiale. Non solo tali tecnologie possono portare enormi benefici, ma la loro diffusione a livello globale è spinta da dinamiche economiche e sociali che non possono essere fermate. Bisogna invece indirizzare l'uso di tali tecnologie verso risultati benefici, e prevenirne i possibili esiti negativi con adeguate misure politiche, giuridiche e tecnologiche.

Importanti norme in tema di informazione, profilazione e decisione automatica sono contenute nel nuovo Regolamento Europeo sulla protezione dei dati (GDPR).

Il GDPR sancisce, infatti, un generale divieto di sottoporre gli individui a processi decisionali completamente automatizzati, compresa la profilazione, in grado di produrre effetti giuridici o di incidere in modo significativo sull'interessato. La profilazione è in linea di principio vietata dal nuovo regolamento, ma vi sono ampie eccezioni. Essa è consentita qualora il trattamento (i) sia autorizzato da una legge o un regolamento, che prevede altresì misure idonee a tutelare i diritti dei soggetti interessati, (ii) sia necessario per la conclusione o l'esecuzione di un contratto tra l'interessato e il titolare o (iii) sia basato sul consenso esplicito dell'interessato. Inoltre, in caso di profilazione, il consenso deve essere distinto rispetto al consenso relativo ad altri tipi di trattamento. Rispetto alle ultime due ipotesi (contratto e consenso) il titolare dovrà, inoltre, attuare misure appropriate per tutelare i diritti, le libertà e gli interessi legittimi dell'interessato.

Al fine di determinare la liceità della profilazione, anche in presenza di una base giuridica, le linee guida indicano i seguenti elementi (i) il livello di dettaglio e la completezza del profilo (se questo descrive solo aspetti parziali della persona interessata, o ne ricostruisce un quadro più completo); (ii) l'impatto della profilazione sull'interessato; e (iii) le misure di sicurezza volte ad assicurare equità, non discriminazione e accuratezza nel processo di profilazione.

Inoltre, nelle ipotesi di profilazione e processi decisionali automatizzati, il GDPR garantisce all'interessato il diritto di esserne informato (art. 22 GDPR e Considerando 71). In particolare, in capo al titolare del trattamento sussiste l'obbligo di informare gli interessati circa (i) le modalità e le finalità della profilazione, (ii) la logica inerente il trattamento e (iii) le conseguenze di tale tipo di trattamento, rispetto ai criteri utilizzati nel processo decisionale. Il diritto di informazione è accompagnato dal diritto di opporsi alla profilazione (articolo 21), di richiedere la cancellazione dei propri dati e del proprio profilo (articolo 17) e di contestare le decisioni automatizzate (articolo 22(3)).

Recentemente sono state adottate alcune iniziative più specificamente dirette alla disciplina dell'intelligenza artificiale. Un primo passo in questa direzione può ravvisarsi nella recente comunicazione della Commissione Europea sulla lotta alla disinformazione online[43], per intensificare gli sforzi volti a contrastare tale fenomeno. In particolare, ci si propone di migliorare la capacità di rilevamento e analisi dell'esposizione dei cittadini alla disinformazione, la cooperazione e la capacità di elaborare risposte

[43] European Commission, Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Action Plan against Disinformation JOIN(2018), consultabile al sito: <https://ec.europa.eu/digital-single-market/en/news/action-plan-against-disinformation>.

comuni a tali minacce in collaborazione con le piattaforme online e l'industria, e la resistenza della società mediante campagne e strumenti di sensibilizzazione.

In particolare, la Commissione ha proposto numerose azioni tra cui il sostegno al giornalismo di qualità, mediante aiuti di stato da parte degli Stati membri, e il coinvolgimento attivo delle piattaforme online e dell'industria pubblicitaria attraverso l'elaborazione di un codice di condotta[44], per aumentare la trasparenza e proteggere i cittadini dalla disinformazione. Tale codice, a propria volta, identifica una serie di azioni che i firmatari devono mettere in atto, tra cui un sistema di valutazione delle fonti e della qualità dei contenuti. Il codice richiede inoltre l'uso di strumenti tecnologici per l'identificazione e la chiusura di profili falsi, per facilitare la diffusione di informazioni pertinenti, autentiche, accurate e autorevoli, per ottenere una maggiore trasparenza. In particole tali strumenti dovranno consentire agli utenti di comprendere perché sono stati presi di mira da una determinata pubblicità politica, anche mediante l'uso di indicatori di affidabilità delle fonti, di proprietà dei media e /o di verifica dell'identità del promotore dell'informazione.

Un'altra interessante iniziativa proviene dalla California, che ha adottato il Senate Bill n. 1001[45], che vieta l'uso dei bot allo scopo di ingannare e fuorviare gli utenti di piattaforme e siti internet. In particolare, la sezione 17941(a) rende illegale per chiunque l'uso di bot per comunicare o interagire online con altri individui in California con l'intento di indurre in errore tali individui circa la propria identità artificiale.

Quelle appena ricordate sono solo le prime timide iniziative per la regolazione dell'intelligenza, ma mostrano che governare l'IA, disciplinandone l'uso, è alla nostra portata. La coscienza dei problemi deve operare quale stimolo per soluzioni socialmente e tecnologicamente adeguate alla misura dei problemi stessi.

[44] Representatives of online platforms, leading social networks, advertisers and advertising industry agreed on a self-regulatory Code of Practice to address the spread of online disinformation and fake news, consultabile al sito: <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation>. Su cui in questo numero speciale: M. MONTI, *La disinformazione online, la crisi del rapporto pubblico-esperti e il rischio della privatizzazione della censura nelle azioni dell'Unione Europea (Code of practice on disinformation),* in *questo fascicolo.*

[45] SB-1001 Bots: disclosure, consultabile al sito: <https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001>.