Rivista Italiana di Filosofia Politica

Firenze University Press
www.fupress.com/rifp

Una prima tassonomia delle sfide politiche poste dall'intelligenza artificiale

An Initial Taxonomy of the Political Challenges Raised by Artificial Intelligence

Gabriele Giacomini

Università degli Studi di Udine
gabriele.giacomini@uniud.it

Abstract. Artificial intelligence (AI) has recently undergone a radical evolution, transitioning from a symbolic nature (e.g., expert systems), where humans imparted explicit rules, to a sub-symbolic one (e.g., machine learning systems) where statistical relationships are sought in large amounts of data (big data). To date, this paradigm shift remains little known to the general public. The article analyses the nature of this "alien intelligence", so termed for its peculiar and, to some extent, obscure way for human beings to process information. Based on these characteristics, it explores three potential issues emerging from the widespread use of AI, emphasising that every innovation has consequences for power balances and social dynamics: the challenge of surveillance and the use of force; that of the quality of the public sphere; and finally, that of intellectual substitution and inequality. The article concludes by highlighting the necessity of a profound ethical and regulatory examination of AI in the coming decades.

Keywords: artificial intelligence, automatic learning, surveillance, public sphere, inequality.

Riassunto. L'intelligenza artificiale (IA) ha compiuto recentemente un'evoluzione radicale, passando da una natura simbolica (ad esempio, i sistemi esperti), dove gli umani impartivano regole esplicite, ad una sub-simbolica (ad esempio, i sistemi di apprendimento automatico) dove si cercano relazioni statistiche su grandi quantità di dati (big data). Ad oggi, questo cambio di paradigma rimane poco noto al gran-

Rivista Italiana di Filosofia Politica 5 (2023): 183-213
ISSN 2785-3330 (print) | DOI: 10.36253/rifp-2327

de pubblico. L'articolo analizza la natura di questa "intelligenza aliena", così denominata per il suo modo peculiare, e in parte oscuro per gli esseri umani, di processare informazioni. A partire da queste caratteristiche, si esplorano tre potenziali questioni emergenti dall'uso diffuso dell'IA, sottolineando che ogni innovazione ha conseguenze su equilibri di potere e dinamiche sociali: la sfida della sorveglianza e dell'uso della forza; quella della qualità della sfera pubblica, infine quella della sostituzione intellettuale e della diseguaglianza. L'articolo si conclude sottolineando la necessità di un profondo esame etico e normativo dell'IA nei prossimi decenni.

Parole chiave: intelligenza artificiale, apprendimento automatico, sorveglianza, sfera pubblica, diseguaglianza.

Premessa

I recenti successi dall'IA dipendono da un cambiamento di paradigma tecno-scientifico che ha permesso di superare la programmazione classica e i sistemi esperti, in cui la macchina veniva istruita dagli umani attraverso la formulazione di regole esplicite, per approdare a un modello di apprendimento automatico, in cui i programmatori si limitano a indicare una direzione e a offrire una quantità enorme di dati, e in cui è l'IA a identificare le relazioni statistiche più efficaci a raggiungere il risultato richiesto[1]. Questo passaggio di paradigma, avviato nell'ultimo scorcio del Novecento e affermatosi da circa una quindicina di anni[2], è perlopiù sconosciuto ai non "addetti ai lavori". Esserne consapevoli è, quindi, un punto di partenza per inquadrare la novità degli odierni sistemi di intelligenza artificiale, e un'assunzione fondamentale per qualsiasi discorso di tipo filosofico, etico e normativo[3].

In questo quadro, il presente articolo si propone di perseguire due obiettivi. Il primo è quello di presentare quella che l'informatico Cristianini ha definito una "intelligenza aliena"[4]: "intelligenza" perché è in grado di agire in un ambiente incerto utilizzando informazioni per prendere decisioni che aumentano le probabilità di successo, "aliena" in quanto uti-

[1] Si è passati da un paradigma in cui è lo specialista del dominio a definire delle "feature" che poi il modello utilizza per i suoi compiti, a uno in cui è l'algoritmo di apprendimento stesso ad apprendere anche tali "feature", e ciò si è dimostrato estremamente utile in casi come il trattamento di informazioni complesse.

[2] In particolare, il cambio si avvia negli anni '80 in base a una intuizione di Frederick Jelinek e del suo gruppo di ricerca. Su questo passaggio storico si consiglia Cristianini, *La scorciatoia*.

[3] Per una prima introduzione filosofica all'IA e alle sue varie applicazioni: Fossa, Schiaffonati e Tamburrini, *Automi e persone*.

[4] Cfr. Cristianini, *La scorciatoia*.

lizza un "linguaggio" radicalmente diverso da quello degli umani, tanto differente da essere opaco e di ardua gestione. Il secondo obiettivo è quello di presentare (o forse si dovrebbe dire anticipare) tre categorie di problemi che potrebbero avere un impatto rilevante sulla società, e che potrebbero insistere in aree su cui il pensiero filosofico-politico ha tradizionalmente proposto sistemi e teorie. Ci si focalizza su probabili problemi non perché si esclude che l'IA sia una tecnologia gravida di straordinarie potenzialità di progresso. Al contrario: l'impressione è che l'IA possa essere in grado di sostenere un "salto qualitativo" per la specie umana, dall'ambito medico[5] a quello economico[6], da quello della mobilità[7] a quello dell'ambiente[8], da quello giuridico-amministrativo[9] a quello della ricerca e della cultura[10]. Piuttosto, vengono evidenziati dei problemi per due motivi. Il primo è una buona notizia: per sviluppare le potenzialità positive servirà lavoro, impegno e applicazione, ma dove i risultati saranno ritenuti comunemente benefici non sarà necessario molto spirito critico. Il secondo è che qualsiasi innovazione, per quanto possa essere entusiasmante come risultato netto, ha dei vincitori (auspichiamo molti) e dei vinti (speriamo il meno possibile), cambia gli equilibri di potere, e può influire nei modi in cui alcuni gruppi di persone vivono concretamente, sia in rapporto con le macchine sia fra di loro.

Questo articolo, invece, non intende avanzare teorie di tipo etico e prescrittivo circa l'IA e la politica: questo lavoro, senza dubbio sterminato data l'ampiezza degli ambiti di applicazione dell'IA, dovrà essere compiuto dalla comunità scientifica nei futuri anni, probabilmente nei prossimi decenni. Per questioni di selezione e di ordine, la letteratura di riferimento prende le mosse dal dibattito nazionale, nella misura in cui riflette gli orientamenti presenti a livello internazionale e senza rinunciare a segnalare pubblicazioni straniere laddove temi rilevanti o risultati significativi non siano ancora penetrati in Italia.

---

[5] Diagnostica precoce delle malattie, analisi di dati sanitari e immagini mediche, sviluppo di nuovi farmaci, previsione di epidemie, robotica assistiva.

[6] Ottimizzazione della catena di approvvigionamento, automazione dei processi produttivi, analisi delle tendenze del mercato, analisi algoritmica per il trading.

[7] Veicoli autonomi, ottimizzazione del traffico, sicurezza stradale, manutenzione predittiva dei mezzi di trasporto.

[8] Gestione delle reti energetiche, rilevamento delle perdite, ottimizzazione delle coltivazioni, previsione delle malattie delle piante, gestione sostenibile delle risorse.

[9] Analisi e ricerca di documenti legali, previsione dei risultati dei processi, automazione dei servizi pubblici, analisi per le politiche pubbliche.

[10] Analisi di grandi set di dati, simulazioni complesse, assistenti virtuali per scrittori e giornalisti, creazione di contenuti personalizzati, traduzioni istantanee, rappresentazioni grafiche, sostegno nella creatività artistica.

La portata della sfida. Intelligenze aliene, opacità, bias

La risposta alla domanda "l'IA è intelligente?" dipende da come scegliamo di definire l'intelligenza. Detto altrimenti, "l'IA è intelligente?" deriva da "che cosa è intelligente?". Nel dibattito scientifico sull'IA possiamo identificare due strade.

La prima definizione di intelligenza è *dipendente* dal confronto con il modello di cognizione umana, ed è legata all'approccio suggerito da Turing a metà Novecento[11]. Nel dibattito contemporaneo, un esponente di questa visione è Luciano Floridi[12]. In questo senso, l'intelligenza di qualsiasi ente è misurata sulla base di quanto essa assomigli agli esseri umani. Alan Turing, nel suo celebre test, suggerì che una macchina può essere considerata intelligente se il suo comportamento si avvicina a quello di un essere umano, fino a diventare indistinguibile. In questo approccio, il confine tra ciò che è considerato intelligente e ciò che non lo è si concentra su capacità tipicamente umane, come l'uso del linguaggio. Si presume che ci debba essere una qualità specifica, come avere un cervello, un linguaggio o una coscienza, che rende un agente "intelligente" davvero tale. Fino a quando le macchine intelligenti erano poco sviluppate, questa definizione operativa proposta da Turing poteva funzionare: il raggiungimento delle capacità cognitive umane da parte da sistemi automatici appariva lontano, dunque il comportamento umano poteva essere un riferimento funzionale per la comparazione con tutti gli altri comportamenti.

La seconda definizione è *indipendente* dal confronto con l'essere umano, ovvero è indifferente alle caratteristiche umane, e appare in Albus[13]. Nel confronto attuale, questa prospettiva è difesa ad esempio da Cristianini[14]. In questo senso, un ente è intelligente se si comporta in determinate maniere, non perché assomiglia agli esseri umani in alcune "caratteristiche chiave". Ciò che è cruciale in questa definizione, invece, è l'interazione tra l'obiettivo dell'agente e l'ambiente circostante (*task-environment*). In particolare, un agente è considerato intelligente se è in grado di agire in un ambiente incerto utilizzando informazioni per prendere decisioni che aumentano le sue probabilità di successo. Secondo questa visione, l'idea che il comportamento sia orientato verso un obiettivo è fondamentale. In biologia, ad esempio, lo scopo primario di un organismo è la sopravvivenza di sé e dei suoi geni. Il punto è che la capacità di comportarsi in modo efficiente in situazioni nuove non è un'esclusiva degli esseri umani, ma si

[11] Turing, "Computing machinery and intelligence," 235, 433-60.
[12] Il suo maggior contributo all'etica dell'IA è Floridi, *Etica dell'intelligenza artificiale*. Sull'intelligenza artificiale "non intelligente" si veda anche Esposito, *Comunicazione artificiale*.
[13] Albus, "Outline for a theory of intelligence," 473-509.
[14] Cristianini, *La scorciatoia*.

manifesta in piante che reagiscono agli stimoli ambientali, in colonie di formiche che prendono decisioni complesse sul luogo ideale per costruire un nido, o in software progettati per adattarsi a nuove informazioni che raccolgono su Internet[15]. Pertanto, secondo questa prospettiva filosofica l'intelligenza non ha solo una "modalità umana" ma può manifestarsi in molteplici forme.

Questi due approcci condizionano la diagnosi sulla "natura" degli attori di intelligenza artificiale. In base alla definizione che scegliamo, daremo una valutazione diversa sulle possibili implicazioni etiche, sociali e politiche dell'IA.

