L'IRCOCERVO
PRIMA RIVISTA ELETTRONICA ITALIANA DI METODOLOGIA GIURIDICA - TEORIA GENERALE DEL DIRITTO E DOTTRINA DELLO STATO
FIRST ITALIAN DIGITAL JOURNAL OF LEGAL METHODOLOGY - GENERAL THEORY OF LAW AND DOCTRINE OF THE STATE

La personalizzazione algoritmica e il suo apparente antropocentrismo come problema filosofico ed etico

Emanuele Brambilla

Università degli Studi di Milano

Abstract: **Algorithmic Personalization and its Apparent Anthropocentrism as a Philosophical and Ethical Problem**
The article deals with the definition and the scope of the so called "algorithmic governmentality". Specifically, it examines how the core of this system lies in the discovery of correlations among data, which are interpreted in order to influence users' lives. Furthermore, it reads the problem of the filter bubble in light of this governmentality, highlighting how the level of content personalization is falsely anthropocentric. In conclusion, it shows the necessity of a strong "algorethics".

Keywords: Algorithms, Neutrality, Filter Bubble, Governmentality, Algorethics.

Sommario: 1. Qualche considerazione preliminare – 2. Gli algoritmi fra neutralità e governamentalità – 2.1. Il significato di neutralità algoritmica – 2.2. Il concetto di governamentalità algoritmica in prospettiva antropologica – 2.3. Una specificazione: la *filter bubble* – 3. La problematica algoritmica: considerazioni tecniche ed etiche – 3.1. Metodi tecnici di intervento sugli algoritmi – 3.2. L'algoretica e l'umanizzazione degli algoritmi.

1. Qualche considerazione preliminare

La *governance* della realtà virtuale è un compito complesso, che richiede sforzi diversificati e su piani molteplici, sebbene coerenti e unitari fra loro. Come ricorda Gambini:

> La grande quantità di dati personali generati dall'odierno mondo elettronico, combinata con le nuove tecnologie a disposizione, che rendono sempre più sofisticati e meno facili da rilevare gli strumenti di raccolta e trattamento delle informazioni personali, aumentano in modo esponenziale i rischi di

esposizione delle persone a forme illegittime di intrusione nella propria sfera privata e di spoglio della propria identità[1].

In un contesto così articolato, gli algoritmi assumono un ruolo centrale, diventando il motore invisibile e fondamentale di ciò che accade nell'"infosfera[2]" e in tutti i sistemi o *software* di intelligenza artificiale. A questa complessità, contribuisce il fatto che gli algoritmi stessi siano multiformi e sfaccettati, poiché molteplici sono le modalità in cui possono essere costruiti e progettati[3].

Inoltre, il modo in cui essi funzionano determina altresì lo scopo e l'ambiente virtuale in cui andranno a operare. Alcune fra le funzioni più comuni sono il riordino e la rielaborazione dei dati, la ricerca di informazioni, la protezione delle comunicazioni tramite la crittografia, la previsione di eventi attraverso l'analisi di casi simili e il riconoscimento di determinati *patterns.* Un salto di qualità in questo ambito è avvenuto con l'introduzione di modelli basati sull'apprendimento automatico che hanno trasformato profondamente l'ideazione degli algoritmi, ottimizzandone lo sviluppo e rendendoli in grado di individuare correlazioni, formulare previsioni e adattare le proprie prestazioni sulla base dei dati a loro disposizione[4].

In ogni caso, a prescindere dalla loro specificità e dalla finalità con la quale vengono implementati, è indiscutibile che la loro presenza ed efficienza siano alimentate da una pluralità di dati, non sempre prodotti o inferiti dall'attività dell'utente. Accanto a questi ultimi, infatti, figurano i dati sintetici, i quali pur non appartenendo ad una determinata persona, possono essere costruiti attraverso modelli o proprietà statistiche derivate da dati personali[5]. Ciò rende manifesto come

[1] M. Gambini, "La protezione dei dati personali come diritto fondamentale della persona: meccanismi di tutela", in *Espaço Jurìdico Journal of Law*, 14 (2013), n. 1, pp. 151-152.

[2] Il termine compare in diversi autori per definire quell'ambiente costituito da scambi informativi e linguistici di vario genere. Nello specifico, per Floridi, l'infosfera è formata sia dall'intelligenza artificiale sia dall'uomo, che diviene un agente informazionale in grado di scambiare dati con altri enti, sia naturali sia artificiali. Per approfondire: K.E. Boulding, *Economics as a Science*, McGraw-Hill, Bombay-New Delhi, 1970, pp. 15-16; A. Toffler, *The Third Wave*, William Morrow and Company, New York, 1980, p. 170; L. Floridi, "Information Ethics: On the Theoretical Foundations of Computer Ethics", in *Ethics and Information Technology*, 1 (1999), n. 1, p. 44.

[3] Per approfondire: S.S. Skiena, *The Algorithm Design Manual*, Springer, Cham, 2020.

[4] Secondo Nello Cristianini, l'introduzione del *machine learning* ha aperto una nuova stagione per l'intelligenza artificiale e per gli algoritmi. "*The shortcuts that enabled AI to rapidly flourish in the past 20 years are made possible by machine learning algorithms and large samples of human behaviour. Once trained on superhuman quantities of data, algorithms such as GPT-3 can complete sentences, predict purchases, answer questions, and assist with translation, all without an understanding of the subject matter*". N. Cristianini, *The Shortcut. Why Intelligent Machines Do Not Think Like Us*, CRC Press, Boca Raton-Londra-New York, p. 64; si veda anche: R. Kitchin, "Thinking Critically about and Researching Algorithms", in *Information, Communication & Society*, 20 (2016), n. 1, pp. 14-29.

[5] Per una ricostruzione del concetto di dati sintetici, intesi come dati generati da modelli matematici con lo scopo di imitare dati reali per finalità statistiche o di addestramento di modelli di *machine*

la società dell'informazione si configuri sempre più come una vera e propria "società algoritmica",

> organizzata sulla base di processi decisionali delegati a software che, in esecuzione di complessi algoritmi, elaborano o comunque adoperano proprio l'informazione non solo a fini decisionali, bensì anche solo per apprendere […], dal momento che l'addestramento dei sistemi intelligenti richiede proprio la fornitura di dati per consentire loro di imparare costantemente e poi continuare a farlo grazie all'esperienza che progressivamente acquisiscono[6].

La persona, dunque, si trova al centro di questo sistema proprio perché contribuisce alla sua stessa costruzione e prosecuzione, spesso anche a prescindere dalla sua volontà. La "fiducia" negli algoritmi, o per lo meno l'accettazione del loro influsso, ne determina un utilizzo sempre più esteso, il quale trasforma a sua volta il modo di vivere nel digitale e nelle varie dimensioni della società[7]. Tale fiducia, in particolare, diviene cruciale quando la complessità dei sistemi ostacola la loro piena comprensione da parte dell'utente, il quale deve decidere se affidarvisi. Una scelta del genere, tuttavia, lungi dall'essere irriflessa, dovrebbe raggiungere un proprio equilibrio valutando esclusivamente le capacità effettive del sistema e non quelle che eccedono il suo perimetro d'azione[8].

L'affidamento, pertanto, non è qualcosa di statico e lineare, ma si evolve nel tempo in base all'interazione concreta con il sistema, che rende utenti e istituzioni consapevoli del suo funzionamento e delle sue criticità[9].

Nei prossimi paragrafi, ci si focalizzerà sul ruolo che gli algoritmi hanno nella vita degli utenti e a quali conseguenze porti il loro uso in ambiti anche molto diversi fra loro, suggerendo un orizzonte etico nel quale inscrivere questa problematica e rinvenire possibili soluzioni.

learning, sia concesso il rimando a: E. Brambilla, "Dati sintetici e 'corpo elettronico': una possibile contraddizione?", in *Ciberspazio e Diritto*, 26 (2025), n. 81, pp. 347-356.

[6] G. Fioriglio, "Intelligenza artificiale, privacy e rapporto di lavoro: una prospettiva informatico-giuridica", in *Lavoro, Diritti, Europa*, (2022), n. 3, p. 3; si veda anche: G. Fioriglio, "La Società algoritmica fra opacità e spiegabilità: profili informatico-giuridici", in *Ars interpretandi*, 26 (2021), n. 1, pp. 53-67.

[7] "Questo confidare negli algoritmi ne determina una presenza sempre più pervasiva, che sembra non conoscere confini, giustificando il parlare di una società che essi contribuiscono a definire nelle sue nuove e significative caratteristiche". S. Rodotà, *Il diritto di avere diritti*, Laterza, Roma-Bari, 2013, p. 402; invece, per un concetto di fiducia in senso normativo-istituzionale si consulti: High-Level Expert Group on Artificial Intelligence, *Orientamenti etici per un'IA affidabile*, 8 aprile 2019.

