Università degli Studi di Milano-Bicocca

Centro Interuniversitario MaCSIS

MaCSIS Working Paper Series

AUTOMATION BIAS

LA SOSTITUZIONE DEL GIUDIZIO E ALTRI ISTUPIDIMENTI

Serena La Rosa

Working Paper n.1/2020

Università degli Studi di Milano-Bicocca

Dipartimento di sociologia e ricerca sociale

Master in Comunicazione della scienza e dell'innovazione sostenibile

Automation Bias: la sostituzione del giudizio e altri istupidimenti

Relatore
Prof. Federico Cabitza

Tesi di
Serena La Rosa
Matricola n. 863541

Anno Accademico 2019-20

Per Veronica, per Arianna, per Clara.
E per Marco, ancora prima.

Indice

Introduzione

La costruzione di una cittadinanza scientifica non può prescindere dalla consapevolezza del rischio legato all'utilizzo dell'Intelligenza artificiale, e in particolare di sistemi di supporto alle decisioni basati su algoritmi di apprendimento automatico. L'obiettivo di questo lavoro è la documentazione del processo di ricerca, scrittura e realizzazione di un podcast in tre episodi, ognuno di circa 15-20', rivolto a un pubblico non specializzato. Al podcast è stato dato il titolo "Pecore elettriche" – ovvio rimando al romanzo di Philip K. Dick – per evocare un approccio alla tecnologia vicino alla cultura popolare e attento ai risvolti sociali. Al lavoro di documentazione attraverso le fonti è stato aggiunto il contributo di specialisti in diversi ambiti della materia sotto forma di interviste, che nel podcast sono state tagliate e montate per rispettare le tempistiche e costruire una narrazione più lineare.

Nel capitolo 1 si definisce il Machine Learning come sottoinsieme dell'Intelligenza Artificiale, si introduce il concetto di Big Data e si accenna alla varietà dei campi di applicazione. Nel capitolo 2 si approfondiscono gli aspetti cognitivi del processo decisionale, analizzando le cause dei bias in un'intervista con la prof. Laura Macchi. Al fine di chiarire le differenza tra decisioni umane e output informatici, si racconta la sfida a scacchi tra Garry Kasparov e il calcolatore Deep Blue. Si introducono quindi sistemi di supporto decisionale (DSS) basati su algoritmi di ML. Il materiale costituisce la base teorica per la scrittura del primo episodio.

Nel capitolo 3 si esaminano gli errori commessi dalle macchine, con particolare attenzione ai pregiudizi inseriti attraverso i dati di input. Si presenta la definizione di "Armi di distruzione matematica" (O'Neil, 2017) e si esamina la giurisprudenza attinente ai sistemi di ML con un'intervista al prof. Amedeo Santosuosso. Il materiale costituisce la base teorica della scrittura del secondo episodio.

Nel capitolo 4 si discutono i concetti di Automation Bias e *deskilling* attraverso il racconto dell'incidente del volo AF477. A una rassegna delle possibili manovre correttive si affiancano le interviste al prof. Federico Cabitza sulle conseguenze politiche e sociali del deskilling e alla prof. Viola Schiaffonati sulle implicazioni etiche legate al tema della responsabilità. Il materiale costituisce la base teorica della scrittura del terzo episodio.

1. Apprendere senza imparare

1.1 Il Machine Learning

La ricerca sull'intelligenza artificiale (AI) nasce negli anni Cinquanta: il termine viene introdotto da John McCarthy alla conferenza di Darthmouth (McCarthy et al., 1955) cui partecipa anche il ricercatore dell'IBM Arthur Samuel, che con i suoi studi sul gioco della dama (Samuel, 1959) contribuirà a rendere popolare il termine Machine Learning (ML). Per capire cosa sia il ML è pertanto necessario prima accennare ai principi dell'AI. Già Alan Turing nel 1950 suggeriva di sostituire alla suggestiva domanda «Le macchine possono pensare?» un interrogativo più concreto: «Le macchine possono fare quello che possiamo fare noi in quanto esseri pensanti?» (Harnard, 2008). Ovvero: se opportunamente istruite, le macchine possono imitare i comportamenti umani? Il Test di Turing – o come lo chiamava lui: *The Imitation Game* – proposto in *Computing Machinery and Intelligence* (Turing, 1950) aveva lo scopo di determinare se, in base alle risposte fornite durante un interrogatorio esclusivamente testuale, una macchina potesse risultare indistinguibile da un essere umano, nel qual caso la macchina si poteva considerare "intelligente" senza necessità di introdurre valutazioni relative a un'eventuale coscienza. Nel corso degli anni l'impianto teorico del test di Turing è stato molto discusso, ma la ricerca sperimentale sull'AI ha cominciato a svilupparsi soltanto quando i computer sono diventati abbastanza potenti da processare i dati a una velocità adeguata.

Definiamo AI ogni programma o dispositivo capace di percepire l'ambiente circostante e agire in modo da ottimizzare le probabilità di portare a termine uno specifico compito. Ovvero: macchine che «replicano il risultato del pensiero umano senza replicare il pensiero» (Carr, 2017). Il ML è un sottoinsieme dell'AI costituito da algoritmi in grado di migliorare autonomamente le proprie prestazioni sulla base di informazioni date. Negli anni Novanta avviene inoltre uno spostamento fondamentale nell'ambito della ricerca: si passa dal cosiddetto *knowledge-driven* ML (dove l'apprendimento è basato sulla conoscenza, ovvero su un set di istruzioni) al *data-driven* ML: un insieme di tecniche statistiche in grado di individuare modelli significativi (*pattern*) in opportune masse di dati.

> In simple terms, machine learning is a statistical method for discovering correlations in past events that can then be used to make predictions about future events. Rather than giving a computer a set of instructions to follow, a programmer feeds the computer many examples of a phenomenon and from those examples the machine deciphers relationships among variables. Whereas most software programs apply rules to data, machine-learning algorithms do the reverse: they distill rules from data, and then apply those rules to make judgments about new situations. (Carr, 2017).

Si capisce quindi che quando parliamo di predizioni non attribuiamo alla macchina mitologiche capacità divinatorie: sono piuttosto stime basate sull'assunto che quello che è successo in passato continuerà a verificarsi in futuro. Il concetto di autonomia, inoltre, rimane da intendersi all'interno di un preciso set di vincoli e obiettivi.

1.2 Che cosa sono i Big Data

Perché i sistemi di *data-driven* ML possano funzionare, ovvero riescano ad approssimare in maniera appropriata la funziona che mette in relazione le variabili di ingresso ai valori in uscita, devono prima essere addestrati su grandi moli di dati. È utile a questo punto precisare che «grandi moli» non significa necessariamente Big Data, ma più semplicemente un numero di casi adeguato a individuare quella funzione. La quantità di dati, inoltre, non determina la qualità dell'algoritmo, anzi: un dataset più piccolo può risultare più completo e/o accurato.

Il termine Big Data viene usato per indicare una raccolta di dati informativi così grande in volume e varietà da richiedere specifiche tecnologie per estrarne informazioni di valore. Possono essere numeri, immagini, grandezze, parole, clic su una pagina web, transazioni finanziarie, spostamenti, o qualunque altro input possa essere digitalizzato e archiviato. Le sorgenti tipiche sono i sensori installati negli oggetti (l'internet delle cose) e tutte le attività che si svolgono sui social network.

A causa dell'enorme quantità di informazioni che viaggiano su internet ogni giorno, è difficile calcolare la quantità esatta di dati creati quotidianamente.