Coloro che condividono la definizione di intelligenza *dipendente* dall'umano tendono ad avere una lettura antropocentrica dell'IA. L'idea è che l'IA esegue con creatività e successo molti compiti senza essere minimamente intelligente (nel senso umano del termine). L'essere umano, con la sua peculiare forma di coscienza, resta il modello imprescindibile, il che rende impossibile "a priori" che la copia (l'intelligenza artificiale) possa "superare" l'originale (l'intelligenza umana). Non a caso, nel 2022 Floridi sostiene che l'IA non abbia ancora superato il test di Turing[16]. Il test di Turing è, dal 1950, la più celebre definizione operativa di intelligenza, e richiede che un computer inganni un giudice umano e gli faccia credere che sta conversando con un'altra persona (e non con un computer). Secondo Floridi l'attuale IA, come ad esempio ChatGPT, non è ancora in grado di rispondere ad alcuni quesiti logici particolarmente avanzati, facendo palesemente errori, e quindi non supera il test. Un problema di questa valutazione è che, in realtà, quote non trascurabili della popolazione umana alfabetizzata non sono in grado di compiere compiti logici e matematici (anche moltiplicazioni e divisioni, per non parlare di rispondere a "trucchi logici"). Probabilmente Floridi prende come modello umano un suo collega di filosofia della scienza all'Università di Oxford: non considera tanto l'intelligenza umana (nella sua media), quanto i suoi "pinnacoli", le vette più alte. Ma questo è un piccolo trucco. Piuttosto, la nostra impressione è che un'IA come ChatGPT non superi il test di Turing perché lo superi troppo. Detto altrimenti, se il criterio per superare il test di Turing è quello di essere confondibile con un essere umano, allora è vero che una IA tradisce la sua differenza, ma perché, quando viene interrogata, dimostra – al netto di possibili bias e "allucinazioni" – di saper maneggiare , in maniera discreta o addirittura buona, troppe questioni su argomenti troppo distanti, e con una velocità palesemente sovrumana.

---

[15] L'obiettivo di superare una concezione antropocentrica dell'intelligenza, riconoscendola a soggetti sia naturali sia artificiali, è presente anche in Giacomelli. Cfr. Giacomelli, *Dall'umanità in poi*.

[16] Quello di Turing, scrive Floridi, "è un test, la cui soglia è davvero molto bassa, eppure nessuna IA l'ha mai superata." (Floridi, *Etica dell'intelligenza artificiale*, 272).

Invece, coloro che propendono per la definizione di intelligenza indipendente dell'essere umano sottolineano l'irriducibile diversità dell'artificiale. Cristianini, in particolare, definisce le IA come "aliene"[17]. Questa definizione ha il pregio di abbandonare concezioni antropocentriche dell'intelligenza, evitando quindi di etichettare come "non intelligenti" azioni che superano di gran lunga specifiche performance cognitive umane[18]. Al tempo stesso, grazie all'idea secondo cui l'intelligenza ha molti modi per manifestarsi, questa prospettiva non rende possibili "graduatorie unidimensionali" fra gli enti intelligenti, ma permette di riconoscere le peculiarità di umani, animali, sistemi artificiali[19]. Entrando più nello specifico, le specializzazioni cognitive degli esseri umani sono pensare in termini di oggetti concreti, di agenti, di causa effetto, di geometria elementare. Gli umani sono in grado di realizzare compiti motori complessi, come allacciarsi le scarpe (compito ad oggi impossibile per l'IA). Gli esseri umani, invece, sono mediamente incapaci di eseguire calcoli che richiedono anche una minuscola "memoria di lavoro"[20], non sono in grado di leggere un codice QR o di comprendere il mondo della fisica quantistica, dove gli oggetti sono sostituiti da onde e non hanno una posizione definita[21]. Ciò ha due implicazioni. La prima è che la realizzazione di una "intelligenza universale", biologica e/o artificiale, sarebbe una questione mal posta[22]: piuttosto, secondo la prospettiva di Cristianini, esistono tanti modi intelligenti di compiere molti tipi di azioni. La seconda è che, così come esistono intelligenze (come l'umana) superiori a quelle di una macchina in vari comportamenti, esistono anche intelligenze "sovrumane" (nel senso stretto di "superiori a quella umana") in diversi compiti specifici.

Questi due approcci divergono non solo nel concepire l'identità umana (e il "posto" dell'intelligenza umana) in rapporto alle tecnologie, ma anche nella valutazione delle possibilità che gli individui e le istitu-

---

[17] Cfr. Cristianini, *La scorciatoia*, 20-4.

[18] Si pensi a sistemi di IA in grado di battere a scacchi i campioni del mondo, di fare diagnosi di cancro in base alla lettura di TAC con una possibilità minore di incorrere in falsi negativi dei medici, oppure in grado di scoprire in una sola notte centinaia di nuove molecole farmacologiche.

[19] Giacomelli, invece, ritiene che l'intelligenza dovrebbe essere concepita come la capacità di un agente qualsiasi di introiettare – con diverse profondità a seconda del soggetto stesso – l'ordine esistente in natura. L'ordine naturale fa così da "gancio ontologico" per una valutazione delle intelligenze biologiche e artificiali. Cfr. Giacomelli, *Dall'umanità in poi*.

[20] La maggior parte delle persone non è in grado di eseguire a mente calcoli elementari, come 45+378:13.

[21] La fisica quantistica è poco intuitiva per la mente umana, eppure si dimostra corretta quando viene applicata.

[22] Facciamo riferimento, ad esempio, allo scenario della "superintelligenza", che vedrebbe una singola intelligenza artificiale avere performance eccellenti in tutti i compiti. Il riferimento è a Bostrom, *Superintelligenza*.

zioni hanno di comprendere le macchine e di governarle. Entrambe le prospettive, naturalmente, sono attente sia alle opportunità sia ai rischi dell'IA, ma pesano in maniera diversa i "margini di manovra" dell'essere umano. Il primo approccio propende verso una concezione "ottimistica": l'etica dell'intelligenza artificiale non solo è necessaria, ma è praticabile. Se l'essere umano è l'unico a essere (per autodefinizione) "veramente intelligente", questo gli garantirebbe la capacità di gestire anche le macchine più raffinate. Si tratterebbe di un compito non banale, ostacolato da problemi come l'opacità degli algoritmi e come i bias, ma sostenibile dalle intelligenze umane. In questa ottica (ad esempio: Floridi), le opacità degli algoritmi non dipenderebbero tanto dal sistema artificiale per come è in quanto tale, ma soprattutto da responsabilità umane (ad esempio, la causa principale starebbe nelle imprese private che tendono a nascondere i loro processi sotto il velo del diritto industriale). In maniera simile, i bias che viziano gli algoritmi (i sistemi di IA possono discriminare le persone in base a genere, etnia, età eccetera) dipenderebbero soprattutto dal fatto che "imparano" da dati prodotti da persone in carne ed ossa: questo stesso problema confermerebbe la preminenza della responsabilità umana, che si porrebbe alla "fonte" del processo (anche se questo porta a pregiudizi ed errori). È sottovalutata, invece, l'ipotesi che l'IA rischi di discriminare perché è insensibile a qualunque conseguenza negativa e poiché, a differenza delle persone, prende decisioni senza paura di essere sanzionata.

Il secondo approccio, riconoscendo la radicale differenza dell'intelligenza artificiale rispetto a quella umana, e sganciando in qualche modo la prima dal confronto con la seconda, sembra più "pessimista", sottolineando le sfide che l'IA porta alla capacità di gestione di individui e istituzioni. L'IA, secondo questi studiosi, porta inevitabilmente con sé un certo grado di "incertezza epistemica" che è insopprimibile del tutto. Proprio essendo intelligenze "aliene", che ragionano in manera diversa dagli esseri umani – ad esempio non secondo il criterio della causalità certa ma secondo quello di correlazione probabilistica, non selezionando le informazioni tramite euristiche ma processando quantità gigantesche di dati – le macchine possono giungere a sapere cose che gli umani possono non comprendere. Questo ridurrebbe il potere degli umani di gestire efficacemente le macchine. I maggiori rischi si manifestano soprattutto quando alle macchine viene chiesto di generare modelli enormi, e quindi non interpretabili, per fare analisi o previsioni in campi che gli esperti umani conoscono poco (il paradosso è che l'IA potrebbe essere utile soprattutto in campi in cui gli umani sono in difficoltà, ma è proprio in questi campi che il supporto artificiale potrebbe trasformarsi nel parere di un "oracolo"). Il governo umano di queste tecnologie è ancora percepito come necessario,

ma ci sono maggiori dubbi sul fatto che ciò sia praticabile, e questo per il modo stesso in cui sono concepite le IA di successo.

Approfondiamo la portata della questione secondo questa seconda prospettiva, che è quella potenzialmente più sfidante per i nostri sistemi sociali e politici. Per quanto riguarda l'opacità, ovvero il primo problema rilevante dell'IA, non è sufficiente fare riferimento al fatto che le aziende digitali sono recalcitranti a condividere informazioni sui loro sistemi (problema comunque diffuso). Bisogna aggiungere la constatazione informatica che le dimensioni delle astrazioni nei modelli d'intelligenza artificiale sono inumane per ampiezza ma anche per "linguaggio". Queste macchine, attraverso algoritmi e reti neurali profonde, individuano pattern e relazioni nei dati nel numero di miliardi, su un'immensa rete di unità interconnesse[23]. Inoltre, non è detto che queste relazioni si basino su categorie interpretabili da noi, ovvero potrebbero non essere riconducibili a nessuno dei concetti che gli umani utilizzano nel linguaggio quotidiano[24]. Il risultato è una sorta di conoscenza che è distribuita su milioni di parametri numerici, rendendo estremamente difficile (se non impossibile) per gli esperti umani comprendere appieno la logica alla base di tali decisioni. L'IA potrebbe operare su concetti "alieni" che non sono interpretabili o traducibili in termini umanamente comprensibili. Un esempio si è manifestato nel 2016, quando AlphaGo, un programma d'IA, ha affrontato nel Go il campione sudcoreano Lee Sedol[25]. Non solo AlphaGo ha vinto la partita, ma durante la mossa 37 della seconda partita ha preso una decisione che, agli occhi degli esperti umani e dello stesso Sedol, sembrava palesemente un errore. Tuttavia, quella mossa, umanamente incomprensibile, ha preparato il terreno per la successiva vittoria della macchina. Ciò evidenzia che la conoscenza "distillata" da AlphaGo durante il suo processo di autoaddestramento può essere indecifrabile e inafferrabile non solo per gli esseri umani in generale, ma per i massimi esperti mondiali di quel settore. Il problema è che sarebbe desiderabile interpretare umanamente le decisioni dell'IA per evitare che non ci siano esiti negativi "nascosti". In alcune decisioni, come quelle giuridiche, la procedura è sostanza. Ma si può pensare anche a procedimenti decisionali basati sull'IA, con conseguenze sociali, che potrebbero essere viziati da discriminazioni non visibili "a occhio nudo". Come sottolinea Cristianini, se gli umani non riescono a seguire le procedure artificiali, insegnare i principi e i valori umani alle macchine potrebbe non essere scontato[26]. Un concetto importante nella

[23] Mitchell, *Artificial Intelligence*.
[24] Burrell, "How the Machine 'Thinks,'" 1-12.
[25] Cosimi, "Lee Sedol."
[26] Cristianini, *La scorciatoia*, 83-98.

filosofia politica, come quello di accountability, potrebbe essere difficilmente applicabile nel mondo dell'IA. Fino a che punto è possibile rendere conto delle azioni di una IA se le motivazioni alla base dell'esito decisionale sono oscure, inaccessibili?