[8] "*Calibration refers to the correspondence between a person's trust in the automation and the automation's capabilities* […] *Overtrust is poor calibration in which trust exceeds system capabilities*". J.D. Lee, K.A. See, "Trust in Automation: Designing for Appropriate Reliance", in *Human Factors*, 46 (2004), n. 1, 2004, p. 55 (si vedano complessivamente le pp. 50-56).

[9] F. Cabiddu, L. Moi, G. Patriotta, D.G. Allen, "Why Do Users Trust Algorithms? A Review and Conceptualization of Initial Trust and Trust Over Time", in *European Management Journal*, 40 (2022), n. 5, pp. 685-706.

2. Gli algoritmi fra neutralità e governamentalità

L'accademica australiana Kate Crawford, circa il ruolo dell'informazione e degli algoritmi nella società scrive che:

> Il modo in cui i dati vengono interpretati, raccolti, classificati e denominati è fondamentalmente un atto di creazione e perimetrazione del mondo […] Il mito della raccolta dei dati come pratica benevola nell'informatica ha oscurato le sue operazioni di potere, proteggendo coloro che ne traggono il maggior profitto ed evitando loro la responsabilità delle sue conseguenze[10].

Con queste parole, la studiosa sottolinea come la raccolta e la rielaborazione dei dati tramite gli algoritmi, non debba lasciare indifferente l'uomo, proprio perché tale potere occulto intacca la sua stessa vita. In questo capitolo, partendo da questa considerazione, si cercherà, anzitutto, di comprendere se e come un algoritmo possa davvero dirsi neutrale, per poi analizzare quali siano le conseguenze del suo influsso sulla vita dell'utente.

2.1. Il significato di neutralità algoritmica

Le varie e sofisticate modalità di costruzione degli algoritmi dimostrano che non solo la loro diffusione sia trasversale in tutti gli ambiti inerenti all'intelligenza artificiale, ma manifestano anche il ruolo di primo piano che hanno nel guidarne il funzionamento e l'uso.

La sociologa Zeynep Tufekci ritiene che:

> *Algorithmic control of content can mean the difference between widespread visibility and burial of content. […] an algorithm can be a strong tailwind or a substantial obstacle. Algorithms can also shape social movement tactics as a movement's content producers adapt or transform their messages to be more algorithm friendly*[11].

[10] K. Crawford, *Né intelligente né artificiale. Il lato oscuro dell'IA*, il Mulino, Bologna, 2024, p. 135.

[11] Sebbene il contesto in cui Tufekci collochi il suo ragionamento sia quello dei movimenti sociali, tale riflessione è applicabile analogicamente anche all'ambito della presente discussione: il controllo algoritmico della visibilità dei contenuti, infatti, incide in modo generale sulle condizioni di accessibilità e circolazione delle informazioni nell'ambiente digitale. Tale meccanismo opera ogni volta che la visibilità di un contenuto dipende dalle modalità con cui una piattaforma lo raccomanda, lo filtra e lo rende accessibile agli utenti. Z. Tufekci, *Twitter and Tear Gas. The Power and Fragility of Networked Protest*, Yale University Press, New Haven-Londra, 2017, p. 154.

Come si nota, Tufekci sottolinea l'enorme portata personale e collettiva che riveste anche una semplice modifica ai *training set* degli algoritmi. Infatti, ciò può portare alla manipolazione di disposizioni e attitudini, a modificare la percezione della realtà, a indurre certe decisioni e alimentare desideri o disillusioni.

Se trasformazioni del genere, seppur minime, possono mutare di molto il risultato degli algoritmi, allora questi ultimi possono essere definiti intrinsecamente neutrali?

Secondo alcuni studiosi, un algoritmo è davvero neutrale solo quando valori esterni alla sua finalità principale non incidono sul modo in cui produce i propri output[12]. Ad esempio, si consideri

> *a hiring algorithm that aims to rank candidates on the basis of qualification. The algorithm is neutral only if values other than qualification play no role in how the algorithm ranks candidates. So, it's not neutral if an applicant's politics, race, or family connections play any role in how it ranks candidates. In such cases, values external to qualification play a role in its rankings*[13].

Secondo questa definizione, la neutralità algoritmica non è l'assenza di ogni valore in generale, bensì l'assenza di valori esterni allo scopo dell'algoritmo stesso, che ne influenzino i risultati. Altri autori ritengono che gli algoritmi possano essere definiti neutrali solo se interpretati come strumenti che svolgono una funzione, al pari di lavatrici e frigoriferi, che sono neutrali e "indifferenti" rispetto al bucato o al cibo[14]. Tuttavia ci si chiede come tale neutralità sia possibile, in quanto da un lato, l'algoritmo, considerato singolarmente, non essendo sufficiente a determinare l'output, rende necessario l'apporto di considerazioni "esterne"[15]; dall'altro, non è neutrale rispetto a ciò che produce, proprio perché (a differenza della lavatrice) si alimenta dello stesso "oggetto" (ossia dei dati) su cui esercita la sua funzione tecnica.

Stando all'esempio esposto pocanzi, se il sistema deve classificare i candidati basandosi sulle loro qualifiche, dovrà valutare i titoli di studio, l'esperienza professionale pregressa, le competenze acquisite e tutta una serie di fattori che

[12] "[…] *an algorithmic system is neutral only if values external to its aim play no role in how it delivers its results*". M. Philips-Brown, "Algorithmic neutrality", in *ArXiv*, p. 17 (https://arxiv.org/pdf/2303.05103).

[13] *Ibidem*.

[14] "Se consideriamo che gli algoritmi altro non sono che tecnologie che svolgono una funzione, la risposta è ovvia: certo che sono neutrali. Una lavatrice è neutrale rispetto al bucato, un frigorifero è neutrale rispetto al cibo. Il problema è che gli algoritmi […] non sono né lavatrici, né frigoriferi: sono tecnologie che agiscono in maniera autonoma e spesso questa autonomia si esprime in un comportamento pregiudizievole". D. Bennato (2017), "Il pregiudizio degli algoritmi, ovvero: come prevenire la dittatura dei software?", in *Agenda digitale*. Recuperato da https://www.agendadigitale.eu/infrastrutture/il-pregiudizio-degli-algoritmi-ovvero-come-prevenire-la-dittatura-dei-software/, [Data di consultazione: 23/02/2026].

[15] "*Because underdetermination is pervasive, neutrality is impossible for many – if not most – algorithmic systems that affect our daily lives*". M. Philips-Brown, *op. cit.*, p. 19.

contribuiscono alla scelta della persona più adatta. Il problema è che siccome questi elementi non bastano per ordinare i candidati in una classifica, bisogna decidere quale peso dare a ciascuno di essi, intervenendo "dall'esterno". Detto altrimenti, è indispensabile stabilire se valgono più i titoli di studio o le esperienze pregresse e classificare ciascuno di essi secondo criteri che ne fissino la pertinenza rispetto alla posizione lavorativa.

Questa valutazione costituisce il momento epistemico, ossia quello in cui i dati disponibili vengono utilizzati per sostenere una certa scelta circa l'idoneità dei candidati. L'evidenza tratta dai dati, tuttavia, non è mai oggettiva, ma può risultare inconcludente, imperscrutabile o fuorviante.

Nel primo caso, l'algoritmo produce una conoscenza incerta, basata su inferenze statistiche e sulla presunzione che ogni correlazione fra dati sia di tipo causale, benché non sia così. Ciò può accadere quando il sistema, basandosi sull'analisi di candidati che in passato hanno ottenuto un punteggio alto grazie ai titoli di studio, conclude che i candidati attuali con caratteristiche simili siano più idonei di altri. In questo caso, l'algoritmo trasforma una semplice correlazione statistica in una certezza causale ingiustificata o, per lo meno, non sufficientemente argomentata.

Nel secondo caso, le ragioni che motivano la conclusione raggiunta dall'algoritmo, non sono intellegibili o verificabili, come quando il punteggio assegnato ad un candidato non è accompagnato dall'esplicazione dei passaggi logici che lo hanno prodotto.

Nel terzo caso, l'affidabilità (e "neutralità") dei dati di partenza vengono riprodotte negli *output*, influenzando la classifica dei candidati. Se l'algoritmo in questione premia con un punteggio alto la continuità lavorativa dei candidati senza considerare alcune situazioni particolari, come interruzioni per maternità o per motivi di salute, penalizza tutti quelli che, pur idonei, non hanno le esatte caratteristiche che rispecchiano il set di partenza[16].

Al momento epistemico segue quello normativo, nel quale gli output prodotti vengono utilizzati per orientare e giustificare le decisioni che si assumono in merito ai candidati, selezionandone alcuni e scartandone altri. Una decisione fondata sulle tre criticità appena esaminate, corre il rischio di risultare rispettivamente ingiustificata, opaca o distorta, causando problematiche concrete alla vita e alla carriera dei candidati[17].