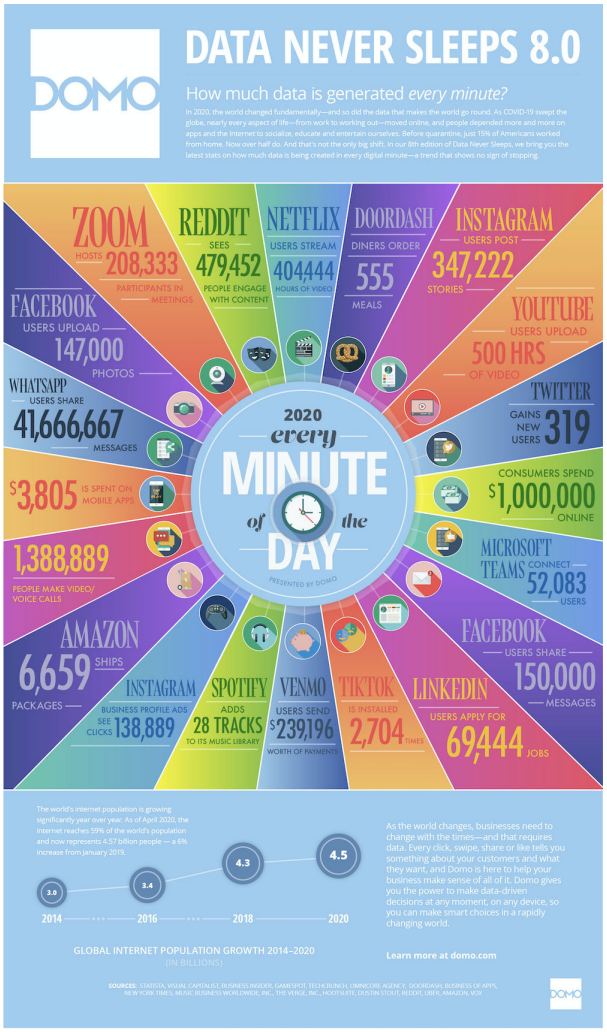


Fig. 1.1 Infografica dei dati generati al minuto nel 2020 per provenienza (Domo, 2020)

Integrando diverse fonti, la società americana Domo ha stimato che nel 2020 ogni individuo ha generato 1.7MB di dati al minuto (Fig.1.1). Per avere un'idea delle grandezze in gioco, si consideri che nel 2020, il volume dei dati dell'intero universo digitale è stato stimato pari circa a 44 ZettaBytes (ZB), dove il prefisso "Zetta" indica la settima potenza di 1000: 10^{21} bytes. Si prevede che nel 2025 la produzione quotidiana di dati arrivi a 463 Exabyte (EB, 10^{18} bytes) (Racounter, 2019) ma è una stima fatta prima della pandemia di Covid-19, che ha di fatto provocato un significativo aumento dell'attività online.

1.3 Le applicazioni del Machine Learning

Il ML viene usato quando si desidera che il sistema migliori le proprie prestazioni nel tempo. Un'applicazione tipica di ML si realizza negli algoritmi di personalizzazione dei servizi web: allo scopo di massimizzare la probabilità d'acquisto di un prodotto e/o il tempo di utilizzo di un dispositivo, il sistema colleziona abitudini e preferenze di tutti gli utenti sulla base delle quali produce raccomandazioni individuali il cui successo informa la generazione delle raccomandazioni successive. All'interno delle tecnologie di ML esiste poi il sottoinsieme del Deep Learning (DL) che utilizza strutture computazionali connesse secondo il modello del cervello umano, ovvero reti neurali artificiali. Le applicazioni più comuni sono quelle legate ai sistemi di riconoscimento vocale o facciale, ai programmi di traduzione automatica, all'analisi diagnostica delle immagini.

Il DL viene utilizzato anche nelle *self-driving cars*, le macchine senza autista. Equipaggiate con termometri laser, trasmettitori radar e sonar, rilevatori di movimento, ricevitori GPS e sistemi di visione a 360° per "vedere" dove stanno andando e percepire l'ambiente circostante nel dettaglio, questi algoritmi possono elaborare istantaneamente tutti gli input, gestire acceleratore e freno con precisione, e rispondere agevolmente agli eventi inaspettati (Carr, 2015). Le automobili Waymo, il ramo aziendale di Alphabet – la società che contiene Google – dedicato alla guida automatica, sono state le prime a ottenere l'autorizzazione per percorrere le strade pubbliche dell'Arizona senza la presenza di un autista umano di sicurezza. La compagnia sostiene di poter contare sul «cervello più avanzato oggi su strada», in grado di modellare non solo operazioni come il riconoscimento degli oggetti sulla carreggiata, ma anche la maniera in cui il comportamento umano influisce sul comportamento della macchina, grazie ai meccanismi di DL addestrati sui dati accumulati durante le sei milioni di miglia percorse nel mondo reale, e gli altri cinque miliardi guidati in simulazione (Hawkins, 2018).

Secondo il primo report pubblicato da Waymo, nel periodo che comprende il 2019 e primi nove mesi del 2020, i veicoli a guida autonoma sono stati coinvolti in 18 incidenti e 29 "mancate collisioni" senza gravi conseguenze per gli esseri umani (Schwall et al., 2020). Ma le cronache recenti riportano incidenti con conseguenze più gravi, come per esempio quando nel 2018 a Tempe, Arizona, una macchina Uber ha investito e ucciso una ciclista. Al di là delle difficoltà tecniche ancora da superare, la guida autonoma – con o senza supervisione umana – pone un complesso problema legale, etico e culturale. Nel caso di Tempe in esempio, il *National Transportation Safety Board* ha determinato che la probabile causa dell'incidente è da riconoscere nell'incapacità dell'operatore umano di monitorare sia l'ambiente di guida che il funzionamento del sistema automatizzato perché distratto dal cellulare (National Transportation Safety Board, 2019). Ma per quale ragione dovremmo lasciar guidare un computer, se poi non possiamo impiegare quel tempo nella maniera che più ci aggrada? La riconfigurazione delle competenze umane in un ambiente sostanzialmente modificato dagli interventi dei sistemi automatici è un nodo innanzitutto culturale, la cui comprensione diventa necessaria allo sviluppo di tecnologie effettivamente migliorative.

2. Errori e pregiudizi

2.1 Errare è umano

Ci sono diversi tipi di errori che intervengono sistematicamente nel processo decisionale umano, ovvero nella modalità di pensiero finalizzata alla valutazione di alternative e all'intrapresa di scelte. Si chiamano bias cognitivi: sono preferenze, inclinazioni, euristiche o distorsioni della percezione spesso dovute alla necessità evolutiva di velocizzare le nostre reazioni, ma che si rivelano perniciose nel lungo termine o in contesti inadatti. A titolo di esempio, alcuni bias tra i più noti sono:

- *Halo Effect* (effetto alone) – la percezione di un tratto positivo dell'individuo influenza la percezione di altri tratti. il caso classico è quello della bellezza, cui può venire associata intelligenza o rettitudine morale.
- *Confirmation Bias* (bias di conferma) – il meccanismo per cui interpretiamo e selezioniamo le informazioni in modo da confermare la nostra ipotesi, ignorando quelle che la contraddicono.
- *In-Group Bias* (favoritismo di gruppo) - la tendenza a prediligere il proprio gruppo di appartenenza su basi etniche, socio-culturali, politiche o religiose.
- *Similarity Bias* (pregiudizio di somiglianza) - il pregiudizio positivo nei confronti delle persone che ci somigliano, significativo nell'ambito della selezione del personale quando il manager uomo tende a preferire candidati maschi.