Un secondo punto è quello dei bias dell'IA, che potrebbero portare a comportamenti imprevisti e dannosi. È noto che i bias possono insinuarsi nei modelli di IA attraverso l'utilizzo di dati "trovati in natura", che dunque riflettono le diseguaglianze esistenti nella società. Ad esempio, alcuni ricercatori hanno scoperto che l'IA può associare certi titoli professionali a concetti con connotazione maschile o femminile sulla base di come vengono presentati nei testi naturali[27]. Parole come "elettricista" o "programmatore" tendono a essere associate a concetti con connotazioni maschili, mentre "assistente" o "nutrizionista" sono legati a concetti con connotazioni femminili. Questo è il risultato di tendenze culturali che vedono certe parole professionali apparire più spesso in associazione con termini maschili o femminili. A ciò si aggiunge un problema più profondo, ovvero il fatto che l'IA è un meccanismo che persegue gli obiettivi assegnatigli in modo indifferente alle conseguenze più ampie. Questo pone un problema particolare quando consideriamo le norme etiche e sociali che guidano le società umane. In molte costituzioni democratiche mature, principi come l'eguaglianza, la non discriminazione e l'accesso paritario sono fondamentali. Tali principi proteggono le persone da discriminazioni basate su età, disabilità, sesso, genere, orientamento sessuale, etnia, religione e credenze. Per garantire il rispetto di questi principi, è essenziale che le decisioni dell'IA non siano influenzate da queste caratteristiche protette. Di primo acchito, una soluzione macchinosa ma percorribile potrebbe essere "scartare" le informazioni protette dagli input offerti all'IA, in maniera tale che quest'ultima non le utilizzi. Tuttavia, la semplice omissione di informazioni sensibili dai dati potrebbe non essere sufficiente. Questo perché le informazioni possono essere implicitamente contenute in altri dati, apparentemente innocui. La combinazione di tali "segnali deboli" all'interno di vasti set di dati e migliaia di parametri può creare impronte statistiche che identificano efficacemente tratti protetti degli individui. Ad esempio, anche se i dati etnici vengono esclusi da un modello di IA destinato alla valutazione finanziaria, l'etnia potrebbe ancora essere dedotta da altre informazioni come il codice di avviamento postale o l'età del matrimonio, e sovrapporsi perfettamente a informazioni legalmente protette senza che nessuno se ne renda conto. I bias non sarebbero mai politicamente neutri e potrebbero essere difficilmente identificabili. Questo potrebbe avere un impatto sulla praticabilità delle teorie etico-politiche circa l'equità o la giustizia.

[27] Caliskan, Bryson e Narayanan, "Semantics," 183-6.

Tre categorie di problemi per la teoria politica

La sorveglianza e l'uso della forza

L'applicazione dell'IA cambia il panorama di diverse questioni al centro della teoria politica. Tre delle maggiori sono il problema della sorveglianza e dell'uso della forza, il problema della qualità della sfera pubblica, il problema della sostituzione intellettuale e della diseguaglianza[28].

Iniziamo con la prima. Le tecnologie cambiano le distribuzioni dei poteri. Ad esempio, nel contesto dell'assedio moderno, l'introduzione della polvere da sparo e la sua applicazione all'artiglieria ha favorito l'attacco, e ha penalizzato i difensori delle mura di cittadine e castelli (tanto da portare all'invenzione delle fortificazioni poligonali, pensate per resistere maggiormente all'artiglieria pesante)[29]. Le tecnologie possono avvantaggiare alcuni attori e penalizzarne altri. Non è ancora chiaro se l'IA possa favorire il potere centrale, quello dei governi e degli stati, oppure quello diffuso della mobilitazione della popolazione. Partendo, però, dal presupposto normativo secondo cui la libertà dei cittadini e la loro capacità di incidere nei processi decisionali collettivi sono un valore, mentre è indesiderabile una concentrazione eccessiva del potere, qualifichiamo come problematica l'eventualità che l'IA possa alimentare la sorveglianza, il controllo ed eventualmente la repressione.

A quasi cinquant'anni dalla pubblicazione del più celebre lavoro di Foucault, *Sorvegliare e punire*[30], forse non esiste descrizione più adeguata di quella foucaultiana per inquadrare i rischi dell'attuale configurazione digitale, in cui centinaia di milioni di cittadini lasciano informazioni su loro stessi. Nelle società moderne, teorizza Foucault, la disciplina si configura come sorveglianza, spesso preventiva, su moltissimi comportamenti degli individui, che in genere sono ben visibili. Il potere adotta le forme della burocrazia ed è nascosto, remoto, anonimo, e anche per questo efficiente nell'attività di sorveglianza. Tutto ciò, in passato, è stato possibile grazie a una tecnologia della comunicazione: la scrittura, la colonna vertebrale del potere moderno (in particolare della burocrazia statale).

---

[28] Questi tre problemi sono esemplificativi e non esaustivi. Ma si tratta di focus classici per la filosofia politica. Uno dei temi emergenti, invece, è quello dell'ambiente, con gli esperti che stanno lavorando affinché l'IA, attraverso l'efficientamento e l'ottimizzazione, diminuisca l'impatto ambientale di molte attività umane (nonostante l'IA sia di per sé energivora). Cfr. Floridi, *Il verde e il blu*. Lavori più specialistici sono: Lyu, W., & Liu, J, "Artificial Intelligence and emerging digital technologies," 117615; Ahmad, Zhang, Huang, Zhang, Dai, Song & Chen, "Artificial intelligence in sustainable energy industry," 125834.

[29] Alfani & Rizzo, *Nella morsa della guerra*.

[30] Foucault, *Sorvegliare e punire*. Interessante a questo proposito Vaccaro, "Tutto il potere agli algoritmi!," in Giacomini, *La politica nel mondo digitale*, 233-54.

È questa tecnica a rendere possibile l'indottrinamento, la registrazione, l'archiviazione[31]. Oggi il punto è che, come sottolinea Ferraris[32], le tecnologie digitali sono formidabili nel registrare (si pensi ai big data). Tanto che, secondo Ferraris, è proprio la gigantesca capacità di registrare qualsiasi attività umana a essere la più peculiare caratteristica dei media digitali, mettendo addirittura in secondo piano le loro funzioni comunicative. È interessante considerare, dunque, il potenziamento che può verificarsi con la tecnologia digitale.

Edward Snowden è un informatico statunitense entrato in possesso, in qualità di *contractor* di agenzie di sicurezza statunitensi come la CIA e la NSA, di documenti riservati su progetti di sorveglianza globale[33]. Secondo questi documenti, i governi statunitense e britannico avrebbero costruito programmi segreti di sorveglianza di massa che, una volta resi noti nel 2013, hanno suscitato un grande scalpore nell'opinione pubblica (è il cosiddetto "Datagate"). Snowden ha fatto pubblicare documenti su programmi di intelligence, tra cui PRISM e Tempora, attraverso la collaborazione con alcuni giornalisti. Il 5 giugno 2013 *The Guardian* ha pubblicato un ordine ad alta segretezza che ingiunge a un ramo dell'azienda di Verizon Communications – un fornitore di banda larga e di telecomunicazioni – di fornire metadati riferiti alle telecomunicazioni domestiche negli Stati Uniti[34]. Sono poi seguiti altri articoli che hanno rivelato l'esistenza di PRISM, un programma di sorveglianza elettronica, *cyberwarfare* e *signal intelligence*, classificato come segreto, usato per la gestione di informazioni raccolte attraverso Internet e altri fornitori di servizi elettronici e telematici, che avrebbe consentito alla NSA di accedere alla posta elettronica, ricerche web, chat audio e video, fotografie e altro traffico in tempo reale. In particolare, la NSA e l'FBI avrebbero attinto informazioni dai server centrali di importanti società Internet e fornitori di servizi digitali[35]. Inoltre, *The Guardian* ha diffuso ulteriori informazioni su Tempora, un programma britannico analogo[36].

Le aziende specializzate in sorveglianza sviluppano prodotti non solo per le agenzie di intelligence ma anche per distretti di polizia. Ad esempio

---

[31] L'intuizione di Foucault è stata poi sviluppata da Deleuze. Nel *Poscritto sulle società di controllo* del 1990, in maniera quasi profetica, Deleuze teorizza il passaggio da sistemi disciplinari basati sull'analogico a sistemi basati sulla cifra (e sull'informatica). Immagina, ad esempio, un meccanismo di controllo che in ogni momento dia la posizione di un elemento in un ambiente aperto (l'attuale GPS). Deleuze, *Poscritto sulle società di controllo*.

[32] Ferraris, *Documanità*.

[33] Cfr. Domscheit-Berg, *Inside Wikileaks*; Chiusi, *Nessun segreto*.

[34] Greenwald, "NSA."

[35] Greenwald e MacAskill, "NSA Prism program taps"; Gellman e Poitras, "British intelligence"; Gellman e Soltani, "NSA infiltrates links."

[36] MacAskill, Borger, Hopkins, Davies e Ball, "GCHQ taps fibre-optic cables."

CellHawk, scrive Richards[37], è un software che sta aiutando i dipartimenti di polizia, l'FBI e gli investigatori privati negli Stati Uniti a convertire le informazioni raccolte dai fornitori di servizi di telefonia mobile in mappe delle posizioni, dei movimenti e delle relazioni delle persone. Il produttore di CellHawk afferma che è in grado di automatizzare con grande velocità processi che in passato richiedevano un lavoro lungo e scrupoloso da parte degli investigatori. Il prodotto basato sul Web può importare record di dettagli delle chiamate, mostrando chi sta parlando con chi. Può anche gestire registrazioni di posizione (i telefoni si connettono a varie torri mentre i loro proprietari si spostano). Si tratta di caratteristiche che fanno sembrare CellHawk più simile a uno strumento per una sorveglianza continua che per un semplice sostegno agli inquirenti. Geofeedia è un altro sistema sotto i riflettori di *The Intercept*, di *The New York Times* e di *Inverse*[38]. Raccogliendo dati dai *social media*, Geoofeedia sembra rispecchiare le paure di Foucault sul potenziale oppressivo della sorveglianza. Si tratta sostanzialmente di un'impresa innovativa specializzata nella raccolta di messaggi di social media con geo-tag, da piattaforme come Facebook, Twitter e Instagram, per monitorare gli eventi (ad esempio manifestazioni pubbliche di protesta) in tempo reale. Traccia la posizione di attivisti e manifestanti e calcola attraverso algoritmi intelligenti "indici di minaccia". L'idea è di valutare il rischio di azioni criminali per prevenzione[39]. Un ulteriore esempio è il software PREDPOL utilizzato negli Stati Uniti, che combina una serie di variabili spazio-temporali per elaborare la probabilità che un crimine possa ripetersi in un luogo determinato[40].