Di conseguenza, la neutralità algoritmica, intesa come assenza di valori, capacità orientative o *bias*, è un falso ideale e uno dei tanti "miti" legati alla

[16] Il rapporto fra *input* e *output* negli algoritmi corrisponde al principio "*garbage in garbage out* ". "*Evaluations of the neutrality of the process, and by connection whether the evidence produced is misguided, are of course observer-dependent*". B.D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, "The Ethics of Algorithms: Mapping the Debate", in *Big Data & Society*, 3 (2016), n. 2, p. 5

[17] Per approfondire come ciascuna delle tre criticità epistemiche abbia precise conseguenze: *Ivi*, p. 5-9.

tecnologia e all'intelligenza artificiale. Tale mito, si fonda sull'idea che la (apparente) rimozione dell'intervento umano e l'utilizzo di un sistema tecnico-calcolante possano garantire totale imparzialità e correttezza. Sul punto, alcuni autori, dimostrano l'impossibilità della neutralità attraverso tre argomentazioni principali. La prima mostra come i dati, già in fase di addestramento, non apprendano la realtà in modo oggettivo e indipendente da criteri umani o distorsioni. La seconda scardina la retorica della "disintermediazione", secondo la quale l'automatizzazione di un processo garantirebbe l'imparzialità in quanto eliminerebbe l'intervento umano. La terza dimostra come sia falsa la pretesa che i risultati del calcolo algoritmico siano oggettivi e impediscano l'alterazione della realtà secondo logiche predeterminate[18].

Ciò, tuttavia, non significa che tutti gli algoritmi con scopi simili a quelli descritti da questo esempio siano di per sé ingiusti, poiché assumono una grande rilevanza il peso che viene dato alle diverse variabili e il modo in cui vengono valutate.

D'altro canto, è imprescindibile una valutazione etica degli algoritmi, sebbene vi siano moltissime sfumature in questo campo. Si pensi, per esempio, agli algoritmi che inducono gli utenti a comprare un determinato prodotto mostrando inserzioni pubblicitarie, create grazie alla cronologia. In questo caso, una valutazione etica adeguata dovrebbe considerare le finalità dell'algoritmo, i mezzi che utilizza per raggiungere il proprio scopo e se sia moralmente riprovevole o meno persuadere subdolamente qualcuno a compiere una determinata azione.

Gli algoritmi vengono ideati, dunque, con un obbiettivo preciso e per ottenere risultati ben definiti; essi hanno un vero e proprio "potere" sull'utente, che molto spesso rimane implicito e difficilmente rinvenibile, come si vedrà a breve. Secondo alcuni, questo potere, non è necessariamente dannoso, ma va considerato alla luce del fatto che gli algoritmi hanno *bias* così come l'essere umano. Piuttosto, la questione principale è quella di comprendere le modalità in cui gli algoritmi esercitano il loro potere sulla persona per poterne chiarire i meriti e i limiti[19]. Tuttavia, ci si domanda quali siano le conseguenze di un potere così esteso, che supera le caratteristiche e gli effetti del singolo algoritmo, per inserirsi in un eco-

[18] Per leggere gli esempi per ciascuna delle tre categorie: M. Airoldi, D. Gambetta, "Sul mito della neutralità algoritmica", in *The Lab's Quarterly*, 20 (2018), n. 4, pp. 33-38; si veda anche: K. Coyle, "La neutralità difficile", in *AIB Studi*, 58 (2018), n. 2, pp. 257-265.

[19] "[…] *algorithmic power isn't necessarily detrimental to people; it can also act as a positive force.* […] *algorithms* […] *operate with biases like the rest of us. And they can make mistakes. What we generally lack as a public is clarity about how algorithms exercise their power over us. With that clarity comes an increased ability to publicly debate and dialogue the merits of any particular algorithmic power.* […] *algorithmic codes are more opaque, hidden behind layers of technical complexity. How can we characterize the power that various algorithms may exert on us? And how can we better understand when algorithms might be wronging us?*". N. Diakopoulos, "Algorithmic Accountability: On the Investigation of Black Boxes", in *Columbia Journalism Review* (https://www.cjr.org/tow_center_reports/algorithmic_accountability_on_the_investigation_of_black_boxes.php).

sistema digitale molto più ampio e complesso. Infatti, il tema non è soltanto quello della valutazione di singoli algoritmi, ma soprattutto quello dell'intero sistema che creano e nel quale la persona umana si trova inserita[20].

2.2. Il concetto di governamentalità algoritmica in prospettiva antropologica

Secondo Berns e Rouvroy, per conferire agli algoritmi un potere di questo tipo sono necessarie due precondizioni, che devono essere raggiunte stabilmente prima di ottenere una "governamentalità"[21] algoritmica completa ed efficace.

Il primo obbiettivo consiste nella raccolta massiva e possibilmente automatizzata (così da velocizzare la procedura) di enormi quantità di dati che in seguito vengono custoditi in magazzini digitali, secondo scopi e tempistiche non sempre chiari. Queste informazioni, raccolte talvolta senza il consenso e la piena consapevolezza dell'utente[22], provengono da fonti differenti e vengono utilizzate per scopi altrettanto eterogenei. Secondo i due studiosi, ciò che risulta veramente incredibile, è che i dati più che essere volutamente ceduti, vengono quasi "abbandonati"[23] con disinteresse o disillusione, come se non si potesse fare altrimenti. A causa di questo comportamento, l'utente viene spersonalizzato e i suoi dati divengono tessere di un mosaico impossibile da completare, poiché pur potendo essere correlati reciprocamente, non restituiscono mai la realtà integrale

[20] B.C. Stahl, "From Computer Ethics and the Ethics of AI Towards an Ethics of Digital Ecosystems", in *AI and Ethics*, 2 (2021), pp. 65-77.

[21] Termine di foucaultiana memoria. M. Foucault, *Sicurezza, territorio, popolazione*, Feltrinelli, Milano, 2005.

[22] Vi sono due strategie di raccolta: la prima incentiva "[…] la fruizione di beni e servizi che, nell'essere utilizzati, in virtù della tecnologia adottata, registrano informazioni. L'altra strategia è invece quella di chiedere informazioni direttamente agli interessati. Le tipologie di dati raccolti sono diverse nei due casi: se ci si rivolge con una richiesta all'interessato, le informazioni sono filtrate da una sua valutazione e verbalizzazione; nell'altro caso, le informazioni raccolte riguardano i suoi comportamenti. La collaborazione richiesta per consentire la raccolta dei dati è dunque diversa nei due casi: mentre nel primo l'interessato deve collaborare attivamente comunicando egli stesso le informazioni, nel secondo la collaborazione è, per così dire, passiva, nel senso che è sufficiente che l'interessato usi un bene o servizio affinché i dati generati da tale comportamento siano suscettibili di registrazione". S. Thobani, "Il pagamento mediante dati personali", in S. Orlando, G. Capaldo (a cura di), *Annuario 2021. Osservatorio Giuridico sulla Innovazione Digitale*, Sapienza Università Editrice, Roma, 2021, pp. 361-379.

[23] I dati "[…] *sont plus abandonnées que cédées, des traces laissées et non des données transmises, mais sans apparaître pour autant comme 'volées', qu'elles apparaissent aussi comme absolument quelconques et dispersées, tout cela donne lieu à une évacuation ou au minimum à un voilement de toute finalité, et à une minorisation de l'implication du sujet, et donc du consentement pouvant être donné à cette communication d'informations*". A. Rouvroy, T. Berns, "Gouvernementalité algorithmique et perspectives d'émancipation. Le disparate comme condition d'individuation par la relation? ", in *Réseaux*, 1 (2013), n. 177, p. 169.

della persona, ma solamente una ricostruzione frammentata e lontana dal vissuto effettivo[24].

Tuttavia, paradossalmente, tale frammentazione della persona fa apparire i singoli dati come oggettivamente veri, poiché assolutizzati e astratti dal loro contesto. Si pensi a come alcuni *software* biometrici riescano a riconoscere le emozioni attraverso espressioni facciali e para-verbali: ogni singolo movimento o gesto viene trasformato in pura informazione, in base alla quale prendere posizione, attuare strategie di mercato e additare qualcuno come sospetto. Rouvroy e Berns mostrano come una dinamica di questo genere marchi l'ingresso nel cosiddetto "*comportementalisme numérique généralisé*", in cui la persona è ridotta a dati "grezzi". La "governamentalità algoritmica", pertanto, non prende avvio da un potere diretto sull'utente, ma dalla sua esclusione e datificazione[25].

Il secondo passaggio riguarda il trattamento dei dati attraverso l'intelligenza artificiale, con lo scopo di conferire loro un senso e farne emergere le possibili connessioni. In tal modo, il dato diviene l'elemento fondamentale e originario per poter istituire correlazioni e svolgere deduzioni che contribuiscano alla scoperta di ulteriori informazioni che non figurano nel *set* posseduto inizialmente. Questa operazione, oltre a porre tutti i singoli dati sullo stesso livello veritativo, mira a realizzare una conoscenza fondata sulle connessioni che sussistono fra le varie informazioni, la quale corre il rischio di essere ambigua e dannosa poiché strettamente intrecciata alla vita delle persone. L'apparente limpidità degli *output*, perciò, non deve essere considerata una prova della loro attendibilità e correttezza. Infatti, dalla correlazione di due o più elementi non si può dedurre a priori che fra loro sussista un rapporto causale e nemmeno che tale connessione rispecchi la realtà concreta, come si è accennato in precedenza in merito alle criticità degli usi epistemici.