Per approfondire cause e manifestazioni dell'errore umano, e capire come questi fattori intervengono nel processo decisionale, abbiamo intervistato Laura Macchi, professore ordinario di Psicologia del pensiero della decisione della comunicazione all'Università degli Studi di Milano-Bicocca[1].

Che cosa sono i bias?

Si usa il termine specifico bias – che in inglese significa inclinazione, quindi pregiudizio – proprio per sottolineare il fatto che non si tratta di errori legati a

[1] La conversazione con la prof. Macchi, avvenuta con la sottoscritta via Skype il 12/02/21, è stata di seguito sintetizzata e redatta. Le domande sono riportate in neretto.

situazioni accidentali come la stanchezza o un calo dell'attenzione, quanto piuttosto alle caratteristiche strutturali del nostro sistema cognitivo. Rappresentano un sostanziale scarto, una distanza fra la performance del soggetto umano – quello che di fatto il soggetto valuta a livello di probabilità degli eventi futuri, oppure di opzione migliore in un contesto decisionale – rispetto a un parametro normativo che è esterno alla psicologia, ma può essere tratto dalla logica classica, nel caso del pensiero deduttivo, o dalla teoria della massimizzazione dell'utilità, nel caso delle decisioni.

Questo vuol dire che gli uomini non sono in grado di giudicare oggettivamente?

Ci sono esperimenti sia sociali, in ambito applicativo, sia sperimentali in ambito di laboratorio, che danno esito a due rappresentazioni differenti del nostro modo di funzionare. In un caso risultiamo molto proni ai bias e propensi a commettere errori; nell'altro invece siamo in grado di risolvere i problemi creativamente. L'enigma della ragione si compone di queste due facce, e allora conviene soffermarsi su quali siano le caratteristiche del funzionamento cognitivo che danno esito, a seconda delle condizioni, ai bias piuttosto che alle soluzioni creative. Quella che ci distingue dai computer è una *Working Memory* relativamente modesta: una limitata capacità di prestare attenzione ed elaborare informazioni che ci impedisce di lavorare in parallelo, almeno a livello di focus attentivo consapevole. Per procedere, noi dobbiamo restringere il campo: non potendo essere esaustivi nell'elaborazione, quando le informazioni sono troppe lavoriamo a ciò che ci appare più rilevante. Siamo abituati a valutare la salienza dell'informazione, e a cercare di cogliere gli indizi informativi che provengono sia dal contesto che dal contenuto. Ma se pensiamo al ragionamento logico, bisognerebbe prescindere dai contenuti e guardare solo la forma. Invece noi ci siamo adattati a elaborare queste informazioni di contenuto e contesto nel modo più fruttuoso possibile, per cogliere relazioni che diano senso e ci consentano di interpretare i dati collegandoli l'uno all'altro. Ecco, i bias si producono quando non si tiene conto del modo che abbiamo di elaborare le informazioni: spesso non sono che il frutto della discrepanza fra queste caratteristiche umane e quello che ci viene richiesto in una determinata situazione. Possono sembrare errori, di ragionamento o di valutazione, quando in realtà sono

fraintendimenti. Penso ad esempio alla questione del linguaggio: se in una situazione decisionale il linguaggio utilizzato è troppo distante dal linguaggio naturale, allora potremmo avere come risultato un bias apparente.

Nell'attività di elaborazione di contenuto e contesto, qual è il ruolo del corpo?
È un ruolo molto importante, ma va considerato tutto il meccanismo con cui si sta in un ambiente. C'è tutta una letteratura sul *nudge* e l'architettura delle scelte: come predisporre gli ambienti secondo un design decisionale affinché il soggetto sia favorito in certi comportamenti e non ostacolato. È quello che succede con l'apertura delle porte: a seconda che si mettano maniglie oppure placche viene indotto un certo tipo di comportamento, spingere o tirare. Tutto contribuisce a veicolare un messaggio, e questo messaggio deve essere il più possibile ben interpretato. Il corpo ovviamente interagisce con l'ambiente fisico, ma generalizzando un po' direi anche che sta dentro al nostro armamentario, ad esempio quello linguistico, per interagire con l'ambiente linguistico che incontriamo: pensiamo a tutte le istruzioni e indicazioni che gli operatori incontrano sotto forma di segnali quando si trovano a pilotare un aereo. Tutto quanto sia da utilizzare a livello ergonomico – perché c'è il corpo in ballo – ma anche a livello interpretativo deve tener conto di come funzioniamo, di come percepiamo. L'idea del corpo è molto interessante, ma va inserita all'interno della naturale predisposizione del soggetto umano a interagire con un esterno. Il corpo è un altro canale di scambio informativo, e in fondo la regola è un po' la stessa. Così come per evitare i bias devo parlare in un certo modo, acciocché l'altro non incorra in fraintendimenti, lo stesso succede a livello ergonomico: non basta dire "tiri quella leva" se va contro al modo naturale di utilizzare l'oggetto. E neanche è sufficiente avvertire che in quello specifico caso si deve usare diversamente, perché poi diventa un compito di apprendimento e non più una situazione che induce a reagire velocemente. Quando si disegna l'ambiente decisionale bisogna fare uno sforzo per tener conto di tutte queste caratteristiche.

Inducendo nelle persone un comportamento automatico si favorisce anche l'apprendimento o è invece necessario introdurre un attrito?
Perché una funzione diventi routine è necessario che la fase di immagazzinamento

sia lunga. È richiesto un certo sforzo perché avvenga una sedimentazione effettiva, altrimenti si rischia che anche l'apprendimento scivoli via. La teoria dei processi duali ipotizza che ci siano due velocità di apprendimento e di comportamento: un pensiero più intuitivo, automatico e routinario, che è più veloce nel dare risposte; e un pensiero più riflessivo, analitico e cosciente. Se pensiamo al travaso degli apprendimenti avvenuti attraverso l'attivazione del sistema più riflessivo nel sistema più intuitivo – per esempio: un maestro di scacchi che riesce a giocare in contemporanea più partite – si vede che il processo di acquisizione è stato comunque lento e laborioso. Quanto più una funzione è sedimentata, tanto più diventa un automatismo efficace, mentre in caso contrario la routine che si crea può risultare inadatta in certe situazioni. Quando si agisce in condizioni di fretta o di pressione, infatti, oppure ci si focalizza su altri aspetti in quel momento urgenti, quella procedura automatica può decadere. Questo succede a meno di non essere molto attenti nelle operazioni di autocontrollo – cosa non facile – oppure introducendo degli automatismi "fino a un certo punto", che richiamino l'attenzione sui parametri essenziali da considerare per avere un quadro completo.

2.2 Uomo contro macchina

Gli scacchi sono da sempre considerati una prova di intelligenza: un gioco di grande strategia ma anche di creatività, in cui è fondamentale mantenere, oltre alla razionalità, lucidità e visione d'insieme. Per questo, negli anni Novanta, il match tra l'allora imbattuto Garry Kasparov e il computer IBM Deep Blue diventa il simbolo di un confronto universale: quello tra intelligenza artificiale e intelligenza umana. Nel 1996, a Philadelphia, Kasparov perde sì il primo incontro, ma poi vince la partita al meglio dei sei giochi: la prima sconfitta di un campione del mondo contro un calcolatore può essere archiviata come incidente di percorso. Ma Kasparov offre a Deep Blue la rivincita e a New York, nel 1997, il più grande giocatore di sempre perde la sfida. Un brusco risveglio per l'umanità? È più complicato, anche molto meno grave, di così.