Il dubbio è che queste azioni possano rafforzare forme illegittime di controllo e, in alcuni casi, di discriminazione, dato il problema dei bias dell'IA. La tendenza del potere a essere disciplinare, illuminata da Foucault, potrebbe trovare un'occasione nelle tecnologie digitale per "stringere la morsa" sulla propria cittadinanza, e di rendere ciò accettabile anche grazie all'opacità della sua azione (che implica asimmetria di potere fra controllati e controllori). L'opacità – lo ricordiamo – non riguarda solo le modalità di utilizzo dell'IA da parte dei suoi possessori, ma è una caratteristica propria dell'IA probabilistica. Per presentare la portata di ciò

---

[37] Richards, "Powerful mobile phone surveillance tool."
[38] Lee Fang, "The CIA;" Bromwich, Isaac e Victor, "Police Use Surveillance Tool;" Knefel, "Your Social Media Posts."
[39] Le finalità specifiche possono essere varie: combattere l'evasione e la frode fiscale, esercitare un controllo delle frontiere, informare la polizia di eventuali rischi per la sicurezza. Benbouzid, "Des crimes et des séismes."
[40] Il controllo potenziato da sistemi di IA non viene solo dai poteri pubblici, ma anche da soggetti privati come i datori di lavoro. Si pensi, ad esempio ai sistemi di performance manageriale nelle aziende, oppure alla sorveglianza algoritmica nei posti di lavoro (Rosenblat, *Uberland*).

facciamo un esperimento mentale su un caso futuribile di applicazione dell'IA alle attività di polizia investigativa. Francesco viene convocato in Questura perché, mentre stava facendo una camminata in montagna con la moglie, la moglie cade da un dirupo e muore. Il software di deep learning in adozione alla Polizia, processando una marea di dati su Francesco, la moglie, i loro spostamenti e i loro comportamenti, segnala agli agenti che Francesco, con un'alta probabilità statistica, ha spinto la moglie e ha commesso omicidio. Lui, in realtà, è innocente. Non c'è alcuna prova "causa-effetto" che lui abbia spinto la moglie. Eppure, il software segnala l'omicidio in base ad alcune correlazioni. Magari ha considerato il fatto che Francesco, negli ultimi anni, ha avuto tre amanti. Oppure che ha esitato qualche secondo "di troppo" prima di chiamare i soccorsi. Il profilo caratteriale che emerge dai dati del social network è associabile a quello degli iracondi, e così via. Quali dati ha davvero considerato l'IA, in riferimento a quali altri dati contenuti nel gigantesco archivio, quale peso ha attribuito loro? Come è possibile escludere che non ci siano bias presenti nel processo decisionale della macchina? Assumiamo che spetti ad un umano prendere l'ultima decisione, dunque Francesco è stato scagionato perché non c'è nessuna prova "umana" di omicidio, ma se non ci fosse stata l'IA non sarebbe stato neanche convocato (ha subito comunque un danno). Noi umani in genere riteniamo di avere il "diritto alla spiegazione", ovvero ottenere informazioni comprensibili sulla logica adottata in una qualunque decisone che ci riguarda. Ad oggi, tuttavia, non è ancora ben chiaro come sarebbe possibile "proteggere" gli individui da possibili effetti negativi causati dai bias e dall'opacità dell'IA, soprattutto su questioni che possono avere effetti legali o analogamente rilevanti.

L'AI può essere utilizzata dal potere anche per l'esercizio della forza[41]. Data la storia delle tecnologie digitali, da Internet agli schermi touch nati in ambito militare[42], è lecito aspettarsi che l'IA verrà sviluppata per aggiornare gli arsenali delle maggiori potenze[43]. Una delle evoluzioni prossime potrebbe riguardare l'uso di droni automatici per pattugliare aree specifiche e per attaccare eventuali minacce[44]. Questi droni, piuttosto che essere guidati da un singolo sistema centralizzato di controllo, potrebbero operare con l'IA nelle forme dell'"intelligenza dello sciame". Questo schema si basa su regole decentralizzate dove ogni drone, o membro dello sciame, si muove in relazione ai movimenti degli altri, simile al comportamento di uno stormo di uccelli. L'idea è che, grazie al loro numero con-

[41] Per una panoramica su come l'IA cambia l'uso della forza nel settore militare: De Luca, "Hic sunt drones," 41-56.
[42] Cfr. Mazzucato, *Lo stato innovatore*.
[43] Boulanin e Verbruggen, *Mapping the Development of Autonomy*.
[44] Un testo classico sul tema è Chamayou, *Teoria del drone*.

siderevole e alla capacità di coordinamento decentralizzato, questi sciami possano sopraffare le minacce nemiche[45]. Tuttavia, le armi autonome, una volta attivate, possono selezionare e attaccare un obiettivo anche senza intervento umano. Questa autonomia solleva serie questioni etiche e legali. Uno degli argomenti centrali nel dibattito è se esista la necessità di introdurre nel diritto internazionale, e in particolare nell'ambito del concetto di "guerra giusta"[46], delle restrizioni o delle proibizioni d'uso su alcune o su tutte le armi autonome. A parere di Sharkey, lo sviluppo dell'IA non può offrire alcun supporto all'idea che le armi autonome possano conformarsi alle norme internazionali sulla guerra meglio di quanto riescono a fare gli esseri umani[47]. Secondo fonti giornalistiche, i bombardamenti della Striscia di Gaza da parte di Israele iniziati nell'ottobre del 2023 vengono proposti da un'intelligenza artificiale chiamata Gospel e poi selezionati dai soldati. Il sistema utilizza un vasto database, sviluppato negli ultimi anni dalle divisioni di intelligence israeliane, che comprende dati personali e biometrici dei palestinesi raccolti senza il loro consenso, e non è noto come vengano determinati gli obiettivi considerati attaccabili[48]. Come possiamo essere sicuri di aver "insegnato" alle macchine le procedure corrette, se queste vengono interpretate dall'IA in maniere che a noi possono sfuggire? Inoltre, immaginiamo un contesto di guerra convenzionale e simmetrica. Se la motivazione primaria per l'uso di armi automatiche sarà la loro velocità superiore rispetto all'intervento umano, come può un operatore umano esercitare un controllo efficace? Prendere il tempo per valutare se intervenire su un'arma autonoma in azione potrebbe significare concedere un vantaggio cruciale al nemico.

Sistemi di IA potrebbero essere utilizzati da regimi autoritari anche contro la propria popolazione, nel momento in cui venisse percepita come una minaccia, ad esempio nel caso di proteste o di sollevazioni. Il diritto di ribellarsi a un potere autoritario fa parte della cultura liberale e democratica: già alla fine del Seicento, John Locke teorizzò che, se uno Stato abusa dei suoi cittadini, questi hanno il diritto di ribellarsi[49]. Il problema è che le tecnologie digitali sembrano offrire potenti strumenti agli autocrati minacciati. Le rivoluzioni fallite che si sono svolte nell'era digitale in Paesi come Myanmar, Iran, Egitto, Hong Kong e Bielorussia portano alla luce i grandi sforzi, spesso coronati da successo, dei regimi autoritari di utilizzare le nuove tecnologie per la sorveglianza, l'oppressione, la propaganda, la censura e la soppressione dei diritti fondamentali. Si tratta sia di

[45] Verbruggen, "The Question of Swarms," 1-16.
[46] Walzer, *Guerre giuste e ingiuste*.
[47] Sharkey, "Saying No to Lethal Autonomous Targeting," 369-89.
[48] Carboni, "Come funziona l'intelligenza artificiale."
[49] Locke, *Il secondo trattato sul governo*.

azioni reattive, come staccare Internet per qualche giorno a ridurre la banda delle connessioni, sia proattive, ad esempio identificando grazie all'IA i ribelli, minacciandoli, oppure potenziando la propaganda[50]. Il rischio di una deriva verso il dispotismo dovrebbe spingerci a chiederci quali competenze, regole e istituzioni possano aiutare i cittadini a difendere la propria libertà quando questa è minacciata, anche nell'epoca dell'IA.

La qualità della sfera pubblica

Tornando al contesto delle democrazie di lungo corso, Habermas ha illuminato il ruolo che ha la sfera pubblica per le istituzioni democratiche, dal punto di vista sia fattuale sia ideale. Secondo Habermas, la nascita dell'opinione pubblica, fra il Settecento e l'Ottocento, ha segnato la fine della società corporativa e del regime di privilegi dell'epoca feudale[51]. Al tempo stesso, si è sviluppata la nozione di uguaglianza formale di tutti gli individui di fronte alla legge, promuovendo la crescita dei processi di comunicazione essenziali per la democrazia. A partire dall'analisi delle origini della sfera pubblica, Habermas ha identificato i criteri che, nel suo modello, sono essenziali per realizzare un'ideale situazione democratica. In particolare, il livello normativo consiste in un modello di sovranità popolare intesa come processo discorsivo razionale, dove la volontà generale si esprime nella sfera pubblica come esito di confronti dialogici, grazie ai quali i cittadini hanno potuto maturare un'opinione il più possibile riflessiva[52]. Tuttavia, come si potrebbero realizzare questi standard in presenza dell'applicazione dell'IA ai mezzi di comunicazione?

Da un lato, nel Web dinamico troviamo una maggiore facilità per i soggetti "periferici" a emergere. Sulle app dei social media ognuno può, facilmente e a costi molto bassi, pubblicare, condividere e diffondere idee. Dall'altro lato, il Web dinamico permette a nuovi attori particolarmente influenti (i neointermediari possono essere sia le piattaforme digitali sia aziende di comunicazione[53]) di manipolare "dall'alto", con inediti strumenti, il flusso delle comunicazioni. In Internet il pubblico non è astratto e anonimo ma può essere studiato approfonditamente attraverso la raccolta dei dati personali e l'analisi attraverso l'IA, questo potrebbe abilitare forme diffuse e pervasive di agire strategico (manipolazione e propaganda)[54].

---

[50] Giacomini, *The Arduous Road to Revolution*.
[51] Habermas, *Storia e critica dell'opinione pubblica*.
[52] Habermas, *Teoria dell'agire comunicativo*; Habermas, *Fatti e norme*.
[53] Sul ruolo dei neointermediari si rimanda a Giacomini, "Verso la neointermediazione," 457-68.
[54] Cfr. Giacomini, "Habermas 2.0," 31-50. Si veda anche Maffettone, "Etica pubblica e IA," 179-94.

Emblematico dell'attività di profilazione mirata degli utenti è stato il Caso Cambridge Analytica, scoppiato nella primavera del 2018. È stato rivelato che un ricercatore, sviluppando un'applicazione per raccogliere ed elaborare attraverso un sondaggio le attività online svolte dagli utenti, ha avuto accesso ai profili Facebook di circa 87 milioni di utenti, catturandone i dati e le preferenze. Questo "materiale" è entrato poi in possesso della società inglese Cambridge Analytica che lo ha utilizzato per perfezionare le strategie comunicative di alcune campagne elettorali (come quella del candidato repubblicano alla presidenza americana Trump nel 2016). La tecnica di profilazione utilizzata, denominata "psicografica", nasce dalla combinazione fra *big data* e IA. La psicografia serve per descrivere i tratti degli esseri umani[55]. L'elaborazione, con il sostegno dell'IA, permette di definire il grado di apertura, coscienziosità, estroversione, disponibilità, nevrosi dei soggetti. Dopodiché gli psicologi determinano da che cosa gli individui sono motivati e spinti ad agire. A quel punto il *team* creativo, specializzato in comunicazione, confeziona messaggi specifici (video, audio, immagini), ideati appositamente per certi tipi di personalità, attraverso il procedimento del *micro-targeting* comportamentale[56].