È necessario, pertanto, fare propria la consapevolezza che ciò che emerge dall'analisi dei dati, non sia sufficiente per prendere una qualsiasi decisione, specialmente se di carattere giuridico o politico, e sviluppare un senso critico nei confronti di un sapere che si basa su profilazioni automatizzate[26].

[24] Le tecniche di profilazione "[…] determinano il rischio che l'io venga frammentato, a sua insaputa, in una molteplicità di banche dei dati, offrendo così una raffigurazione parziale e potenzialmente pregiudizievole della persona, la quale verrebbe così ridotta alla mera sommatoria delle sue proiezioni elettroniche". G. Resta, *Dignità, persone, mercati*, Giappichelli, Torino, 2014, p. 333.

[25] I dati sono costitutivi "[…] *d'un comportementalisme numérique généralisé dès lors qu'elles expriment ni plus ni moins que les multiples facettes du réel, le dédoublant dans sa totalité, mais de manière parfaitement segmentée, sans faire sens collectivement* […] *l'évolution même des capacités technologiques renforce cette sorte d'objectivité de la donnée qui échappe à toute subjectivité* […] *de telles données ont comme principale caractéristique d'être parfaitement anodines, de pouvoir rester anonymes, et d'être non contrôlables.* […] *elles semblent ne pas mentir, c'est-à-dire pouvoir être considérées comme parfaitement objectives! Cette inoffensivité et cette objectivité sont l'une et l'autre dues à une sorte d'évitement de la subjectivité*". A. Rouvroy, T. Berns, *op. cit.*, pp. 169-170.

[26] Si è di fronte ad una produzione di saperi statistici "[…] *constitués de simples corrélations à partir d'informations non triées, et donc parfaitement hétérogènes*" la quale "[…] *ne réclamant qu'un*

Come si diceva, il raggiungimento di questi due stadi propedeutici, sfocia nell'ottenimento della governamentalità algoritmica, in cui gli algoritmi esercitano un'importante influenza sulla vita degli utenti, anche in aspetti che possono passare inosservati.

A differenza del passaggio precedente, dove il fulcro decisionale risiedeva nelle correlazioni dei dati, in questo caso viene a costituirsi una vera e propria "normatività" occulta, capace di condizionare sottotraccia tutti gli utenti. I tratti personali e caratteriali, le preferenze, le opinioni e tutto quanto è registrabile e analizzabile grazie alle suddette capacità di correlazione, viene utilizzato per anticipare comportamenti e desideri oppure per effettuare proposte mirate e personalizzate. Il successo di questa normatività trasversale, ossia presente in tutto l'ambiente digitale, si basa su due fattori, sinergetici fra loro.

Il primo è il rapporto direttamente proporzionale che va istituito fra quantità e qualità dei dati: più i dati sono attinenti alla realtà, più si assottiglia la discrepanza fra quest'ultima e il profilo dell'utente organizzato algoritmicamente.

Il secondo è il nesso fra la predittività comportamentale e la reattività dell'ambiente in cui l'individuo è inserito. A causa della diffusione dei dispositivi intelligenti, infatti, l'ambiente stesso viene reso capace di registrare ogni aspetto della realtà e di adattarsi alla mutevolezza di ciò che accade. In tal modo, viene rimodellato lo spazio stesso in cui le persone vivono, rendendo i loro comportamenti sempre più conformi alle impostazioni algoritmiche, senza che sia necessaria un'azione diretta su di loro.

Questo "controllo", implicitamente così stringente, indossa tuttavia la maschera della cura verso la specificità di ogni persona, generando l'impressione che alcune categorie discriminanti vengano tralasciate a favore di una maggiore attenzione per le preferenze di ciascuno. Così facendo, non si esclude *ex post* ciò che non rispetta certi canoni, ma si previene fin dal principio ogni possibile novità o imprevedibilità che non rientri nella logica della correlazione[27]. Dunque, più che

minimum d'intervention humaine […] *nous nous trouvons à nouveau face à l'idée d'un savoir dont l'objectivité pourrait paraître absolue* […]. *Les normes semblent émerger directement du réel lui-même.* […] *ces 'savoirs' ne sont cependant constitués 'que' de corrélations, ce qui n'est pas en soi un problème, si l'on n'oublie pas* […] *de conserver un doute, d'entretenir une méfiance par rapport à la suffisance des corrélations, de maintenir la distinction entre corrélation et cause* […] *d'éviter que des décisions produisant des effets juridiques à l'égard de personnes ou les affectant de manière significative ne soient prises sur le seul fondement d'un traitement de données automatisé* […]. *Il semble important de rappeler ceci face à l'évolution vers un monde qui paraît de plus en plus fonctionner comme s'il était constitué lui-même de corrélations, comme si celles-ci étaient ce qu'il suffit d'établir pour en assurer le bon fonctionnement*". *Ivi*, pp. 170-171.

[27] Vi è una differenza fra i dati personali e il sapere prodotto dalla profilazione: l'azione anticipatoria sui comportamenti si limita a intervenire "[…] *sur leur environnement, a fortiori dès lors que l'environnement lui-même est réactif et intelligent, c'est-à-dire qu'il récolte lui-même des données* […]. *De la sorte à nouveau, on évite toute forme de contrainte directe sur l'individu pour préférer rendre, au niveau même de son environnement, sa désobéissance ou certaines formes de marginalité toujours plus improbables.* […] *le profil 'lié' au comportement d'un individu pourrait lui être adapté de façon parfaitement efficace, de par la démultiplication des corrélations employées, au point de*

profili corrispondenti al vero, l'obbiettivo è creare profili coerenti e dettagliati, che rispondano agli scopi per cui vengono realizzati.

Il risultato di tutto ciò è un apparente e curioso "antropocentrismo", dove ogni servizio, consiglio e previsione sembrano cuciti su misura per rispettare e far fiorire l'unicità di ciascuna persona. Si consideri, per esempio, a come Netflix e Spotify, grazie ad algoritmi addestrati tramite il *machine learning*, riescano ad esaminare moltissimi dati per consigliare all'utente film, serie tv, podcast o canzoni adatte ai suoi interessi. Tramite l'incrocio dei dati, quasi sempre impliciti e di natura caratteriale o comportamentale, è possibile riconoscere una serie di *patterns* così da offrire un'esperienza unica e altamente personalizzata[28]. Tuttavia, queste modalità di profilazione, elaborano dati privi di senso per l'individuo e, come si anticipava, li assolutizzano per prevedere scelte o addirittura per indirizzare alcune preferenze individuali. Dunque, lungi dal porre al centro il soggetto e le sue peculiarità, si assottigliano le sue capacità decisionali attraverso la creazione automatica di un profilo personale, costantemente aggiornato e migliorato in base ai *click*.

In tal modo, si viene a creare una sorta di soggettività guidata dai dati, in cui la persona, ancor prima che possa esprimersi, viene inquadrata da processi algoritmici, che delimitano le sue capacità decisionali a ciò che le viene proposto.

Tuttavia, tale analisi, non può essere riferita indistintamente a ogni ambito in cui operano gli algoritmi, ma riguarda soprattutto quei contesti, fondati sulla profilazione comportamentale e la deduzione delle preferenze, in cui la personalizzazione restringe il campo delle opzioni visibili all'utente per ottenere gli scopi appena esaminati. In altri ambiti, come quello medico o diagnostico, tale adattamento alle caratteristiche soggettive è fondamentale per la personalizzazione delle cure da somministrare al paziente o per individuare gravi malattie in base a fattori ricorrenti[29].

sembler éviter l'usage de toute catégorie discriminante, et de pouvoir même prendre en compte ce qu'il y a de plus particulier dans chaque individu […] *la cécité des algorithmes relativement aux catégorisations* […] *est d'ailleurs l'argument récurrent que brandissent ceux qui sont favorables à leur déploiement en lieu et place de l'évaluation humaine* […]. *Le datamining et le profilage algorithmique, dans leur rapport en apparence non sélectif au monde, semblent prendre en considération l'entièreté de chaque réel jusque dans ses aspects les plus triviaux et insignifiants, mettant tout le monde à égalité* […]. *Il ne s'agit plus d'exclure ce qui sort de la moyenne, mais d'éviter l'imprévisible, de faire en sorte que chacun soit véritablement lui-même*". *Ivi*, pp. 171-173.

[28] Alcuni parametri considerati sono: la scelta iniziale dei contenuti, il percorso seguito per pervenire a tale scelta, il tempo passato davanti allo schermo, gli orari di accesso alla piattaforma, il genere del film o della canzone. C.A. Gomez-Uribe, N. Hunt, "The Netflix Recommender System: Algorithms, Business Value, and Innovation", in *ACM Transactions on Management Information Systems*, 6 (2015), n. 4, pp. 1-19.