La strategia attuata da Deep Blue si basa sul cosiddetto approccio della "forza bruta": grazie alla sua potenza computazionale il computer può esaminare tutte le

possibili mosse per poi scegliere la più efficace. Non impara dai suoi sbagli, non prevede il comportamento dell'avversario, ma ogni volta riparte da capo: è proprio questa «smemorata oggettività», scriverà Kasparov (2019), a fare dei computer «eccellenti strumenti di analisi – e avversari pericolosi».

> Technology can advance in leaps, and IBM had invested heavily. I lost that game. And I couldn't help wondering, might it be invincible? Was my beloved game of chess over? These were human doubts, human fears, and the only thing I knew for sure was that my opponent Deep Blue had no such worries at all (Kasparov, 2017).

Ed è vero, Deep Blue non conosce stress o preoccupazione. Ma la sua vittoria ha più a che vedere con l'astuzia umana che con l'infallibilità della forza bruta. I programmatori IBM intendono usare ogni mezzo pur di vincere, pertanto programmano la macchina perché proceda nel gioco in maniera emotiva, inserendo inutili ritardi prima di muovere, simulando indecisione, oppure rispondendo con tale prontezza da lasciar temere all'avversario di essere caduto in una trappola. Lo snodo è la 36esima mossa del secondo gioco, quando invece di giocare nella maniera più razionale – come Kasparov si aspetta – Deep Blue adotta una strategia più raffinata, quasi umana. Kasparov è destabilizzato: non sa più chi ha davanti, si deconcentra, si arrende. Arrivati in parità alla sesta e decisiva partita, Kasparov commette un errore grossolano e dopo poche mosse concede. IBM rifiuta di offrire la rivincita e Deep Blue viene ritirato, ma a tutta la ricerca sull'IA viene impressa una straordinaria accelerazione (Levy, 2017).

Nel 2016 il programma di Google AlphaGo sconfigge Lee Sedol, pluricampione mondiale di Go: un gioco molto più complesso degli scacchi. AlphaGo non usa trucchi per destabilizzare psicologicamente l'avversario, ma un sistema di reti neurali e apprendimento automatico per rinforzo (nel *Reinforcement Learning* ogni decisione viene valutata per incoraggiare i comportamenti corretti). Possiamo quindi supporre che, con sufficiente disponibilità di tempo, denaro e potenza computazionale, non ci sia sfida che un computer non possa vincere. Per questo, in ambiti ben più rilevanti del gioco di strategia, si conclude che affidarsi a sistemi di ML per minimizzare gli effetti degli errori umani – siano essi dovuti a un pregiudizio inconscio o a una distrazione emotiva – sia tutto sommato una buona idea.

Tabella 2.1 Livelli di automazione della decisione e della selezione dell'azione (Parasuraman et al., 2000)

HIGH	10. The computer decides everything, acts autonomously, ignoring the human.
	9. informs the human only if it, the computer, decides to
	8. informs the human only if asked, or
	7. executes automatically, then necessarily informs the human, and
	6. allows the human a restricted time to veto before automatic execution, or
	5. executes that suggestion if the human approves, or
	4. suggests one alternative
	3. narrows the selection down to a few, or
	2. The computer offers a complete set of decision/action alternatives, or
LOW	1. The computer offers no assistance: human must take all decisions and actions.

2.3 La sostituzione del giudizio

Ogni decisione deriva da un giudizio, formulato all'interno di un processo cognitivo che coinvolge elementi di percezione, esperienza, aspettativa e valori personali, e si traduce nell'intrapresa di un corso di azioni tra diverse alternative disponibili. Come abbiamo visto (cfr. 2.1), il giudizio umano è soggetto a errori e bias: in medicina si valuta che l'incidenza degli errori diagnostici sia intorno al 10-15% (Graber, 2013); nell'ambito delle risorse umane si stima che il processo di assunzione, che è quello con il più alto impatto economico, fallisca nel 50% dei casi (Sullivan, 2017). Pertanto a partire dagli anni Ottanta, in contesti organizzativi o aziendali, si intensifica il ricorso a sistemi di *Machine Learning-Decision Support System* (ML-DSS) per operare scelte sia operative che di policy making. In presenza di un numero finito di variabili da valutare, il computer suggerisce la decisione migliore: una diagnosi, un'assunzione, la valutazione di un rischio.

I DSS supportano gli esseri umani in compiti decisionali complessi, come per esempio la gestione di grandi impianti di produzione o il monitoraggio degli impatti del cambiamento climatico su specifiche aree, ma anche la valutazione del rischio di recidiva di un criminale. Sono tipicamente composti da: un set di dati; un software,

cioè un modello matematico con i criteri di valutazione; un'interfaccia utente. E possono essere introdotti a qualunque livello dell'interazione uomo-macchina.

Se con automazione intendiamo la sostituzione parziale o completa di una funzione precedentemente svolta dall'operatore umano, il suo utilizzo può variare in uno spettro continuo di valori compresi tra il "completamente manuale" e il "completamente automatico", come schematizzato nella Tabella 1 (Parasuraman et al., 2000). Al livello 2 e 3, il computer offre una serie di opzioni lasciando la decisione all'operatore. Al livello 4 e 5, il computer suggerisce un'alternativa e la esegue previa approvazione: è il confine su cui si gioca la maggiore rilevanza del giudizio umano, in una condizione di consapevolezza aumentata dall'intervento della macchina. Ai livelli successivi, l'azione dell'operatore viene progressivamente limitata fino alla sua completa rimozione dal processo. Questo significa che il giudizio umano viene completamente sostituito da quello automatico, inserendo il rischio di una progressiva deresponsabilizzazione dell'operatore.

3. Macchine che sbagliano

3.1 Dai Big Data ai Deep Data

> "Whenever you solve a problem you usually create one. You can only hope that the one you created is less critical than the one you eliminated." – Legge di Wiener #29 (Croft, 2013)

Se per rimuovere il problema dell'errore umano si sostituisce al giudizio degli uomini quello delle macchine, si inserisce nel processo decisionale il problema dell'errore delle macchine. E non è detto che sia meno grave: i computer sono intelligenti «allo stesso modo in cui può esserlo una radiosveglia» (Kasparov, 2019) e perciò si limitano a esaminare o correlare un numero enorme di informazioni per conseguire un risultato assegnato, senza costruire una reale forma di conoscenza. La loro forza – quella che Kasparov descrive come "smemorata oggettività" – è anche la loro debolezza perché «a differenza della scacchiera, il mondo è un posto senza confini, e trovargli un senso richiederà sempre qualcosa in più di calcoli matematici o statistici» (Carr, 2017).

Chiamiamo Machine Bias l'errore sistematico per cui un algoritmo produce output iniqui. Un sistema di ML che funziona correttamente riflette la realtà descritta dai dati che elabora, secondo uno schema che può efficacemente sintetizzarsi in *Garbage In, Garbage Out*: se i dati in ingresso sono sporchi, lo saranno anche quelli in uscita. In questo senso – e solo in questo senso – si può dire che «i dati parlano da soli»: in quanto artefatti umani, raccontano la storia di come e dove sono stati ricavati e analizzati. I bias che si nascondono nelle operazioni di raccolta e analisi influiscono sul risultato finale tanto quanto i dati stessi (Crawford, 2013).