Le basi per le previsioni comportamentali sono modelli di personalità come il Big Five model, il DISC e il Myers-Briggs Type Indicator. Probabilmente il Big Five è il modello più diffuso attualmente. McCrae e Costa hanno postulato cinque grandi dimensioni di personalità: estroversione-introversione, gradevolezza-sgradevolezza, coscienziosità-negligenza, nevroticismo-stabilità emotiva, apertura mentale-chiusura mentale[57]. Una persona che è circondata di amici farà registrare un alto grado di estroversione, chi tende a pianificare la propria giornata avrà un punteggio alto in termini di coscienziosità. In ambito digitale, ad esempio, gli utenti con elevato grado di apertura mentale tendono a mettere "mi piace" a quadri di Dalì o alle conferenze TED[58]. Questo ancoraggio alle azioni quotidiane rende il metodo applicabile ai contesti digitali, in quanto le associazioni fra comportamenti rilevati e tratti comportamentali sono facili da individuare (anche dall'intelligenza artificiale). Grazie al digitale, la valutazione basata sul Big Five model è molto potente, dato che può contare su ampie disponibilità di informazioni e grandi capacità di calcolo. Oltre alla disponibilità di dati, infatti, serve la capacità di interpretazione, e questa è

[55] Per una rassegna sulle applicazioni della psicografia si veda Wells, "Psychographics," 196-213.
[56] Sulla procedura di costruzione dei messaggi, che vede collaborare da vicino informatici e psicologi, si veda Kaiser, *Targeted*.
[57] Cfr. McCrae, Costa, "Validation," 81-90.
[58] Cfr. Youyou, Kosinski, Stillwell, "Computer-based personality judgments,"1036-40.

garantita dai sistemi di IA: gli operatori umani danno una direzione operativa e l'IA "digerisce" i big data.

Già nel 2013, uno studio aveva rilevato che i 'mi piace' di Facebook potevano stimare in modo automatico una vasta gamma di attributi personali generalmente ritenuti privati, come orientamento sessuale, etnia, convinzioni politiche e religiose, carattere, intelligenza, felicità, divorzio dei genitori, addirittura l'uso di sostanze che generano dipendenze[59]. Sembra attualmente possibile entrare negli "abissi comportamentali" più intimi[60]. Queste potenzialità applicate al mondo politico si trasformano in ciò che Soro[61] ha definito "pedinamento degli elettori". Come scrive Byung-Chul Han, «ogni passo nella rete è osservato e registrato. La nostra vita si riflette completamente nella rete digitale. Le nostre abitudini digitali offrono una copia esatta della nostra persona, del nostro animo, forse persino più precisa o completa dell'immagine che abbiamo di noi stessi […]. Nelle campagne elettorali statunitensi, big data e data-mining si dimostrano nei fatti l'uovo di Colombo. I candidati dispongono di uno sguardo a 360 gradi sugli elettori: da fonti diverse vengono raccolte, anzi comprate, immense masse di dati, poi connesse fra loro in modo da produrre dei profili estremamente precisi degli elettori. Si ricorre al micro-targeting per rivolgersi ai votanti in modo mirato, con messaggi personalizzati, per influenzarli»[62].

Come sottolinea Han, non si tratta tanto di un semplice servizio, quanto di una manipolazione, operata spesso senza che gli individui siano consapevoli. Il divario di potere fra un normale cittadino e un'organizzazione che utilizza l'IA può essere notevole, in quanto un agente di IA impara dal comportamento di miliardi di utenti, ha accesso a numerose informazioni personali su ognuno di loro e può scegliere i suoi messaggi da un catalogo quasi infinito. Gli obiettivi di queste aziende possono non coincidere con quelli degli utenti[63]. La propaganda non è di certo una novità nella storia (si può sostenere che la filosofia platonica e aristotelica, intesa come episteme opposta alla doxa, sia nata in funzione antisofistica, ovvero con l'obiettivo di contrastare derive relativistiche e retoriche nella polis greca[64]). Tuttavia, l'IA sembra essere in grado di potenziare le capa-

[59] Cfr. Kosinski, Stillwell e Graepel, "Private traits and attributes," 5802-5.
[60] Alcuni studi che mostrano la capacità di penetrare in aspetti intimi sono: Chen, Tsai e Chen, "A user's personality prediction approach," 913-37; Tandera, Suhartono, Wongso e Prasetio, "Personality prediction system," 604-11.
[61] Soro, *Persone in rete*.
[62] Han, *Psicopolitica*, 75-77.
[63] Burr, Cristianini & Ladyman, "An Analysis," 735-74.
[64] Si tratta di una lettura standard nella storia della filosofia, addirittura manualistica (cfr. Bonazzi, *I sofisti* e Erler, *Platone*).

città manipolatorie, aumentando l'asimmetria fra chi utilizza queste tecnologie e chi le subisce. Per non contare che oggi, grazie all'IA applicata alla grafica, è possibile creare immagini e video indistinguibili da quelli reali. Negli ultimi anni si è parlato, anche nella cronaca, di *fake news*: si tratta di notizie false divulgate su Internet con lo scopo deliberato di ingannare, grazie anche all'IA che "anima" i bot automatici e che contribuisce a verificare quali contenuti (magari emotivi, sensazionalistici, cospiratori) sono più efficaci in base al target.

Infine, sembra che la strutturazione dell'ambiente che le grandi piattaforme costruiscono attraverso il ricorso ai big data e all'IA rischi di limitare la diversità di opinioni presenti nella sfera pubblica online, portando a una maggiore polarizzazione e alla formazione di "bolle" informative. Profilare i gusti degli utenti e personalizzare l'offerta di contenuti è alla base del business di molte big tech. In particolare, gli algoritmi di raccomandazione delle piattaforme hanno il compito di studiare i gusti degli utenti, e di proporre loro contenuti apprezzati, che spingano le persone a rimanere connesse (e quindi ad aumentare i profitti delle aziende digitali). Questi sistemi, apprendendo automaticamente, tendono a perseguire qualsiasi misura di "coinvolgimento" senza una comprensione dei potenziali effetti collaterali sugli individui o sulla qualità dell'opinione pubblica. Questo potrebbe favorire la creazione di *filter bubbles* (gli individui si trovano in bolle in cui si ricevono soprattutto informazioni che confermano ciò che credono[65]) o di *echo chambers* (metafora di un ambiente chiuso che riflette e rafforza ciò che si trova già al suo interno[66]). Fino a quando siamo in una bolla che ci suggerisce i viaggi che desideriamo fare, o le scarpe che ci piacciono di più, ciò è soddisfacente. Ma potrebbe diventare un problema nel caso in cui si trattasse di opinioni politiche, disabituando al fatto (e al valore democratico) del pluralismo[67]. Questo tipo di dinamiche potrebbero avere conseguenze politiche generali, poiché tendono a "chiudere le comunità"[68].

Ciò potrebbe essere un problema per la democrazia, la quale dal punto di vista normativo necessità di una conciliazione fra posizioni diverse, in maniera tale da continuare nella cooperazione sociale. Autori come Rawls e Arendt hanno proposto modelli politici per conciliare la libertà e la pluralità di idee, stili di vita e valori con la stabilità e la

---

[65] L'ideatore del concetto è Parisier, *The Filter Bubble*.
[66] Ne parla Sunstein, *#Republic*.
[67] Arendt e Rawls, fra i molti, hanno sottolineato l'importanza di un pluralismo "ragionevole" e non "estremista", per la democrazia moderna. Cfr. Arendt, *The human condition*; Rawls, *Political liberalism*.
[68] Del Vicario, Bessi, Zollo, Petroni, Scala, Caldarelli e Quattociocchi, "The spreading of misinformation on line," 554-9.

giustificazione dei sistemi democratici liberali. Si pensi al concetto di "ragionevolezza" in Rawls[69] (le persone hanno differenti dottrine "comprensive", ma al tempo stesso devono essere ragionevoli, mantenere una postura reciprocamente aperta e disponibile, al fine di convergere su un sotto-insieme di valori politici e continuare, così, nella cooperazione sociale) o a quello di "in-between" di Arendt[70] (ogni individuo dovrebbe coltivare la libertà, sia di azione sia di pensiero, per abitare pienamente lo spazio della politica, ovvero quello spazio intermedio che resta tra i soggetti, e che permette loro di confrontarsi al netto di appartenenze, affiliazioni, credenze). Al tempo dell'IA e delle sue capacità profilatorie e personalizzanti, ci si interroga se il rischio non sia quello di aumentare le bolle nella sfera pubblica democratica[71], aggravando i conflitti sociali e politici che sono già in corso.

La sostituzione intellettuale e la diseguaglianza

La democrazia non si basa solo su aspetti formali, ma anche materiali. La dimensione del lavoro è centrale, e potrebbe essere impattata a breve dall'IA. Finora siamo abituati al fatto che le tecnologie sostituiscono soprattutto il lavoro manuale, routinario, quello con tanta fatica e poco valore aggiunto[72]. La storia è costellata da innovazioni tecniche che hanno sollevato l'essere umano dal fardello della fatica meccanica, operativa (si pensi all'aratro, all'aspirapolvere, al trattore, alla leva, alla ruota, ai camion). L'IA potrebbe portare a una rottura di questa tendenza, andando a sostituire quote di lavoro intellettuale, creativo. Sul lato calcolistico-matematico la "sostituzione intellettuale" [73] da parte delle macchine è quasi completamente avvenuta. Sul lato creativo l'IA sta avanzando in maniera significativa[74]. Le macchine potenziano le capacità di alcuni lavoratori e al tempo stesso sollevano gli esseri umani da mansioni che non necessiteranno più dell'intelletto naturale.

---

[69] Cfr. Rawls, *Political liberalism*. Per una interpretazione della ragionevolezza al tempo del digitale: Sala, "Ragionevolezza e irragionevolezza nell'uso del web", 207-20.
[70] Cfr. Arendt, *The human condition*.
[71] Damiano Palano parla di "bubble democracy" (Palano, *Bubble democracy*).
[72] Moretti, *La nuova geografia del lavoro*.
[73] Per "sostituzione" si intende che una quota rilevante di lavoratori umani, impiegati in una specifica attività, vengono sostituti progressivamente da macchine.
[74] Come puntualizza Casilli, in realtà l'output prodotto dall'IA si basa a sua volta su attività umane sottopagate nel cosiddetto Sud globale, in luoghi come le click farm, dove vengono addestrati i sistemi intelligenti stessi. Ciò, tuttavia ha effetti diversi nelle società dell'Occidente, con l'IA che va a potenziare, affiancare e in parte sostituire molte attività tradizionalmente da "colletti bianchi". Cfr. Casilli, *Schiavi del clic*; Brynjolfsson e McAfee, *The second machine age*.

Già oggigiorno, oltre alle raccomandazioni di prodotto personalizzate e al marketing, l'IA si occupa di assistenza clienti automatizzata e di controllo della qualità, per non parlare della gestione del personale, della logistica e della catena di approvvigionamento[75]. L'IA sta avendo un impatto in aree come la legge (con programmi che possono effettuare ricerche legali o analizzare contratti), la finanza (con algoritmi che possono analizzare grandi quantità di dati molto più velocemente di quanto possano fare gli esseri umani) e la medicina (con sistemi di diagnosi assistita da computer). Un settore professionale ad alto rischio di "sostituzione intellettuale" è quello delle traduzioni. Per esemplificare, DeepL è un recente servizio di traduzione automatica basata su reti neurali profonde. Le traduzioni non sono perfette, ma testi descrittivi, non troppo complessi, vengono trasposti in maniera eccellente, in decimi di secondo, a costo praticamente nullo. Prossimamente, invece di richiedere agli umani di eseguire traduzioni, è possibile che gli umani vengano impiegati principalmente per revisionare e controllare le traduzioni effettuate automaticamente. Ciò che in passato poteva essere fatto da cento specialisti, nei prossimi anni potrebbe richiederne dieci "potenziati" dall'IA. L'evoluzione dell'IA potrebbe addirittura erodere i lavori creati dalla stessa rivoluzione informatica: i programmatori. L'IA sembra sempre più capace di generare codice autonomamente. Immaginiamo ChatGPT unito a un software di programmazione di siti Web, a sua volta collegato a un'IA specializzata in grafica e design (come DallE). In chat si potrebbe chiedere all'IA di proporre un layout del proprio sito Web, e poi si potrebbero chiedere delle modiche semplicemente interloquendo. L'IA potrebbe sottoporre all'umano alcune proposte di logo per il nuovo sito, noi potremmo sceglierne una e anche offrire dei suggerimenti migliorativi. L'IA potrebbe creare contenuti per il sito generando testo, immagini e video automaticamente in base alle nostre indicazioni[76].