[29] Per approfondire: N. Hallowell, S. Badger, A. Sauerbrei, C. Nellåker, A. Kerasidou, "'I Don't Think People Are Ready to Trust These Algorithms at Face Value': Trust and the Use of Machine Learning Algorithms in the Diagnosis of Rare Disease", in *BMC Medical Ethics*, 23 (2022), n. 112, pp. 1-14; K.B. Johnson, W.Q. Wei, D. Weeraratne, M.E. Frisse, K. Misulis, K. Rhee, J. Zhao, J.L. Snowdon, "Precision Medicine, AI, and the Future of Personalized Health Care", in *Clinical and*

2.3. Una specificazione: la *filter bubble*

In base a quanto esplicitato, risulta evidente che uno dei problemi principali in questo campo, dunque, non sia soltanto quello della salvaguardia della privacy, ma soprattutto il fatto che il profilo dell'utente parli al suo posto[30]. Ciò che tale profilo rivela, inoltre, si proietta sempre verso il futuro, ossia verso un calcolo delle probabilità di ciò che potrebbe accadere o che l'utente potrebbe fare, presentando come plausibile esclusivamente ciò che si adatta a determinate aspettative[31].

In questo contesto, si inserisce la dinamica della cosiddetta *filter bubble*, ovvero quell'effetto creato dagli algoritmi per cui nei motori di ricerca e nei social media vengono proposte o mostrate informazioni, immagini e video coerenti con gli interessi, i gusti e i comportamenti della persona.

La superficie e le coordinate della bolla vengono costruite in base alla correlatività fra i dati e sono in grado di escludere tutti i contenuti non pertinenti e dissonanti rispetto al profilo del fruitore. Secondo Eli Pariser, la *filter bubble* non va meramente ad inserirsi nella naturale tendenza a leggere e ad interessarsi a quegli aspetti che sono più congeniali con le proprie opinioni, ma apporta tre novità fondamentali. In primo luogo, isola quasi completamente l'utente da ciò che abita al di là delle pareti della bolla, non rendendolo visibile al suo interno. Inoltre, anche se un certo interesse è condiviso con altri utenti, questi ultimi non risiedono nella bolla tutti assieme, dato che il livello di personalizzazione e le differenze individuali possono portare, al massimo, a trovare punti di contatto, ma mai una completa identificazione fra bolle diverse.

Secondariamente, la bolla è invisibile non tanto perché non ci si accorga della sua presenza o influenza, ma perché non sono limpidi i criteri con cui filtra le informazioni che ritiene pertinenti agli interessi dell'individuo. Non sempre, infatti, nelle *news feed* di Google o nei *reels* di Facebook, vengono proposti argomenti o video che catturano l'attenzione.

In terzo luogo, non si decide di entrare o meno in una bolla, poiché tale ingresso è inevitabile, essendo insito nella natura stessa del *web*. Dunque, si sceglie

Translational Science, 14 (2021), n. 1, pp. 86-93; V. Francesco, *Algoritmi diagnostico-terapeutici in medicina interna*, Edra, Milano 2021.

[30] Le normative sulla privacy erigono "[…] *autour de l'individu, une série de 'barrières' aux tonalités essentiellement défensives et restrictives. Sans considérer ceci comme vain, nous voulons signaler ici avec force l'indifférence de ce 'gouvernement algorithmique' pour les individus, dès lors qu'il se contente de s'intéresser et de contrôler notre 'double statistique', c'est-à-dire des croisements de corrélations, produits de manière automatisée, et sur la base de quantités massives de données*". A. Rouvroy, T. Berns, *op. cit.*, pp. 179-180.

[31] O meglio, a quanto ci si aspetta in base alle conclusioni tratte dall'analisi per correlazione dei vari dati che compongono il profilo dell'utente.

inconsapevolmente in quale bolla entrare e, conseguentemente, quali escludere totalmente o parzialmente[32].

Se si legge la problematica delle bolle informative alla luce del concetto di *gouvernementalité algorithmique* delineato pocanzi, diviene chiaro come la bolla non sia cucita attorno alla persona, ma attorno al suo profilo, ossia a ciò che l'algoritmo ha deciso che quella persona è. Piuttosto, ciò che la persona reale percepisce, sono le conseguenze di tale filtro, come la selettività e superficialità delle conoscenze, la difficoltà di pensare fuori dagli schemi, il sentirsi confermata sempre nelle proprie opinioni senza saper argomentare a favore o contro.

A tal proposito, alcuni ricercatori, hanno condotto degli studi specifici sui cosiddetti *shorts* proposti da alcune piattaforme come TikTok o YouTube, provando, con l'aiuto dei *Large Language Models*, a simulare i meccanismi algoritmici che li selezionano. La novità di questo approccio risiede nel porre l'attenzione sulla nascita della bolla informativa come conseguenza dell'unione fra il sistema di raccomandazione e i *feedback* indicati dal fruitore.

L'obiettivo è quello di mostrare come la creazione della *filter bubble* sia connaturata a questi sistemi, benché, nonostante ciò, vi siano delle modalità per mitigare la restrizione progressiva dei contenuti. Anzitutto, i ricercatori costruiscono la loro simulazione creando due componenti e integrandole fra loro. La prima è la simulazione di un profilo utente con caratteristiche demografiche e attributi come età, sesso, interessi personali e disponibilità economiche (calcolate approssimativamente tramite il prezzo del cellulare). La seconda è l'ideazione di un modulo di raccomandazione, attraverso cui proporre agli utenti *shorts* personalizzati che rispondano alle loro caratteristiche individuali[33]. Ciò stabilito, all'utente vengono presentati cinque video ai quali deve reagire in uno dei seguenti modi: *watch*, *like*, *comment*, *collect*, *skip* e *dislike*. Ciascuna di queste opzioni ha un peso differente (sia negativo sia positivo), che va moltiplicato per altri fattori in una data funzione.

Per misurare la formazione e l'andamento statistico della *filter bubble*, vengono prese in considerazione tre metriche fondamentali, calcolate differentemente. La prima è il rapporto fra le tipologie di video guardati e il numero totale di categorie di ciascun video, ad un dato "livello", ossia a una categoria specifica con cui il video è classificato[34]. La seconda, detta "entropia", riguarda la

[32] E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You*, Penguin Books, Londra-New York, 2011, pp. 8-9.

[33] Gli utenti vengono dotati di ragioni per cui guardare un video e di una "personalità". Le motivazioni sono: "*Social Interaction, Entertainment, Information-Seeking, Browsing/Variety Seeking, Escapism*" La personalità consta di cinque dimensioni: "*Openness to Experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism* […]. *For the purposes of simulation, these dimensions are randomly generated on a scale from 0 to 1 for each user-agent*". N. Sukiennik, H. Wang, Z. Zeng, C. Gao, Y. Li, "SimTok: Simulating Filter Bubble on Short-video Recommender System with Large Language Model Agents", in *ArXiv*, p. 2 (https://arxiv.org/pdf/2504.08742).

[34] I tre livelli presi come esempio sono: sport, calcio, squadra che gioca. Le categorie sono corrispondono alle motivazioni, indicate nella nota precedente. *Ibidem*.

quantità informativa delle categorie di video visti, per ciascuno dei tre livelli[35]. La terza, chiamata "soddisfazione", misura il rapporto tra i contenuti apprezzati e quelli proposti, sempre a seconda del livello in cui sono localizzati.

I risultati mostrano che, nelle prime interazioni degli utenti, basate sulle motivazioni per cui si guardano gli *shorts*, l'entropia cala bruscamente, segno che vengono esposti a contenuti sempre più omogenei. A lungo andare, invece, si verifica una lieve aumento, sebbene la *filter bubble* resti particolarmente evidente al primo livello, dove la generalizzazione è maggiore[36]. Parallelamente, la soddisfazione dell'utente decresce insieme alla varietà dei contenuti, suggerendo che l'esposizione a video simili porti a noia e disinteresse.

Invece, l'entropia calcolata in base alla personalità degli utenti, decresce continuamente, raggiungendo un risultato molto basso. Ciò significa che, la personalità, rispetto alla motivazione per cui si guarda un video, è un fattore che contribuisce molto più sensibilmente alla formazione di bolle informative[37].

Per mitigare il filtro informativo gli autori propongono due strategie differenti. La prima fa leva sull'elevata corrispondenza iniziale fra interessi dichiarati dagli utenti e video mostrati, proponendo di aumentarne la varietà per stimolare la scoperta di nuovi interessi e la crescita della soddisfazione personale.

La seconda, suggerisce di attribuire ai suddetti *feedback* pesi diversi e scalari, dando più rilevanza ai commenti e ai video salvati, piuttosto che ai *like*, ottenendo così una varietà e una soddisfazione più elevate[38].