Le tecniche di riconoscimento dei pattern statistici presentano diversi tipi di errori che possono influire sulle prestazioni. Il *data leakage* (fuga di dati) è un fenomeno che si verifica quando una variabile inclusa nel set di training contiene più informazioni di quelle che si avrebbero a disposizione nella pratica. Per esempio, nel 2008 un sistema di ML per la rilevazione del cancro nelle mammografie prevedeva l'addestramento del modello su un set di dati contenente anche gli ID dei pazienti. Ma questi ID erano stati assegnati consecutivamente, quindi potevano essere sfruttati per determinare la fonte dei dati e accrescere il potere predittivo. Nella pratica gli ID

dei pazienti sono casuali: un algoritmo così costruito, nonostante le ottime prestazioni in fase di test, funzionando nel mondo reale avrebbe fornito risultati più scarsi. Un altro fenomeno deleterio è quello del *dataset shift* (trasformazione del dataset): si verifica a quando le condizioni in cui il modello viene addestrato sono diverse da quelle in cui viene distribuito. Per esempio: un sistema di riconoscimento delle immagini addestrato su una serie prodotta in condizioni di luce controllata potrebbe fallire quando utilizzato su immagini prodotte in condizioni di luce variabili (Gretton, 2018).

C'è poi un altro tipo di errore, legato alla raccolta dei dati, che Crawford (2013) chiama *signal problem*, e che si verifica quando «si presume che i dati riflettano accuratamente il tessuto sociale, e invece presentano lacune significative perché da certe comunità proviene un segnale meno potente, o nessun segnale».

> Ad esempio, a Boston c'è un problema di buche per strada: se ne chiudono circa 20.000 all'anno. Per allocare le risorse in modo efficiente, la città di Boston ha rilasciato l'eccellente app per smartphone StreetBump, che attinge ai dati dell'accelerometro e del GPS per rilevare passivamente le buche, segnalandole immediatamente. Sebbene sia certamente un approccio intelligente, StreetBump ha un signal problem. Le persone nei gruppi a basso reddito negli Stati Uniti hanno meno probabilità di avere uno smartphone, e questo è particolarmente vero per i residenti più anziani, dove la penetrazione degli smartphone può arrivare fino al 16%. In città come Boston, questo significa che i set di dati raccolti dagli smartphone non ricevono input da parti significative della popolazione, spesso le stesse che hanno le minori risorse. Fortunatamente, l'Ufficio di New Urban Mechanics di Boston è a conoscenza di questo problema e collabora con una serie di accademici per tenere in conto le questioni di accesso equo e digital divide. (Crawford, 2013)

Per minimizzare questo genere di problemi, è necessario che per ogni ML-DSS vengano dichiarate le specifiche circostanze in cui si sono misurate le prestazioni – come per esempio le caratteristiche tecniche o demografiche del set di training – e segnalate le condizioni in cui il sistema non funziona al meglio. È necessario considerare più approfonditamente la qualità dei dati che vengono utilizzati, dargli maggior spessore, affiancando alla raccolta una rigorosa ricerca qualitativa, per trasformare i Big Data in *deep data*.

3.2 Weapons of Math Destruction

Come abbiamo visto (cfr. 3.1) anche le macchine possono avere pregiudizi: i nostri. Ma a causa dell'uso ormai massivo e talvolta indiscriminato che se ne fa per valutare e predire i comportamenti delle persone, questi bias possono avere conseguenze

sociali devastanti. L'autrice Cathy O'Neil – già professoressa di matematica, quindi consulente finanziaria esperta di fondi speculativi, oggi attivista di Occupy Wall Street e fondatrice di una società di consulenza «che aiuta le compagnie e le organizzazioni a gestire e valutare il rischio algoritmico» (ORCAA, 2018) – individua una categoria particolarmente critica di modelli che chiama *Weapons of Math Destruction* (WMD) ovvero: armi di distruzione matematica – con un gioco di parole intraducibile in italiano, basato sull'assonanza tra *math* e *mass*, e quindi con le armi di distruzione di massa – che descrive nel suo libro uscito nel 2016.

> Come fossero divinità, questi modelli matematici erano misteriosi, e i loro meccanismi invisibili a tutti, tranne che ai sommi sacerdoti della materia: matematici e informatici. I loro giudizi – anche se sbagliati o pericolosi – erano incontestabili e senza appello. E se da una parte penalizzavano i poveri e gli oppressi della nostra società, dall'altra aiutavano i ricchi ad arricchirsi sempre di più. (O'Neil, 2017)

Secondo la classificazione di O'Neil, le WMD sono:

- misteriose: funzionano sul modello della scatola nera
- incontestabili, inappellabili e spesso sollevate da ogni responsabilità
- costruite per rinforzare ricorsivamente ogni pregiudizio, modificando la realtà attraverso le loro conseguenze in un pernicioso ciclo di feedback
- particolarmente punitive nei confronti delle categorie più povere e svantaggiate

Sono sistemi che esercitano un impatto decisivo su tutte le persone che nella vita si trovino a dover cercare un lavoro, chiedere un mutuo, stipulare un'assicurazione, essere giudicati da un tribunale, prenotare un biglietto aereo, iscriversi a un'università a numero chiuso – praticamente l'intera società. E a differenza degli algoritmi che vengono usati, per esempio, in ambito medico, per i quali si prevede un confronto tra il risultato previsto e quello reale, e quindi la possibilità di individuare e correggere eventuali bias iniziali, le WMD non tengono conto degli errori che commettono e non sono sottoposte ad alcuna verifica.

Le conseguenze economiche di questo genere di sistemi possono essere drammatiche, ma la questione è anche profondamente etica. I sistemi decisionali devono essere progettati secondo principi di equità e di trasparenza, così da poter capire secondo quali percorsi da certi dati in ingresso si arriva a una determinata uscita. Un funzionamento di tipo oracolare non è accettabile in un sistema

democratico. Le predizioni, inoltre, devono essere continuamente sottoposte a verifica, per individuare ed eventualmente correggere i bias presenti nello storico dei dati. Altrimenti, quando il valore di ogni individuo viene quantificato in un punteggio, in una posizione all'interno di una classifica assoluta – privata del contesto – quel numero, quella posizione in classifica, rischia di diventare un destino.

3.3 La legge in Italia

Dal momento che i sistemi automatici per il supporto alle decisioni possono commettere degli errori, è necessario prevederne la gestione e tutelare i cittadini dagli eventuali malfunzionamenti sotto il punto di vista legale. Per analizzare le conseguenze giuridiche dell'AI abbiamo intervistato il professor Amedeo Santosuosso, che insegna Diritto, scienza e nuove tecnologie all'Università di Pavia; ICT e diritto presso l'Istituto Universitario di Studi Superiori di Pavia, ed è direttore scientifico dello *European Center for Law Science and New Technologies*[2].

Come sono regolamentate in Italia le applicazioni di intelligenza artificiale?
In Italia questi sistemi non hanno una regolamentazione diretta, quindi la possibilità che ci siano degli errori sistematici – i bias – rientra nel rischio generale. Quando sono coinvolti i diritti delle persone può entrare in gioco il Regolamento europeo sulla protezione dei dati (GDPR) che prevede il divieto di adottare delle decisioni su base esclusivamente automatica. Però se parliamo delle operazioni di intelligenza artificiale a scopo conoscitivo, non focalizzato alla decisione riguardo a una specifica persona, non c'è una regolamentazione specifica. C'è poi un discorso di alfabetizzazione tecnica: è chiaro che io non potrò mai essere un ingegnere informatico, però se mi occupo di tecnologia, devo farmi carico di conoscere alcuni di questi aspetti; così come chi lavora sul versante tecnico deve farsi carico di alcuni vincoli giuridici, come per esempio il fatto che non c'è solo la responsabilità che deriva dalla violazione delle norme sulla privacy, ma ci può essere anche quella che deriva dalle norme generali sulla responsabilità. Mi spiego: se io ho sul mio balcone un vaso di fiori bellissimo, ma non lo assicuro bene e finisce sulla testa di un

[2] La conversazione con il prof. Santosuosso, avvenuta via Skype con la sottoscritta il 12/02/21, è stata di seguito sintetizzata e redatta. Le domande sono riportate in neretto.

passante, io sono responsabile civilmente di questa cosa. Alla stessa stregua può accadere che un sistema di intelligenza artificiale non sia congegnato in modo da non arrecare danno, e quindi produca una responsabilità giuridica prevista dal codice civile. Invece c'è l'idea che quello che riguarda le regolamentazioni dell'intelligenza artificiale rientri nel concetto di privacy: è vero, ma non è tutta la verità. Io non ho una visione terrificante del diritto delle applicazioni tecnologiche, tutt'altro, però bisogna essere consapevoli che esistono dei vincoli.