Nel mirino dell'innovazione, per la prima volta in maniera così diretta, ci sono i mestieri intellettuali, creativi e anche artistici (scrittori, musicisti, grafici). Una obiezione a questo potrebbe essere che, ad esempio,

---

[75] In piattaforme come Deliveroo o Just Eat l'IA è fondamentale per dirigere i rider, per raccogliere le ordinazioni, organizzare informazioni e comunicazioni (mentre il maggior numero degli esseri umani impegnati nell'ambito largo dell'organizzazione stanno ai fornelli a cucinare o pedalano sotto la pioggia e sotto il sole).

[76] A chi potrebbero andare i diritti di loghi, immagini o grafiche costruire dall'IA, ma che nascono dalla rielaborazione di lavoro umano? All'azienda proprietaria dell'IA? All'operatore umano che avanza le richieste all'IA? O alla comunità dei creativi? Questa comunità potrebbe essere definita in che modo? Questo problema sta già emergendo. Ad inizio 2023 alcuni tra i maggiori disegnatori italiani hanno sottoscritto un manifesto per chiedere all'Unione europea la tutela delle loro opere "saccheggiate" dagli algoritmi generativi, i quali elaborano le immagini trovate sul web per produrre nuovi contenuti. Per non parlare dei problemi, nei casi di immagini e video, legati all'indistinguibilità tra verità e fantasia/menzogna.

anche i fotografi "tradizionali", pochi decenni fa, hanno subito l'innovazione digitale. Tuttavia, in passato si trattava soprattutto dell'aspetto pratico e operativo di avere un numero limitato di scatti, di dover sviluppare le pellicole e così via. La differenza è che ora si tratta proprio di una sostituzione del lavoro artistico, dell'atto creativo "in sé" di produrre belle immagini. Un altro esempio riguarda la professione giornalistica (sistemi di IA come i *large language models* sono specializzati proprio nella scrittura). L'attività tradizionale che utilizzava la macchina da scrivere è stata erosa (in termini di numero di addetti) dai programmi di videoscrittura, e poi dai social. Ma la differenza è che ora non si tratterebbe di facilitazioni "tecniche" alla scrittura (quando si è in missione non bisogna più dettare al collega, via telefono, l'articolo, ma lo si può scrivere direttamente sullo smartphone), ma della scrittura vera e propria, dell'atto intellettuale in quanto tale. Naturalmente, gli umani dovranno gestire l'IA sia in entrata (dovremo essere esperti nel dare i comandi e gli spunti giusti) sia in uscita (gli output vanno controllati, in quanto contengono errori, inesattezze, addirittura discriminazioni), e ciò porterà alla creazione di nuove professioni. Resta il punto di quante persone occuperanno queste professioni.

Il rischio è la creazione di nuove *rust belt* con annessi disagi sociali e reazioni estremiste. "Rust belt" è un termine usato negli Stati Uniti per descrivere una regione industriale, principalmente situata nel nord-est e nel Midwest, che a causa di fattori come la globalizzazione e l'automazione ha visto un declino significativo nella sua base manifatturiera e nei posti di lavoro nell'industria a partire dagli anni '70. Quando pensiamo a una potenziale *rust belt* nel settore dei colletti bianchi e dei lavori intellettuali, intendiamo che l'IA potrebbe causare una riduzione significativa dei posti di lavoro in settori tradizionalmente associati a lavori mediamente o altamente qualificati e ben remunerati. Esistono preoccupazioni legittime che, man mano che l'IA e altre tecnologie continuano a svilupparsi, ci potrebbero essere meno opportunità per i lavoratori in questi settori, portando a situazioni simili a quelle osservate nella *rust belt* del recente passato. Mentre l'IA aumenta la disoccupazione (o la sottoccupazione) per grandi masse di lavoratori, potrebbe accrescere la domanda di specifiche (e magari numericamente limitate) professionalità ad alta qualifica legate alla tecnologia. Ciò potrebbe portare a una polarizzazione del mercato del lavoro, con un crescente divario salariale tra diverse professioni. Inoltre, potrebbe verificarsi una concentrazione di capitale, in quanto le aziende che sviluppano e controllano tecnologie basate sull'IA possono essere poche e possono accumulare enormi profitti[77]. Approcci neomarxiani[78] sottolineano

[77] Piketty, *Le Capital au XXIe siècle*.
[78] Si veda Casilli, *Schiavi del clic*.

non solo che la *gig economy* è una forma moderna di sfruttamento, ma che il valore dei contenuti prodotti dagli utenti sulle piattaforme supera di gran lunga la compensazione che ricevono (e costruisce il potere delle piattaforme). Inoltre, oggi i modelli di maggior successo di IA sono sviluppati da grandi aziende private e utilizzano risorse economiche enormi, difficilmente sostenibili da imprese "normali" o anche da molti Stati, il che rafforza ancora di più la posizione delle poche big tech del settore.

Il tema delle diseguaglianze connesse all'innovazione digitale (e a fenomeni come i "lavori *gig*" e i "micro-lavori"[79]) è legato alla stabilità della democrazia. In particolare, la benzina del cosiddetto "populismo" sembrano essere la rabbia e il risentimento, comprensibili alla luce delle crescenti disuguaglianze nelle società occidentali (i populismi hanno in comune il fatto di esprimere il disagio di coloro che hanno la percezione di essere "periferia")[80]. In una situazione difficile per miriadi di persone, il successo delle grandi aziende digitali si fonda sull'efficiente capacità di far incontrare diverse categorie di utenti, anche grazie all'apprendimento automatico (si pensi alla centralità degli algoritmi di raccomandazione per le piattaforme commerciali). Questa capacità è alimentata proprio dai dati offerti dagli utenti: però l'umanità, il più delle volte, lavora gratuitamente per le piattaforme. Come sostiene Ferraris[81], ogni registrazione è capitalizzazione attuale o potenziale, dunque valore. In cambio della gratuità dei servizi, le persone cedono i loro dati e su questi ultimi – al netto del costo dell'erogazione del servizio – le piattaforme si arricchiscono enormemente: è un "plusvalore documediale" (il debito alla tradizione marxiana è ancora evidente). La rivoluzione informatica ha trasformato anche l'ozio in attività economicamente spendibili. Ma, se l'input nel processo economico è costituito dall'attività di centinaia di milioni, se non miliardi, di persone, ciò andrebbe riconosciuto. Per questo Ferraris sottolinea l'importanza (e il senso) che avrebbe ridistribuire il "plusvalore documediale" attraverso una tassazione che generi un *webfare*, ovvero un welfare digitale per far fronte ai problemi generati dall'automazione e dalla sua velocità. In particolare, Ferraris ritiene che ci si debba focalizzare non tanto sulla mera redistribuzione di reddito, quanto su una nuova protezione sociale che metta al centro l'educazione e la cultura.

Questa idea ci ricollega al fatto che l'innovazione deve essere gestita politicamente, secondo ideali di equità e di giustizia. Se gli effetti fosse-

[79] Cfr. Woodcock e Graham, *The gig economy*.

[80] Interessante, a questo proposito, un numero monografico della rivista *Stato e mercato*. In particolare: Diamanti, "Alla periferia della crisi," 117-26 e Reyneri, "Le basi sociali del populismo," 141-8.

[81] Cfr. Ferraris, *Documanità*. Un testo successivo è Ferraris, "Dalla tirannia del merito alla democrazia del bisogno," 83-102.

ro redistribuiti e se ci fosse sostegno per i lavoratori in difficoltà, la sostituzione intellettuale potrebbe essere una opportunità di sviluppo sociale. Uno dei maggiori economisti del Novecento, Keynes, aveva sognato per i suoi nipoti la liberazione dai lavori più duri, faticosi e ripetitivi[82]. Un secolo dopo, possiamo sognare la liberazione dallo stress mentale, dai lavori sedentari, dalla frustrazione delle attività amministrative.

Alcune note per la riflessione etico-politica che verrà

Davanti ai rischi del futuro, a rassicurarci c'è l'idea che si possa "staccare la spina" all'IA. Potrebbe essere una speranza non ben riposta. "Staccare la spina" potrebbe avere costi enormi. La motivazione non dipende soltanto dalle straordinarie potenzialità dell'IA, a cui si dovrebbe rinunciare almeno in alcuni settori (quelli meno sicuri o più rischiosi). Ma anche dal fatto che, con il passare del tempo, la società umana sarà probabilmente sempre più dipendente dell'IA, e tornare indietro potrebbe avere dei costi molto alti da affrontare, come oggi sarebbe difficile in una situazione emergenziale affrontare un repentino calo della disponibilità dei combustibili fossili. I costi delle conseguenze in termini di povertà diffusa, crisi dei servizi, fragilità del sistema socioeconomico rischierebbero di essere superiori ai problemi per cui si pensa di "staccare" l'IA.

Questo significa che la politica ha la responsabilità di governare l'IA e i suoi effetti. Molti ricercatori stanno lavorando all'identificazione dei principi che dovrebbero ispirare l'IA e la sua regolamentazione. Fra i gruppi di lavoro recenti si segnalano i "Principi di Asilomar per l'IA" (2017), la "Dichiarazione di Montreal per l'IA responsabile" (2017), la "Dichiarazione su intelligenza artificiale, robotica e sistemi autonomi" (2018), i "Cinque principi generali per un codice di intelligenza artificiale" (2018), le "Linee guida etiche per l'IA affidabile" (2019). Floridi, forse il più importante esperto di etica dell'informazione al mondo, ha tentato una sintesi dei principi ricorrenti e trasversali nei documenti prodotti negli ultimi anni. Essi sono:

1) Beneficenza (l'IA deve promuovere il benessere, preservare la dignità e sostenere il pianeta).
2) Non maleficenza (l'IA deve rispettare la privacy, essere sicura ed evitare usi impropri).
3) Autonomia (l'IA deve promuovere l'autonomia degli esseri umani, che possono sempre scegliere come e se delegare le decisioni alla macchina).

[82] Keynes, *Democrazia e mercato*. Una lettura di Keynes applicata al digitale è Floridi, "Perché nel mondo digitale abbiamo bisogno di un progetto umano," 103-14.

4) Giustizia (l'IA deve sostenere la prosperità dei popoli, preservare la solidarietà ed evitare l'iniquità).
5) Esplicabilità (l'IA dovrebbe essere il più possibile trasparente e intelligibile).