In definitiva, la centralità del soggetto e l'estrema personalizzazione dei contenuti, hanno portato ad un suo decentramento dal punto di vista della capacità di scelta e di opposizione, sebbene le tecniche appena presentate possano invertire tale rotta.

3. La problematicità algoritmica: considerazioni tecniche ed etiche

Questa apparente centralità dell'uomo fa sorgere importanti interrogativi etici sulla natura, lo sviluppo e l'uso degli algoritmi, tanto che ci si domanda in quali modi si possa favorire il bene comune e individuale della persona umana[39], oltre che conciliare l'antropocentrismo previsto dalle carte etiche e dalla normativa europea con una governamentalità che ha lo scopo esattamente opposto. Difatti, come si è

[35] Vedi nota precedente.
[36] Questo perché il primo livello, ossia quello dei video di sport in generale, presenta moltissime varianti e, di conseguenza, vi è una maggiore possibilità di restringimento dei contenuti.
[37] Anche le caratteristiche demografiche influiscono sulla severità della bolla. Sebbene siano più complicate da calcolare, si stima che agli utenti più giovani e con dispositivi più economici vengano presentati contenuti meno vari.
[38] Anche qui si nota come intervengano fattori esterni all'algoritmo, che ne impediscono la neutralità.
[39] Circa il significato di bene comune si legga: G. Samek Lodovici, *La socialità del bene*, ETS, Pisa, 2017, pp. 74-75 e 179-199.

visto, dinamiche come la *filter bubble* e fenomeni analoghi, non sono effetti collaterali del *web*, ma già intrinsecamente presenti nella sua struttura.

Cathy O'Neil, raggruppa i cosiddetti algoritmi "cattivi" in quattro livelli gerarchici che ne descrivono differenti tipologie e finalità.

Il livello meno grave, è rappresentato da quegli algoritmi che non riflettono intenzionalmente pregiudizi o false generalizzazione e che vengono, ad esempio, addestrati dalle ricerche degli utenti su Internet[40].

Il secondo livello si riferisce a quegli algoritmi che diventano problematici a causa della negligenza nel migliorarne i difetti, come accade nei sistemi di etichettatura delle fotografie, quando non sono stati addestrati con *dataset* molto ampi e variegati oppure in algoritmi che impostano automaticamente turni di lavoro, senza tenere conto delle caratteristiche famigliari e personali degli impiegati.

La terza tipologia, riguarda quegli algoritmi subdoli che creano pubblicità mirate a seconda della categoria sociale in cui rientrano gli utenti oppure quelli che calcolano gli interessi di un prestito e stabiliscono le tariffe assicurative[41].

Al quarto livello, infine, vi sono gli algoritmi che vengono volontariamente utilizzati per monitorare i cittadini e porre in essere una sorveglianza di massa[42].

È curioso notare come, la maggior parte degli algoritmi che rientrano in queste tipologie, specialmente nelle prime tre, sia perfettamente conforme alla legge. Infatti, come emerge anche da alcune carte etiche[43], non sempre ciò che è legale è automaticamente etico. Per questa ragione, si rende necessaria una riflessione che tenga conto di entrambi gli aspetti, ma che sottolinei soprattutto come l'etica debba essere posta come base per la legislazione.

[40] "*In many of these cases, we have reason to suspect that programmers are not explicitly writing biases toward marginalized demographics into their software's code. Instead, it appears the biases in some sense implicitly emerge from the algorithms' operating on the data, mimicking the biases reflected in the data themselves*". G.M. Johnson, "Algorithmic Bias. On the Implicit Biases of Social Technology", in *Synthese*, 198 (2021), p. 9943.

[41] Esempi come questo rendono evidente il rischio che l'uomo della società algoritmica possa diventare una "categoria merceologica". A. Quarta, U. Mattei, *Punto di svolta. Ecologia, tecnologia e diritto privato. Dal capitale ai beni comuni*, Aboca Edizioni, Sansepolcro, 2018, p. 15.

[42] Uno dei peggiori algoritmi di questo genere, è stato utilizzato da "[…] *Volkswagen in 11 million vehicles worldwide to deceive the emissions tests, and in particular to hide the fact that the vehicles were emitting nitrogen oxide at up to 35 times the levels permitted by law. And although it seemed simply like a devious device, this qualifies as an algorithm as well. It was trained to identify and predict testing conditions versus road conditions, and to function differently depending on that result.* […] *it was designed to deceive*". C. O'Neil (2017), "How can we stop algorithms telling lies?", in *The Guardian*. Recuperato da https://www.theguardian.com/technology/2017/jul/16/how-can-we-stop-algorithms-telling-lies, [Data di consultazione: 15/03/2026].

[43] High-Level Expert Group on Artificial Intelligence, *Orientamenti etici per un'IA affidabile*, 8 aprile 2019, pp. 13-14.

3.1. Metodi tecnici di intervento sugli algoritmi

In generale, vi sono tre categorie di intervento per alleviare l'influenza dei *bias* e migliorare l'efficacia algoritmica[44].

La prima è il *pre-processing* che interviene direttamente sui *dataset* di addestramento, rendendoli più eterogenei grazie a differenti metodologie. Una prima modalità, consiste nella modifica delle etichette di classificazione dei dati, in particolare per quelle istanze che si trovano in prossimità del confine decisionale. In tal modo, si interviene direttamente sul modo di "ragionare" degli algoritmi, riequilibrando le probabilità di ricevere un trattamento equo. Tuttavia, il pregio di questo metodo potrebbe anche essere considerato il suo difetto: infatti, la sua semplicità e immediatezza, aggiusta in tempi brevi alcune discriminazioni, ma al contempo è applicabile quasi esclusivamente in contesti dove l'algoritmo ha a disposizione poche azioni fra cui scegliere. Una seconda tecnica, consiste nella ripesatura delle istanze, a seconda del gruppo di appartenenza delle persone che l'algoritmo considera[45]. Ai gruppi ad alto rischio di penalizzazione, viene conferito un peso maggiore così da rendere più equilibrato il modo in cui l'algoritmo li percepisce rispetto ad altri. Il pregio di questo approccio è il mantenimento dell'integrità del set di dati originario, mentre un possibile difetto consiste nel dare troppa rilevanza alla scelta del peso, il che può causare squilibri inaspettati.

La seconda categoria di alleviamento consiste nell'*in-processing*, ossia nel tentativo di modificare il modello durante l'addestramento, incidendo di volta in volta sui *bias* riscontrati. Uno degli approcci di *in-processing* più utilizzati, è quello di intervenire sui criteri che guidano la suddivisione dei nodi nei modelli ad alberi di decisione, capaci di classificare le istanze di enormi quantità di dati. Affianco ad un criterio meramente classificatorio, alcuni propongono di porre l'accento sull'impatto della suddivisione sulle categorie di persone che rientrano nel *dataset*. L'obbiettivo è quello di bilanciare l'accuratezza della classificazione con l'equità decisionale, per evitare che la prima prevalga sulla seconda nei casi più critici. Una seconda tecnica, prevede l'aggiunta di vincoli specifici, per scongiurare che l'accuratezza previsionale minacci il trattamento simile di individui simili. Tale somiglianza, viene stabilita grazie a caratteristiche rilevanti, senza far ricorso alle

[44] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab, "Bias in Data-driven Artificial Intelligence Systems. An Introductory Survey", in *WIREs Data Mining and Knowledge Discovery*, 10 (2020), n. 3, pp. 6-7.

[45] Per comprendere un'applicazione specifica del cosiddetto *reweighting*, si legga il seguente studio: F. Hoitsma, G. Nápoles, Ç. Güven, Y. Salgueiro, "Mitigating Implicit and Explicit Bias in Structured Data Without Sacrificing Accuracy in Pattern Classification", in *AI & Society*, 40 (2024), pp. 2551-2570.

solite categorie possibilmente discriminatorie, con l'obbiettivo di minimizzare il diverso trattamento di individui simili[46].

L'ultima modalità è quella di *post-processing*, che permette di intervenire su un modello già addestrato, analizzandone i risultati critici per correggerli e migliorare le prestazioni dell'algoritmo oppure intervenendo sulla struttura interna del modello. Quest'ultimo approccio, detto *white box*, è molto meno utilizzato, soprattutto grazie alle possibilità di intervento di *in-processing* già in grado di evitare previamente possibili distorsioni. Al contrario, il primo approccio, ossia quello di *black box*, può essere considerato una sorta di correttore del lavoro di *pre-processing* e di *in-processing*.

La varietà di tutti questi metodi è un segno evidente della complessità degli algoritmi e del loro campo d'azione. Per questo motivo, va accompagnata da una seria riflessione etica, per evitare che il contributo umano sia meramente di carattere tecnico e per promuovere la centralità della persona, a prescindere dalla decisionalità algoritmica[47].