In Italia esistono i sistemi di supporto alle decisioni giuridiche?

No, in Italia questa cosa non c'è. Purtroppo – dico io. Certo andrebbe gestita con intelligenza, e stiamo lavorando perché si facciano dei passi avanti. Racconto la storia del caso Loomis perché è significativo per capire gli effettivi contorni della questione. È il caso di un signore arrestato dalla polizia alla guida di una macchina usata durante una sparatoria. Non c'è dubbio sul fatto che abbia commesso un reato, il problema nasce con la valutazione del rischio di recidiva. Il giudice consulta un sistema che si chiama Compas, sviluppato da una società privata, e applica la pena di sei anni suggerita automaticamente. L'avvocato va dal giudice a chiedere spiegazioni, e vuole sapere come funziona questo software. Ma Compas è un software proprietario: la società ha i diritti di privativa industriale e nessuna intenzione di dire come funziona, perché su questa cosa fa profitti. Questo è il problema. Anche in Italia i giudici fanno dei piccoli conti, quando si tratta di applicare le varie componenti della pena: a volte nelle camere di consiglio ci sono le calcolatrici. Questi conti vanno fatti sulla base di alcuni criteri che sono indicati dalla legge, per esempio l'ambiente familiare, la ripetizione e la frequenza di alcuni reati, e via dicendo. Ammettiamo che il Ministero della Giustizia italiano, che è quello che fornisce le attrezzature tecniche, crei un software in cui vengono inseriti i dati storici di quello che è accaduto statisticamente sulla base dei criteri indicati dalla legge. Ammettiamo che questo software a disposizione di tutti i giudici italiani sia anche frutto di un confronto con gli avvocati, e che sia rivedibile, a seconda dei problemi in fase di applicazione. Ammettiamo infine che il giudice abbia la possibilità di discostarsi dall'esito prodotto dal software – così come ce l'aveva il giudice americano, peraltro. A questo punto: sarebbe un problema? Se in Italia i giudici

venissero dotati di un sistema che in modo uniforme, basato su dati e criteri condivisi tra la categoria dei giudici e la categoria degli avvocati, sarebbe una cosa mostruosa? Io dico di no. Il vero problema, nel caso Loomis, è che si trattava di un software proprietario. Il problema era il diritto di proprietà industriale, che portava a un risultato segreto. C'è un'altra obiezione che può essere fatta, e cioè che la macchina tende a confermare quello che i giudici hanno fatto prima, e quindi riduce il tasso di possibile innovazione, nel senso di valutazione calibrata: questo è vero. Ma la macchina cerebrale umana è capace di bias persino peggiori: il giudice applica criteri che in alcuni casi discendono da riferimenti normativi, sì, ma in altri possono essere il risultato di un'attitudine mentale. Il nemico della libertà di decisione è la routine. La perfezione non esiste, ma bisogna saperci lavorare dentro. Com'è giusto che i giudici tengono viva la capacità critica rispetto alle loro stesse convinzioni, così devono essere capaci di valutare l'apporto di un ausilio tecnico.

Esiste il rischio di una dequalificazione professionale per giudici o avvocati?

È un problema di formazione professionale. Prenda un pilota d'aereo: mentre prima doveva imparare solo a muovere la cloche e leggere alcuni dati, ora deve saper guardare criticamente anche il risultato della macchina. La stessa cosa vale per il giudice, e per tutti i soggetti chiamati a prendere decisioni. Ma c'è anche un altro problema, che è quello della costruzione dei processi decisionali. Stefano Quintarelli, che è stato nel gruppo di esperti autori dello *White Paper on Artificial Intelligence at the service of citizens*[3], ha presentato allo *High-level Expert Group on AI* della Commissione europea la proposta del *Redress by design*[4], ovvero di prevedere già in fase di progetto un sistema di ridondanze. Cosa succede quando l'umano e la macchina hanno opinioni diverse? In quei casi è saggio progettare un secondo sistema di elaborazione e controllo. È una soluzione onerosa, certo, ma le tecnologie non sono una bacchetta magica: vanno gestite con intelligenza. La proposta di Quintarelli è una proposta assolutamente ragionevole, che lui ha elaborato in ambito medico. In un articolo di

[3] cfr. https://ia.italia.it/assets/whitepaper.pdf [18/02/2021]

[4] cfr. https://blog.quintarelli.it/2019/04/we-need-redress-by-design-for-ai-systems.html [18/02/2021]

prossima uscita su *Agenda digitale*[5] provo a trasferire questo approccio in ambito giuridico.

Chiudiamo con una domanda tratta dal suo libro *Intelligenza artificiale e diritto* (Santosuosso, 2020): lei preferirebbe essere giudicato da un robot o da un umano?
L'intelligenza artificiale, al momento, non è in grado di giudicare nulla, perché il giudizio, in ambito giudiziario e non, richiede degli input consapevoli e inconsapevoli molto più numerosi e ampi, quindi io preferirei essere giudicato da un giudice umano. Ma vorrei che fosse un umano colto, non solo nei limiti delle necessità professionali, ma anche nelle capacità di gestire la macchina. Su questo ci sono due orientamenti: c'è una parte degli scienziati – uno di questi era Stephen Hawking – che paventa il momento in cui le macchine supereranno le capacità degli umani, quando si verificherà la famosa *singularity*. E poi c'è un'intervista stupenda a Roger Penrose – matematico, premio Nobel per la fisica, amico e mentore di Hawking – in cui dice che «il pericolo è credere che queste macchine possano essere più intelligenti di noi e quindi inchinarsi e fare tutto quello che dicono»[6]. La mia opinione è che le macchine e gli esseri umani co-evolvono: bisogna fare questo salto concettuale. Non è garanzia di niente di buono, ma neanche la dannazione per definizione. È una partita aperta.

[5] cfr. https://www.agendadigitale.eu/documenti/giustizia-digitale/giustizia-predittiva-ecco-i-tre-pilastri-per-capire-gli-impatti-della-tecnologia-sul-diritto [03/03/2021]
[6] cfr. https://youtu.be/dpSpwzyO0vU?t=407 [18/02/2021]

4. L'elemento umano

4.1 Fidarsi troppo

Lo studio dell'interazione umana con i DSS è un fattore essenziale nella progettazione di sistemi che siano contemporaneamente efficienti ed equi. Nel sostituirsi ad alcune attività umane, infatti, l'automazione modifica significativamente, in maniera spesso imprevista e non intenzionale, non solo il comportamento degli umani ma anche le stesse attività. Un fattore cognitivo che condiziona l'interazione uomo-macchina è l'Automation Bias (condizionamento dell'automazione), cioè la tendenza dell'essere umano a porre eccessiva fiducia (*over-reliance*) nei suggerimenti di un supporto tecnologico, anche in presenza di segnali o percezioni in contraddizione con il giudizio della macchina. La decisione viene così operata senza aver analizzato in maniera approfondita tutte le informazioni disponibili (Parasuraman & Manzey, 2010).