Immaginiamo, in base a questi principi, come potrebbe essere "costruito" un sistema di IA nel campo della sorveglianza. Innanzitutto, per quanto riguarda la beneficenza, l'IA dovrebbe lavorare per identificare minacce reali e tangibili, piuttosto che monitorare le persone in modo indiscriminato. Per non creare danno, il sistema dovrebbe evitare usi impropri, ad esempio proteggendo i dati contro tentativi di manomissione, e non dovrebbe conservare informazioni non pertinenti o eccedenti rispetto alle sue funzioni di sicurezza. Poi dovrebbe difendere l'autonomia delle persone garantendo la possibilità per gli individui di decidere se vogliono essere monitorati. Ad esempio, in un contesto urbano, dovrebbero esistere zone in cui la sorveglianza è chiaramente indicata, dando alle persone la possibilità di evitarla. L'IA, per essere giusta, dovrebbe come minimo evitare di discriminare o mostrare pregiudizi basati su etnia, genere, classe sociale o altre categorie, ad esempio monitorando in maniera più stringente persone di colore. Infine, promuovere l'esplicabilità potrebbe significare fornire registri e report dettagliati delle sue attività di monitoraggio, rendendo chiaro come e perché ha preso certe decisioni.

Abbiamo poi visto che i sistemi di raccomandazione possono "involontariamente" amplificare l'effetto *echo chambers*, se non controllati correttamente. Un sistema di IA etico nell'ambito della comunicazione dovrebbe probabilmente ridurre questa tendenza, consigliando anche contenuti non allineati ai gusti degli utenti, e che siano istruttivi e arricchenti, evitando messaggi divisivi o, peggio, diffamatori e violenti. Inoltre, il sistema non dovrebbe essere "aggressivo" nell'utilizzo dei dati, il che significa che dovrebbe evitare di raccogliere informazioni eccessive e garantire la possibilità di rimozione o modifica[83]. Inoltre, sempre per quanto riguarda la non maleficenza, l'IA dovrebbe evitare di manipolare gli utenti per scopi commerciali o politici, ad esempio evitando la promozione eccessiva di contenuti sponsorizzati senza una chiara indicazione. Per proteggere l'autonomia degli utenti bisognerebbe permettere loro di decidere in che misura desiderano che il sistema di raccomandazione automatica personalizzi i contenuti per loro (ad esempio, gli utenti potrebbero scegliere di ricevere una gamma più ampia di contenuti per evitare *echo chambers*). Ciò sarebbe collegato anche con la dimensione della giustizia e della non discriminazione, nel senso che promuovere contenuti di creatori di diver-

[83] L'UE, tramite il GDPR, ha cominciato a muoversi con una regolamentazione in questa direzione. A proposito: Dunn e Durante, "Disinformazione e Internet," 57-82.

se parti del mondo o di differenti contesti socioeconomici, per garantire una rappresentanza equa, significherebbe non essere prevenuti verso particolari gruppi di persone basandosi su caratteristiche come l'etnia, il genere o la religione. La trasparenza potrebbe essere perseguita permettendo agli utenti di capire perché un certo contenuto è stato raccomandato (ad esempio, gli utenti potrebbero avere accesso a un pannello di controllo che mostra come le loro interazioni influenzano le raccomandazioni).

Infine, la questione economica e del lavoro. In questo campo, per quanto riguarda la beneficenza, l'IA dovrebbe essere utilizzata in modo complementare ai lavoratori umani, potenziando le loro capacità piuttosto che sostituendole. Bisognerebbe evitare di "deumanizzare" settori professionali attraverso la completa automazione, in quanto la dimensione umana ha un valore redistributivo e di benessere personale e sociale. A specchio, "non maleficenza" significherebbe non ricorrere all'IA per abbassare deliberatamente i salari. L'autonomia potrebbe essere favorita attraverso la formazione, pensata per aiutare i lavoratori a collaborare efficacemente con l'IA e adattarsi ai nuovi ruoli. Qui emerge l'importanza del contesto politico: avere una voce attiva nelle decisioni relative all'implementazione dell'IA, attraverso la partecipazione partitica e sindacale, andrebbe sicuramente nella direzione dell'autonomia. Anche la giustizia è una dimensione politica. Ad esempio, bisognerebbe investire nei programmi di formazione e riqualificazione per aiutare i lavoratori a transitare in nuovi ruoli se i loro lavori attuali sono a rischio, e promuovere nuove forme di welfare (salario minimo, reddito minimo) in modo da contrastare la crescita delle diseguaglianze.

Al netto di queste proiezioni, due elementi segnalano il fatto che siamo solo all'inizio di una riflessione etico-politica sull'IA. Il primo è che i primi quattro principi sono talmente generici da poter essere considerati poco più che auspici facilmente condivisibili, e che non è chiaro in che modo promuovere concretamente questi principi (è sufficiente l'autoregolamentazione e la tutela dei consumatori? Oppure lo Stato deve intervenire in maniera più diretta e incisiva?). Il secondo è che il principio dell'esplicabilità è corretto dal punto di vista normativo, ma non si capisce ancora fino a che punto possa essere applicabile. Floridi sottolinea che l'oscurità dell'IA non deve diventare un alibi per l'inazione, non deve, cioè, fornire una scusa alle aziende digitali per nascondere le loro procedure interne ai ricercatori, agli organismi di controllo e in generale alle istituzioni democratiche. Floridi ha pienamente ragione, ma sembra dimenticare il fatto che l'IA sia una *black box* anche per gli informatici che l'hanno costruita. Dal momento in cui sono state abbandonate le regole formali a favore di complessissime relazioni statistiche, potrebbe essere difficile spiegare i motivi dietro una decisione di una macchina. Eppure, l'ultimo principio

è il più basilare. Potremmo sostenere che si tratta di un pre-principio, nel senso che soltanto se l'IA è esplicabile possiamo controllare fino in fondo che sia benefica, non malefica, che rispetti l'autonomia degli umani e che promuova la giustizia sociale. Per questo Cristianini, che è un informatico, sottolinea che si stanno sperimentando dei modi per migliorare la verifica dei sistemi di IA, come l'uso di "stress test" o la creazione di "checkpoint" interni in cui si possano fare alcuni controlli[84]. Tecnicamente, questo nuovo campo di studio si chiama "explainable AI (XAI)", e mira a migliorare la fiducia e la trasparenza dei sistemi basati sull'intelligenza artificiale, affinché quest'ultima continui a compiere progressi e il più possibile sicuri. Da questi progetti dipenderà il grado e di controllo con cui potremo addestrare eticamente un'IA.

Sappiamo che l'IA, oltre a essere opaca, può avere dei bias, e prendere decisioni indesiderate. Una domanda che probabilmente sarà al centro della riflessione etico-politica nei prossimi decenni è: chi dovrebbe essere ritenuto responsabile di una azione di un sistema di IA che causa dei danni? Il sistema può anche essere definito intelligente (secondo la definizione *indipendente* di intelligenza) ma può difficilmente essere considerato un agente morale, almeno nel senso tradizionale del termine (anche una colonia di formiche si comporta in maniera intelligente, ma tendiamo a non ritenerla responsabile, almeno nel senso pieno del termine, per aver costruito il formicaio nella nostra casa). In linea teorica, diventa importante che ci sia un "controllo umano significativo". Per significativo si intende che sono escluse condizioni e forme di controllo dell'essere umano sulla macchina puramente nominali. Il controllo deve essere invece sostanziale, ovvero l'agente umano deve essere in grado di esprimere un giudizio ponderato sulle operazioni che il sistema sta compiendo, e di poter intervenire in tempo utile in caso di imprevisti. Inoltre, l'operatore deve essere addestrato a non sovrastimare le capacità dei sistemi informatici. Tuttavia, in pratica, si tratta di un compito complesso, forse impossibile da realizzare pienamente, a causa dell'opacità interna riguardante i processi di calcolo complessi e la difficoltà di interpretare le informazioni che inducono il sistema artificiale a prendere una certa decisione o compiere un'azione. L'operatore che ha attivato il sistema potrebbe non essere in grado di esercitare un controllo efficace sul suo comportamento, sia a causa della difficoltà cognitiva di comprendere i meccanismi decisionali dell'IA (come detto, l'IA è una *black box*) sia per via dei tempi lunghi di reazione degli esseri umani che potrebbero rilevarsi inefficaci per un

[84] Alcuni articoli che presentano il tema della XAI sono: Adadi e Berrada, "Peeking inside the black-box," 52138-52160; Došilović, Brčić e Hlupić, "Explainable artificial intelligence," 0210-0215; Vilone e Longo, "Notions of explainability and evaluation approaches," 89-106.

intervento tempestivo (effetto "buoi scappati"). Ma soprattutto, chi sono i "veri" responsabili? Potrebbero essere: gli ingegneri informatici che hanno progettato il sistema, i responsabili dell'azienda produttrice, i consulenti o i dipendenti dell'organizzazione che ha acquistato il sistema, i responsabili dell'organizzazione che ha acquistato il sistema, il capo dell'ufficio specifico che ha utilizzato il sistema, il lavoratore singolo che monitorava direttamente il sistema. Il rischio, non così remoto, è quello di concludere che nessuno ha dato un contributo davvero significativo al verificarsi del danno[85], oppure si rischia di attribuire per legge questa responsabilità diffusa a una singola figura, magari retribuita proprio per accollarsi legalmente le eventuali responsabilità negative. Si potrebbe oscillare, in sintesi, fra l'effetto "tutti colpevoli, nessun colpevole" e l'effetto "capro espiatorio".

La riflessione etico-politica sull'IA si trova ancora nelle sue fasi iniziali. Si tratterà probabilmente di trovare un equilibrio tra l'innovazione tecnologica e la protezione dei valori espressi dai diritti umani, facendo in modo che questa potente tecnologia possa migliorare il benessere umano e limitando il più possibile gli effetti indesiderati. Per realizzare questo ambizioso obiettivo sarà importante, in primo luogo, essere consapevoli delle caratteristiche dell'IA e delle sfide che pone alla gestione da parte degli esseri umani. In secondo luogo, si tratta di avviare una collaborazione sempre più stretta fra umanisti, scienziati sociali, giuristi ed informatici, in maniera tale da eliminare (se possibile), o ridurre al massimo, i problemi dell'opacità e dei bias. In terzo luogo, probabilmente bisognerà esplorare rinnovati modelli normativi, e avviare una discussione su quali possono essere gli approcci etico-politici più desiderabili in un mondo sempre più caratterizzato dalla presenza dell'IA.

Bibliografia

Ahmad, Tanveer, Zhang, Dongdong, Huang, Chao, Zhang, Hongcai, Dai, Ningyi, Song, Yonghua & Huanxin Chen. "Artificial intelligence in sustainable energy industry: Status Quo, challenges and opportunities." *Journal of Cleaner Production* 289 (2021): 125834. https://doi.org/10.1016/j.jclepro.2021.125834

Albus, James. "Outline for a theory of intelligence." *IEEE Transactions on Systems, Man, and Cybernetics* 21, no. 3 (1991): 473-509. https://doi.org/10.1109/21.97471

Alfani, Guido & Mario Rizzo. *Nella morsa della guerra*. Milano: FrancoAngeli, 2013.

[85] Cfr. Nissenbaum, "Accountability," 25-42.

Arendt, Hannah. *The human condition*. Chicago: University of Chicago Press, 1998.

Benbouzid, Bilel. "Des crimes et des séismes. La policie prédictive entre sciences, techniques et divination." *Réseaux* 6, no. 206 (2017): 95-123. https://doi.org/10.3917/res.206.0095

Bonazzi, Mauro. *I sofisti*. Roma: Carocci, 2010.