3.2. L'algoretica e l'umanizzazione degli algoritmi

Il concetto di "algoretica" compare per la prima volta negli scritti di Luigi Vallauri[48], con lo scopo di aprire un nuovo campo di riflessione che riguardi le relazioni e i punti di contatto fra bioetica e roboetica. Queste due discipline, apparentemente molto distanti, secondo Vallauri hanno in comune l'impiego di algoritmi: da una parte nella struttura del DNA, dall'altra nelle modalità di

[46] "*We capture fairness by the principle that any two individuals who are similar with respect to a particular task should be classified similarly. In order to accomplish this individual-based fairness, we assume a distance metric that defines the similarity between the individuals. This is the source of 'awareness'".* C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, "Fairness Through Awareness", in *Innovations in Theoretical Computer Science Conference*, Association for Computing Machinery, New York, 2012, p. 215.

[47] "*Advanced algorithms inscrutable to human inspection increasingly do the work of labeling us as combatant or civilian, good loan risk or future deadbeat, likely or unlikely criminal, hireable or unhireable. For these reasons, a contemporary theory of ethics* […] *must include an explicit conception of how to live well with technologies, especially those which are still emerging and have yet to become settled, seamlessly embedded features of the human environment. Robotics and artificial intelligence, new social media and communications technologies, digital surveillance, and biomedical enhancement technologies are among those emerging innovations that will radically change the kinds of lives from which humans are able to choose in the 21st century and beyond. How can we choose wisely from the apparently endless options that emerging technologies offer?* […] *Are we prepared to choose well? This question involves the future, but what it really asks about is our readiness to make choices in the present*". S. Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*, Oxford University Press, Oxford, 2016, p. 3.

[48] L. Lombardi Vallauri, "Algoretica. Le due sfide cruciali nell'era tecnologica: bioetica, roboetica", in *Atti e memorie dell'Accademia Toscana di Scienze e Lettere "La Colombaria"*, vol. 82, Leo S. Olschki Editore, Firenze, 2017, pp. 353-376.

funzionamento dell'intelligenza artificiale e dei robot. Egli, a questo proposito argomenta:

> Bioetica e roboetica si possono allora unificare nel concetto di algoretica: l'algoritmo biologico si riflette nell'algoretica umana, animale e ambientale, cioè l'etica del vivente, del bios; l'algoritmo non biologico si riflette nell'algoretica dei robot e dei sistemi di Intelligenza Artificiale; ossia nell'etica concernente gli algoritmi stessi[49].

L'algoretica, pertanto, si configura come l'etica che riflette sul ruolo e sull'impatto degli algoritmi nel mondo biologico e in quello artificiale. Per quanto riguarda il primo, l'attenzione si focalizza sull'uso di algoritmi "biologici", come nel caso del codice genetico. L'algoretica delle macchine, invece, affronta l'uso di algoritmi informatici, di qualsiasi tipo siano.

L'unione fra questi due ambiti, a suo parere, è fondamentale per evitare di cadere nell'errore di considerare gli algoritmi e i dispositivi elettronici come cattivi o discriminatori in sé stessi. Piuttosto, è l'uomo ad essere intrinsecamente malvagio e a contagiare gli algoritmi, tanto che, secondo lui, è talvolta dubbio che egli sia moralmente superiore ad essi[50].

Una prospettiva differente, invece, è quella assunta da padre Paolo Benanti, secondo il quale lo sviluppo dell'algoretica dipende fortemente dal concetto di "valore"[51]. Gli algoritmi, ad esempio, non comprendono cosa sia il valore morale di un azione, essendo in grado di lavorare solo su valori puramente numerici. Una possibile soluzione, ad uno sguardo superficiale, potrebbe essere quella di tradurre i principi etici nel loro corrispettivo numerico, così che la macchina possa comprenderli e addestrarsi attraverso di essi. Tale obbiettivo, tuttavia, è irraggiungibile poiché l'intelligenza artificiale e gli algoritmi, non hanno coscienza, moralità e semantica.

Alcuni ricercatori dell'istituto *Future of Life*, guidati da Vincent Conitzer, tuttavia, stanno cercando di approfondire le modalità in cui l'essere umano arrivi a compiere scelte morali, per poi applicare tale processo anche all'interno della struttura delle intelligenze artificiali. Secondo questo studio, la replica del processo decisionale, dovrebbe passare attraverso l'utilizzo del *machine learning* per identificare quali *patterns* emergano e determinare con quale estensione si possano

[49] L. Lombardi Vallauri, "Algoretica e Informatica giuridica", in *I-lex*, 15 (2022), n. 1, p. 32.
[50] Vallauri pare ignorare completamente gli algoritmi addestrati non intenzionalmente in base alle ricerche degli utenti, nonché la capacità umana di migliorare gli algoritmi problematici. *Ivi*, p. 33.
[51] Il termine ricorre anche nella Dichiarazione di Montréal, nell'enunciazione del principio di equità: "L'attività digitale degli utenti dei sistemi IA e dei servizi digitali deve essere riconosciuta come lavoro che contribuisce al funzionamento degli algoritmi e crea valore". Dunque, le attività degli utenti possono rientrare nei set degli algoritmi per favorire il loro addestramento positivo. Il termine "valore" indica qui la capacità di creare un bene e ottenere un certo effetto, ma possiede altresì una connotazione morale. *Dichiarazione di Montréal sullo sviluppo responsabile dell'intelligenza artificiale*, 2018, p. 13.

replicare le decisioni umane nelle macchine[52]. Ad un'attenta analisi, però, si nota come il punto principale di questa ricerca non sia la comprensione di come la persona scelga, ma la capacità di predire quale tipo di decisioni si verifichino in certe circostanze, per poi tradurre tutto ciò nel linguaggio del computer. Dunque, tale carattere probabilistico, rende evidente come le modalità decisionali non siano replicabili dal punto di vista dell'interiorità, ma (al massimo) soltanto da quello del comportamento esteriore[53].

Data questa impossibilità, Benanti propone di introdurre nelle macchine un imperativo centrale, ossia "dubita di te stessa"[54]. Ci si potrebbe chiedere, però, come sia possibile implementare qualcosa che, di fatto, fa parte della capacità di pensare, caratteristica esclusivamente umana[55].

Secondo lo studioso francescano, questa direttiva "si ottiene introducendo dei paradigmi statistici all'interno delle AI", tanto che "tentativi di questo tipo sono portati avanti da Google e Uber con delle librerie statistiche speciali. Deve essere questa capacità di incertezza il cuore decisionale della macchina"[56].

L'incertezza, dunque, porta la macchina a bloccarsi e interrogare l'uomo, richiedendo il suo intervento, realizzando così quell'*human-centred AI* tanto auspicata dalle più importanti carte etiche, come si vedrà a breve. L'algoretica, di conseguenza, dovrà sviluppare un proprio linguaggio che tenga conto di molteplici aspetti e in particolare di queste quattro caratteristiche: anticipazione, trasparenza, *customization*, adeguamento.

La prima, è una qualità tipicamente umana, che emerge quando due persone collaborano fra loro: l'una anticipa le azioni dell'altra, riuscendo ad intuirne le intenzioni. Allo stesso modo, le macchine dovrebbero essere messe in grado di comprendere ciò che gli uomini vogliono o non desiderano fare, assecondandone la volontà e adattandovisi[57].

[52] R. Kumar (2016), "AI Researcher Vincent Conitzer", in *Future of Life*. Recuperato da https://futureoflife.org/ai-researcher-profile/ai-researcher-vincent-conitzer/, [Data di consultazione: 12/03/2026].

[53] Il comportamentismo è una componente fondamentale di alcuni studi sull'intelligenza artificiale e sulla possibilità di replicare la mente umana. J.R. Searle, "Minds, Brains and Programs", in *Behavioural and Brain Sciences*, 3 (1980), pp. 417-457; J.R. Searle, "L'intelligenza artificiale e gli studi sulla coscienza", in Aa. Vv., *L'uomo alla ricerca della verità. Filosofia, scienza, teologia: prospettive per il terzo millennio*, Vita e Pensiero, Milano, 2005, pp. 201-211.

[54] P. Benanti (2019), "Algor-etica: l'incertezza ci salverà", in *Paolo Benanti*. Recuperato da https://www.paolobenanti.com/post/2019/02/13/algor-etica-ricerca-universale, [Data di consultazione: 13/03/2026].

[55] Chi dubita, infatti, pensa e gli algoritmi non possono pensare. Si veda Cartesio su dubbio e pensiero: R. Decartes, *Meditazioni metafisiche*, Laterza, Bari, 1997, p. 47.

[56] P. Benanti, "Algor-etica: l'incertezza ci salverà", cit.

[57] Infatti, "è possibile definire un tipo di responsabilità in cui l'agente influenza le scelte di un altro agente. Questo non diminuisce la responsabilità piena dell'agente che ha compiuto l'azione ma determina una qualche forma di responsabilità anche per l'agente che lo ha influenzato nel suo comportamento e nelle sue scelte. Questo secondo tipo di responsabilità può essere messo in relazione con i sistemi algoritmici. Infatti, allo stato attuale dello sviluppo tecnologico, non si può

Il secondo aspetto, strettamente legato alle suddette procedure di *pre* e *in-processing*, raccomanda di non porre come obbiettivo primario delle macchine l'ottimizzazione delle loro azioni, ma la massima spiegabilità e intuitività degli algoritmi che le guidano.