L'Automation Bias può provocare diversi tipi di errore. L'*omission error* (errore di omissione) si verifica quando l'operatore non compie tempestivamente l'azione appropriata perché non controlla con sufficiente frequenza/attenzione gli output della macchina (non ci accorgiamo che il correttore automatico non rileva un errore di ortografia) oppure non riconosce un falso negativo della macchina nonostante la presenza di altri segnali di allarme giudicati meno affidabili (rileggendo, confermiamo l'ortografia scorretta anche se "suona" male). Il *commission error* (errore di commissione) è invece quello che si verifica quando l'operatore esegue il suggerimento sbagliato della macchina, cioè non riconosce un falso positivo (accettiamo il suggerimento del correttore per sostituire la sua ortografia sbagliata alla nostra corretta) (Skitka, 2011).

Ci sono molteplici fattori che contribuiscono all'Automation Bias. L'avarizia cognitiva è la tendenza umana a compiere il minor sforzo possibile per prendere una decisione, cercare scorciatoie e risparmiare energie. La raccomandazione di un sistema automatico può quindi diventare il criterio di scelta, rimpiazzando processi più faticosi di analisi e valutazione delle informazioni. C'è poi la *complacency*, ovvero l'inclinazione a credere che i sistemi tecnologici abbiano capacità di analisi superiori

Tabella 4.1 Interazioni tra le prestazioni di un sistema e la risposta dell'utente (Gretton, 2018)

User response	System performance			
	True positive	False positive	True negative	False negative
Agree	Appropriate reliance	Commission errors	Appropriate reliance	Omission errors
Disagree	Under-reliance	Appropriate reliance	Under-reliance	Appropriate reliance

(Lee & See, 2004) e il cosiddetto *Human substitution bias* per cui «qualunque sia il compito, una macchina può farlo meglio di un umano» (Greenhalg, 2013). Un altro pregiudizio storicamente noto è il *Pro-innovation bias* (Rogers, 2010) per cui ogni novità è di per sé migliore dell'esistente. Inoltre, quando compiti decisionali e di sorveglianza si svolgono in collaborazione con una macchina, bisogna considerare anche l'aspetto della dispersione della responsabilità, in maniera simile a quanto avviene quando gli umani collaborano con altri umani: il *Social loafing* (Karau & Williams, 1993) è quel tipo di noncuranza che interviene quando il gruppo di lavoro presenta delle ridondanze e non ci sono responsabilità individuali. Nella misura in cui l'operatore umano percepisce il DSS come un membro della squadra, potrebbe sentirsi meno responsabile della riuscita del compito e quindi ridurre il proprio sforzo nell'analizzare e valutare tutte le informazioni disponibili (Parasuraman & Manzey, 2010).

Ci sono poi caratteristiche sia del sistema che dell'operatore in grado di influenzare l'Automation Bias nella specifica interazione, come l'affidabilità (reale o percepita) del DSS, l'esperienza individuale dell'operatore con altri sistemi automatici, e il suo atteggiamento personale nei confronti della tecnologia. Anche il fenomeno dell'*under-reliance*, ovvero della sfiducia aprioristica nei sistemi automatici, può diventare causa di errore. La Tab 4.1 (Gretton, 2018) riporta uno schema delle possibili interazioni tra le prestazione di un sistema e la reazione dell'operatore.

Quello che va sempre considerato è che, paradossalmente, tanto più il sistema è autonomo e affidabile – o come direbbero in un film: intelligente – maggiore è la difficoltà degli esseri umani nel gestirlo e controllarlo. Ogni volta che si sceglie di avvalersi di un DSS, pertanto, occorre guardare alle sue prestazioni all'interno della dinamica di interazione uomo-macchina, per coltivare aspettative realistiche sui

miglioramenti che si possono ottenere, e valutare l'effettivo costo economico e sociale dell'implementazione.

4.2 Strumenti brillanti, menti ottuse[7]

Alle ore 2:14 UTC della notte tra il 31 maggio e il primo giugno 2009 il volo di linea Air France 447, un Airbus A330-200 in servizio dall'aeroporto Galeão di Rio de Janeiro all'aeroporto Charles de Gaulle di Parigi, precipita nell'oceano Atlantico causando la morte di tutte le 228 persone a bordo: 216 passeggeri, tre piloti e nove assistenti di volo. Il report finale del *Bureau d'enquêtes et d'Analyses pour la sécurité de l'aviation civile* indica come cause dell'incidente errori dei piloti e guasti tecnici (BEA, 2012). Ma il guasto tecnico, di per sé, sarebbe stato trascurabile.

Alle 2:10 UTC a causa del congelamento dei tubi di Pitot – un problema tecnico noto, che aveva già causato qualche incidente ed era in via di risoluzione – i sensori di velocità smettono di funzionare: in presenza di valori incoerenti, il pilota automatico si disattiva. «Una risposta necessaria, minima e logica da parte della macchina», scrive William Langewiesche in un articolo su *Vanity Fair*. Poi prosegue:

> So here is the picture at that moment: the airplane was in steady-state cruise, pointing straight ahead without pitching up or down, and with the power set perfectly to deliver a tranquil .80 Mach. The turbulence was so light that one could have walked the aisles—though perhaps a bit unsteadily. Aside from a minor blip in altitude indication, the only significant failure was the indication of airspeed—but the airspeed itself was unaffected. No crisi existed. The episode should have been a non-event, and one that would not last long. The airplane was in the control of the pilots, and if they had done nothing, they would have done all they needed to do (Langewiesche, 2014).

Da questo momento e per i successivi tre minuti, fino allo schianto nell'oceano, quello che avviene è una tempesta perfetta di errori e disgraziate coincidenze.

Quando si disattiva il pilota automatico, il comandante dell'aereo Marc Dubois sta riposando: in cabina ci sono soltanto i due co-piloti, privi di significativa esperienza di volo non assistito. Il pilota ai comandi, Pierre-Cédric Bonin, colto di sorpresa, probabilmente per rimediare alla perdita di altitudine tira bruscamente indietro la cloche: una reazione «paragonabile al raggomitolarsi istintivo in posizione fetale» (Langewiesche, 2014). È una manifestazione da manuale del cosiddetto *Startle*

[7] Vivek Haldar, sviluppatore di Google (Carr, 2015)

effect: richiamato all'attenzione da un segnale di pericolo inaspettato, l'essere umano subisce una temporanea diminuzione delle abilità cognitive per cui non è in grado di elaborare una risposta. La sorpresa dei due piloti è peraltro particolarmente profonda, perché non hanno idea di come funzioni l'Airbus sotto la superficie dei comandi automatici: negli aerei di "quarta generazione" come l'A330-200 la guida avviene via cavo (*fly-by-wire*) attraverso un sistema di comando elettronico digitale che sostituisce i comandi diretti, meccanici e idraulici. Senza feedback sensoriali, e con un'idea molto approssimativa della complessità attivata dalle loro decisioni, al di fuori della normale routine i piloti non hanno elementi per potersi fidare della macchina.

Per una discussa scelta di progetto dell'Airbus, inoltre, i comandi dei due piloti non sono collegati tra loro e non si muovono all'unisono. Questo significa che il secondo pilota, David Robert, non ha idea né della manovra effettuata né del cambiamento di assetto del velivolo, il cui muso è puntato all'insù. Ma in aerodinamica la portanza sviluppata (cioè la forza per sostenersi in volo) cresce al crescere dell'angolo di incidenza fino a un valore massimo, detto angolo di stallo, superato il quale la portanza diminuisce velocemente, e l'aereo precipita.