Bostrom, Nick. *Superintelligenza. Tendenze, pericoli, strategie*. Torino: Bollati Boringhieri, 2014.

Boulanin, Vincent & Maaike Verbruggen. *Mapping the Development of Autonomy in Weapon Systems*. Stoccolma: SIPRI Report, 2017.

Bromwich, Jonah Engel, Victor, Daniel e Mike Isaac. "Police use surveillance tool to scan social media, A.C.L.U. says." *The New York Times*. 11 ottobre 2016.

Brynjolfsson, Erik & Andrew McAfee. *The second machine age*. New York: Norton & Company, 2014.

Burr, Christopher, Cristianini, Nello, & James Ladyman. "An Analysis of the Interaction Between Intelligent Software Agents and Human Users." *Minds & Machines,* 28 (2018): 735-74. https://doi.org/10.1007/s11023-018-9479-0

Burrell, Jenna. "How the Machine 'Thinks'. Understanding Opacity in Machine Learning Algorithms". *Big Data & Society* 3, no. 1 (2016): 1-12. https://doi.org/10.1177/205395171562251

Caliskan, Aylin, Bryson, Joanna J., & Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356, no. 6334 (2017): 183-86. https://doi.org/10.1126/science.aal4230

Carboni, Kevin. "Come funziona l'intelligenza artificiale che Israele usa per bombardare Gaza." *Wired*. 1° dicembre 2023.

Casilli, Antonio A. *Schiavi del clic*. Milano: Feltrinelli, 2020.

Domscheit-Berg, Daniel. *Inside Wikileaks*. Venezia: Marsilio, 2011.

Chamayou, Grégoire. *Teoria del drone*. Bologna: DeriveApprodi, 2014.

Chen, Tsung Yi, Tsai, Meng Che, & Yuh Min Chen. "A user's personality prediction approach by mining network interaction behaviors on Facebook". *Online Information Review* 40, no. 7 (2016): 913-37. https://doi.org/10.1108/OIR-08-2015-0267

Chiusi, Fabio. *Nessun segreto*. Milano: Mimesis, 2011.

Cosimi, Simone. "Lee Sedol, il campione di Go si ritira per sempre: 'L'intelligenza artificiale è imbattibile.'" *La Repubblica*. 28 novembre 2019.

Cristianini, Nello. *La scorciatoia*. Bologna: il Mulino, 2023.

De Luca, Stefno. *Hic sunt drones, La guerra nell'era digitale*. In *La politica nel mondo digitale*, 41-56. A cura di Gabriele Giacomini & Luca Taddio. Milano: Mimesis, 2023.

Del Vicario, Michela, Bessi, Alessandro, Zollo, Fabiana, Petroni, Fabio, Scala, Antonio, Caldarelli, Guido, Stanley, Eugene H. & Walter Quattrociocchi. "The spreading of misinformation on line". *PNAS* 113, no. 3 (2016): 554-9. https://doi.org/10.1073/pnas.1517441113

Deleuze, Gilles. "Poscritto sulle società di controllo." In *Pourparler*, pp. X-X. Macerata: Quodlibet, 2000.

Diamanti, Ilvo. "Alla periferia della crisi." *Stato e mercato* 38, no. 1 (2018): 117-26. https://doi.org/10.1425/89852

Došilović, Filip Karlo, Brčić, Mario & Nikica Hlupić. "Explainable artificial intelligence: A survey." In *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018: 0210-0215. https://doi.org/10.23919/MIPRO.2018.8400040.

Dunn, Pietro & Massimo Durante. "Disinformazione e Internet. Sfide e prospettive regolatorie in Europa." In *La politica nel mondo digitale*, 57-83. A cura di Gabriele Giacomini & Luca Taddio. Milano: Mimesis, 2023.

Erler, Michael. *Platone*. Torino: Einaudi, 2008.

Esposito, Elena. *Comunicazione artificiale*. Milano: Egea, 2022.

Ferraris, Maurizio. "Dalla tirannia del merito alla democrazia del bisogno." In *La politica nel mondo digitale*, 83-102. A cura di Gabriele Giacomini & Luca Taddio. Milano: Mimesis, 2023.

Ferraris, Maurizio. *Documanità*. Roma-Bari: Laterza, 2021.

Floridi, Luciano. "Perché nel mondo digitale abbiamo bisogno di un progetto umano." In *La politica nel mondo digitale*, 103-14. A cura di Gabriele Giacomini & Luca Taddio. Milano: Mimesis, 2023.

Floridi, Luciano. *Etica dell'intelligenza artificiale*. Milano: Raffaello Cortina, 2022.

Floridi, Luciano. *Il verde e il blu*. Milano: Raffaello Cortina, 2020.

Fossa, Fabio, Schiaffonati, Viola, & Guglielmo Tamburrini, eds. *Automi e persone*. Roma: Carocci, 2021.

Foucault, Michel. *Sorvegliare e punire*. Torino: Einaudi, 2014.

Gellman, Barton & Laura Poitras. "British intelligence mining data from nine U.S. Internet companies in broad secret program." *The Washington Post*. 7 giugno 2013.

Gellman, Barton & Ashkan Soltani. "NSA infiltrates links to Yahoo, Google data centers worldwide, Snowden documents say." *The Washington Post*. 30 ottobre 2013.

Giacomelli, Fabrizio. *Dall'umanità in poi*. Milano: Mimesis, 2023.

Giacomini, Gabriele. "Habermas 2.0." *Ragion pratica* 1 (2020): 31-50.

Giacomini, Gabriele. "Verso la neointermediazione. Il potere delle grandi piattaforme digitali e la sfera pubblica." *Iride* 3 (2018): 457-68. https://doi.org/10.1414/92394

Giacomini, Gabriele. *The Arduous Road to Revolution*. Milano: Mimesis International, 2022.
Greenwald, Glenn. "NSA collecting phone records of millions of Verizon customers daily." *The Guardian*. 6 giugno 2013.
Greenwald, Glenn & Ewen MacAskill. "NSA Prism program taps in to user data of Apple, Google and others." *The Guardian*. 7 giugno 2013.
Habermas, Jürgen. *Fatti e norme*. Milano: Guerini, 1996.
Habermas, Jürgen. *Storia e critica dell'opinione pubblica*. Roma-Bari: Laterza, 1991.
Habermas, Jürgen. *Teoria dell'agire comunicativo*. Bologna: il Mulino, 1986.
Han, Byung Chul. *Psicopolitica*. Roma: Nottetempo, 2016.
Kaiser, Brittany. *Targeted*. New York: HarperCollins, 2019.
Keynes, John Maynard. *Democrazia e mercato*. Milano: Società aperta, 2022.
Knefel, John. "Your Social Media Posts Are Fueling The Future of Police Surveillance." *Inverse*. 20 novembre 2015.
Kosinski, Michal, Stillwell, David & Thore Graepel. "Private traits and attributes are predictable from digital records of human behavior." *PNAS* 110, no. 15 (2013): 5802-05. https://doi.org/10.1073/pnas.1218772110
Lee Fang. "The CIA Is Investing in Firms That Mine Your Tweets and Instagram Photos." *The Intercept*. 14 aprile 2016.
Locke, John. *Il secondo trattato sul governo*. Milano: Società Aperta 2023.
Lyu, Wenjing & Jin Liu. "Artificial Intelligence and emerging digital technologies in the energy sector." *Applied Energy* 303 (2021): 117615. https://doi.org/10.1016/j.apenergy.2021.117615
MacAskill, Ewen, Borger, Julian, Hopkins, Nick, Davies & James Ball. "GCHQ taps fibre-optic cables for secret access to world's communications." *The Guardian*. 21 giugno 2013.
Maffettone, Sebastiano. "Etica pubblica e IA. Da Habermas alla sfera pubblica digitale." In *La politica nel mondo digitale*, 179-94. A cura di Gabriele Giacomini & Luca Taddio. Milano: Mimesis, 2023.
McCrae, Robert R., & Paul T. Costa. "Validation of the five-factor model of personality across instruments and observers." *Journal of Personality and Social Psychology* 52, no. 1 (1987): 81-90. https://doi.org/10.1037/0022-3514.52.1.81
Mitchell, Melanie. *Artificial Intelligence. A guide for thinking humans*. London: Pelican, 2020.
Moretti, Enrico. *La nuova geografia del lavoro*. Milano: Mondadori, 2017.
Nissenbaum, Helene. "Accountability in a Computerized Society." *Science and Engineering Ethics* 2, no. 1 (1996): 25-42. https://doi.org/10.1007/BF02639315
Palano, Damiano. *Bubble democracy*. Brescia: Scholè. 2020.

Parisier, Eli. *The Filter Bubble*. Londra: Penguin, 2011.
Piketty, Thomas. *Le Capital au XXIe siècle*. Parigi: Seuil, 2013.
Rawls, John. *Political liberalism*. New York: Columbia University Press, 1993.
Reyneri, Emilio. "Le basi sociali del populismo." *Stato e mercato* 38, no. 1 (2018): 141-48. https://doi.org/10.1425/89854
Richards, Sam. "Powerful mobile phone surveillance tool operates in obscurity across the country." *The Intercept*. 23 dicembre 2020.
Rosenblat, Alex. *Uberland. How Algorithms Are Rewriting the Rules of Work*. Oakland: University of California Press, 2018.
Sala, Roberta. "Ragionevolezza e irragionevolezza nell'uso del web." In *La politica nel mondo digitale*, 207-20. A cura di Gabriele Giacomini & Luca Taddio. Milano: Mimesis, 2023.
Sharkey, Noel. "Saying No to Lethal Autonomous Targeting." *Journal of Military Ethics* 9, no. 4 (2019): 369-89.
Soro, Antonello. *Persone in rete*. Roma: Fazi, 2018.
Sunstein, Cass R. *#Republic*. Princeton: Princeton University Press, 2017.
Tandera, Tommy, Hendro, Suhartono, Derwin, Wongso, Rini & Yen Lina Prasetio. "Personality prediction system from Facebook users." *Procedia Computer Science* 116 (2017): 604-11. https://doi.org/10.1016/j.procs.2017.10.016
Turing, Alan M. "Computing machinery and intelligence." *Mind* 56, no. 235 (1950): 433-60.
Vaccaro, Salvatore. "Tutto il potere agli algoritmi!" In *La politica nel mondo digitale*, 233-54. A cura di Gabriele Giacomini & Luca Taddio. Milano: Mimesis, 2023.
Verbruggen, Maaike. "The Question of Swarms Control: Challenges to Ensuring Human Control over Military Swarms." *Non-Proliferation and Disarmament Papers* 65 (2019): 1-16.
Vilone, Giulia & Luca Longo. "Notions of explainability and evaluation approaches for explainable artificial intelligence." *Information Fusion* 76 (2021): 89-106. https://doi.org/10.1016/j.inffus.2021.05.009
Walzer, Michael. *Guerre giuste e ingiuste*. Roma-Bari: Laterza, 2009.
Wells, William D. "Psychographics: A Critical Review." *Journal of Marketing Research* 12, no. 2 (1975): 196-213. https://doi.org/10.2307/3150443_
Woodcock, Jamie & Mark Graham. *The gig economy*. Cambridge: Polity, 2019.
Youyou, Wu, Kosinski, Michal & David Stillwell. "Computer-based personality judgments are more accurate than those made by humans." *PNAS* 112, no. 4 (2015): 1036-40. https://doi.org/10.1073/pnas.1418680112