La *customization* porta la macchina ad adattarsi non solo alla volontà dell'uomo, ma altresì alla sua personalità ed emotività. Così facendo, potrà aggiustare il proprio "comportamento" quando vi coopererà ed evitare che la convivenza risulti insopportabile.

L'ultimo tratto, è l'adeguazione dei fini della macchina a quelli dell'uomo, in ogni specifica situazione. Essa, perciò, dovrà possedere una sorta di "umiltà artificiale" che conferisca la priorità alla salvaguardia della persona umana[58].

Tale impostazione antropocentrica costituisce altresì l'approccio fondamentale dell'Unione Europea alla questione, come è particolarmente riscontrabile negli *Orientamenti etici per un'IA affidabile*. Questo documento ritiene che per costruire un'intelligenza artificiale affidabile per l'uomo siano necessarie tre condizioni che devono sempre intersecarsi fra loro durante l'intero ciclo di vita dei sistemi. La prima è la legalità, che prevede che l'intelligenza artificiale sia conforme alle leggi e ai regolamenti, la seconda è l'eticità, per cui l'intelligenza artificiale deve aderire a valori morali, la terza è la robustezza tecnica e sociale, volta ridurre i possibili danni causati dai sistemi. Su questa base poggiano quattro principi etici, ossia il rispetto dell'autonomia umana, la prevenzione dei danni, l'equità e l'esplicabilità. Da tali principi discendono poi sette requisiti pratici, tra cui l'intervento e la sorveglianza umana, la robustezza tecnica e la sicurezza, la privacy e la *governance* dei dati, la trasparenza, la non discriminazione, il benessere sociale e ambientale e l'accountability.

Tutti questi requisiti e in particolare il primo, che ha lo scopo di mettere in atto il principio del rispetto dell'autonomia, manifestano la volontà di porre al centro la persona umana sia dal punto di vista tecnico sia da quello etico-giuridico, vietando che l'uso dei sistemi ne comprometta la dignità, la libertà e l'autodeterminazione. Nello specifico, la realizzazione di un'intelligenza artificiale veramente *human-centered*, passa attraverso tre vincoli. Il primo, detto *human-in-the-loop*, prevede la possibilità di un intervento umano in ogni ciclo decisionale del sistema intelligente[59]. Il secondo, denominato *human-on-the-loop*, raccomanda che l'uomo possa intervenire durante la progettazione e i test sul funzionamento effettivo della macchina. Il terzo, chiamato *human-in-command*, indica la necessità

ancora riconoscere una responsabilità piena ad un agente artificiale ma si può riconoscere ad esso almeno l'*accountability* (imputabilità) per l'influenza che ha avuto nell'agente umano". A. Picchiarelli, "L'algoretica e lo sviluppo etico dell'Intelligenza Artificiale", in R. Razzante (a cura di), *Manuale sull'Intelligenza Artificiale*, Giappichelli, Torino, 2024, p. 8.

[58] P. Benanti, "Algor-etica: l'incertezza ci salverà", cit.

[59] La priorità operativa è della persona, non dell'algoritmo. P. Benanti, *Human in the loop. Decisioni umane e intelligenze artificiali*, Mondadori, Città di Castello, 2022, pp. 130-131.

che l'uomo abbia sotto controllo l'intera attività del sistema e la possibilità di stabilirne le modalità di utilizzo, il blocco o la sospensione[60].

Secondo alcuni, tuttavia, i riferimenti all'umanizzazione e alla centralità della persona rischiano di rimanere ambigui se non si chiariscono previamente quali concezioni antropologiche e contesti culturali siano effettivamente valorizzati. Sul tema, la lettura decoloniale dell'intelligenza artificiale ritiene che le implicazioni e i significati dietro questa terminologia siano di matrice prevalentemente occidentale o eurocentrica e che, pertanto, non posseggano quell'universalità che pretendono di promuovere[61]. Ad esempio, questa prospettiva, ritiene che non sia sufficiente parlare di "essere umano" in generale, in quanto tale termine è solitamente riferito a un soggetto che è titolare di diritti, libero da coercizioni e inquadrato in governi democratici. Al contrario, il concetto dovrebbe altresì includere le asimmetrie esistenti fra Nord e Sud del mondo, le condizioni materiali entro cui i sistemi sono progettati, addestrati e distribuiti, nonché le forme di datificazione che coinvolgono persone collocate in contesti di grande vulnerabilità[62].

La varietà di tali rilievi, pur suscettibili di diverse valutazioni, non sembra tuttavia scalfire in modo decisivo la prospettiva di padre Benanti, nella misura in cui egli non propone una umanizzazione dell'algoritmo che rispecchi una determinata immagine socio-culturale, ma sottolinei la necessità di subordinare il suo funzionamento alla dignità ontologica dell'uomo[63], alla sua personalità e alla sua emotività.

Un'esigenza simile si riscontra nelle riflessioni della studiosa americana Flynn Coleman, la quale propone una umanizzazione che prenda in considerazione sia l'essere profondo della persona umana sia la sua "multiforme diversità[64]":

> *A human algorithm is the philosophical center of ourselves. It's our personal and collective ethos; it's the DNA of our humanity; it's our conscience. All the things that make us who we are, in all of our misery and majesty, fallibility and brilliance. Just as celestial navigation enables us to chart the heavens and*

[60] High-Level Expert Group on Artificial Intelligence, *Orientamenti etici per un'IA affidabile*, 8 aprile 2019, p. 18.

[61] Al posto di promuovere l'universalismo, si propone come soluzione un "pluriversalismo" che tenga conto di più prospettive diverse ed arricchenti. Per approfondire: S. Mohamed, M.T. Png, W. Isaac, "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence", in *Philosophy & Technology*, 33 (2020), pp. 659-684; per una lettura critica di questa prospettiva si legga: M. Tironi, C. Albornoz, "Decolonizing AI? Lessons from a Failed Experiment", in *Big Data & Society*, 12 (2025), n. 3, pp. 1-6.

[62] A. Birhane, "Algorithmic Colonization of Africa", in *SCRIPTed. A Journal of Law, Technology & Society*, 17 (2020), n. 2, pp. 389-409.

[63] "[…] la dignità umana è propria di tutti gli esseri umani in quanto umani. Ciò conferisce alla dignità umana quell'universalità senza cui perderebbe di senso". F. Viola, "I volti della dignità umana", in *Colloqui sulla dignità umana*, Aracne, Roma, 2008, p. 107; si veda anche: Tommaso d'Aquino, *Summa theologiae*, I, q. 93, a. 4.

[64] Questo secondo aspetto può essere altresì riletto alla luce di quel "pluriversalismo" accennato nella nota 61.

> *find our way, our human algorithms guide us to our values, our kindness, our compassion, our bravery, our humanness. The more multifaceted the diversity, the more human the algorithm: never perfect, yet ever more humane*[65].

Un algoritmo umano, pertanto, è un algoritmo a misura d'uomo, non nel senso de-soggettivante visto sopra, ma piuttosto nel senso che tutto ciò che è umano entra o dovrebbe entrare nella progettazione algoritmica, tenendo a mente che "il rispetto fondamentale per la dignità umana postula di rifiutare che l'unicità della persona venga identificata con un insieme di dati" [66]. Tale concetto, non offre una soluzione tecnica immediata per creare degli algoritmi perfetti, ma sprona a domandarsi chi si vuole essere e che valori si desiderano preservare. Solo quando vi saranno certezze su questi punti fondamentali, si potranno veramente costruire algoritmi di cui fidarsi, proprio perché questi ultimi sono considerabili come un riflesso dell'essere umano[67].

In definitiva, per costruire un algoritmo di questo tipo è necessario decidere quale via intraprendere, accettare la fallibilità dell'umano, domandarsi che rapporto si intrattenga con gli altri e con il creato e dirigersi sempre più, sulla via del bene, verso ciò che si desidera essere[68].

[65] F. Coleman, *A Human Algorithm: How Artificial Intelligence is Redefining Who We Are*, Counterpoint Press, Berkeley, 2019, p. 234.
[66] Francesco, *Messaggio per la LVII Giornata Mondiale della Pace*, VI, 1 gennaio 2025.
[67] Come sottolinea Coleman, è un concetto "metafisico", che non indica una soluzione tecnica immediata. Con "valori", inoltre, l'autrice non si riferisce semplicemente ai valori etici, ma anche a caratteristiche personali come la creatività, l'empatia, la gentilezza, il lavoro di gruppo, la compassione etc. *Ibidem*.
[68] Si veda lo schema di costruzione di tale algoritmo "non tecnico", che parte dal concetto di valore inteso in "senso" etico-esistenziale. *Ivi*, pp. 232-233.