> Almost as soon as Bonin pulls up into a climb, the plane's computer reacts. A warning chime alerts the cockpit to the fact that they are leaving their programmed altitude. Then the stall warning sounds. This is a synthesized human voice that repeatedly calls out, "Stall!" in English, followed by a loud and intentionally annoying sound called a "cricket." A stall is a potentially dangerous situation that can result from flying too slowly. At a critical speed, a wing suddenly becomes much less effective at generating lift, and a plane can plunge precipitously. All pilots are trained to push the controls forward when they're at risk of a stall so the plane will dive and gain speed. The Airbus's stall alarm is designed to be impossible to ignore. Yet for the duration of the flight, none of the pilots will mention it, or acknowledge the possibility that the plane has indeed stalled—even though the word "Stall!" will blare through the cockpit 75 times. Throughout, Bonin will keep pulling back on the stick, the exact opposite of what he must do to recover from the stall (Wise, 2011).

Per uscire dallo stallo, in questo momento, sarebbe sufficiente diminuire l'angolo di incidenza, ma il pilota ai comandi non lo fa, e il secondo pilota non ha contezza della situazione. L'angolo di incidenza, inoltre, non è tra le informazioni visualizzate sul pannello di comando. Ma la comunicazione tra i piloti è insufficiente, l'interfaccia scarsamente informativa, la gerarchia di comando poco chiara come la distribuzione delle responsabilità. E piuttosto incredibilmente, dalle registrazioni della scatola nera sembra che nessuno noti il segnale di allarme, che ripeterà la parola *STALL* ben 75

volte. Quando alle 02.11 UTC il comandante rientra in cabina, i due co-piloti non riescono a fornire una spiegazione di quello che sta succedendo. Ma invece di prendere posto ai comandi per avere un'esperienza diretta della situazione, Dubois si siede dietro ai co-piloti. La cloche di Bonin è ancora tirata tutta indietro, ma nessuno se ne accorge.

> 02:13:40 (Robert) *Remonte... remonte... remonte... remonte...*
>
> 02:13:40 (Bonin) *Mais je suis à fond à cabrer depuis tout à l'heure!*
>
> At last, Bonin tells the others the crucial fact whose import he has so grievously failed to understand himself.
>
> 02:13:42 (Captain) *Non, non, non... Ne remonte pas... non, non.*
>
> 02:13:43 (Robert) *Alors descends... Alors, donne-moi les commandes... À moi les commandes!*
>
> Bonin yields the controls, and Robert finally puts the nose down. The plane begins to regain speed. But it is still descending at a precipitous angle. As they near 2000 feet, the aircraft's sensors detect the fast-approaching surface and trigger a new alarm. There is no time left to build up speed by pushing the plane's nose forward into a dive. At any rate, without warning his colleagues, Bonin once again takes back the controls and pulls his side stick all the way back.
>
> 02:14:23 (Robert) *Putain, on va taper... C'est pas vrai!*
>
> […]
>
> Exactly 1.4 seconds later, the cockpit voice recorder stops. (Wise, 2011).

La considerazione immediata che si ricava dalla cronaca dell'incidente del volo AF447 è che, con un velivolo di quarta generazione e in assenza di criticità, fatti salvi decollo e atterraggio, praticamente chiunque può portare un aereo da Rio de Janeiro a Parigi senza problemi: non sono richieste particolari abilità di controllo manuale, né un'accorta vigilanza. Per questa ragione, negli anni, il numero di componenti dell'equipaggio in cabina è passato da tre a due, le competenze richieste a un pilota sono drasticamente diminuite, i corsi e le ore di addestramento obbligatorio si sono ridotti di conseguenza. D'altra parte, l'implementazione dei piloti automatici ha ridotto in maniera sostanziale il numero di incidenti in volo (Fig. 4.1). Ma se la maggior parte degli eventi viene gestita e risolta senza che l'operatore debba

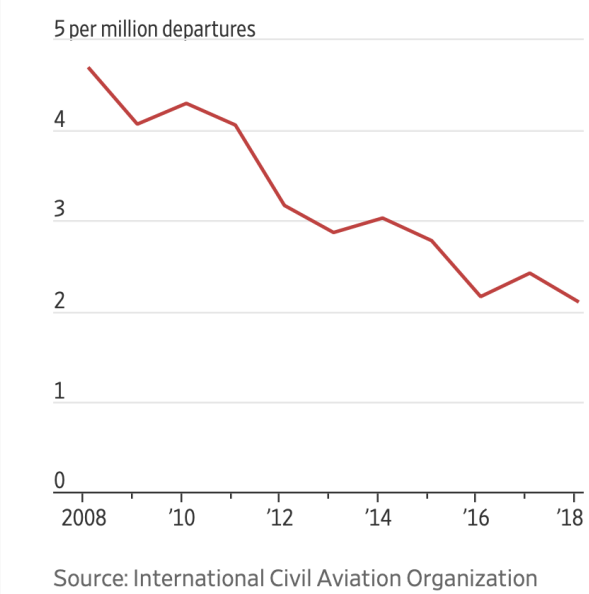
intervenire, quello che rimane è l'imprevedibile, e pochissima esperienza per affrontarlo.

Il deskilling (dequalificazione) è la riduzione del livello di competenza richiesto per svolgere una funzione quando tutte o alcune delle componenti dei compiti corrispondenti sono state automatizzate. Un fenomeno che diventa evidente quando la tecnologia fallisce, o smette anche solo temporaneamente di funzionare, ma che può avere conseguenze negative anche in assenza di errori macroscopici (Cabitza et al., 2017). Non si tratta di una scoperta recente: negli anni Cinquanta del Novecento James Bright, un professore dell'Harvard Business School che studia gli effetti concreti dell'automazione in diversi tipi di industrie, scopre che, nella maggior parte dei casi, l'inserimento delle macchine lascia agli umani un lavoro meno impegnativo e più noioso da svolgere.

> Bright concluded that the overriding effect of automation was (in the jargon of labor economists) to "de-skill" workers rather than to "up-skill" them. "The lesson should be increasingly clear," he wrote in 1966. "Highly complex equipment" did not require "skilled operators. The 'skill' can be built into the machine." (Carr, 2014).

Queste considerazioni possono essere estese a qualunque compito il cui svolgimento sia supportato da sistemi automatici oracolari: dalla medicina all'architettura, dall'attività delle forze dell'ordine all'aviazione, dall'industria manifatturiera alla ricerca, dalla guida autonoma delle automobili fino alla possibilità di girare per una città sconosciuta, anche a piedi, senza l'aiuto di un sistema di navigazione.

Fig. 4.1 Tasso di incidenti aerei per voli commerciali (Pastzor & Wall, 2019)



4.3 Il centauro

Il pericolo del deskilling non è tanto la perdita di alcune abilità, che potrebbero essere aggiornate o sostituite con nuove competenze (*upskilling* e *reskilling*) quanto la perdita di centralità dell'esperienza umana nel processo cognitivo, che sia creativo o decisionale. È infatti conseguenza intrinseca di un'idea di

progettazione che pone l'attenzione principale sulle capacità della macchina, per sfruttarle al meglio, lasciando l'umano a occuparsi di quello che resta. In questo modo l'operatore sarà sempre meno coinvolto e vigile, cioè più portato all'errore, innescando un circolo vizioso di automazione/dequalificazione.

Un aspetto critico della dequalificazione professionale indotta dall'*over-reliance* nell'automazione è la desensibilizzazione al contesto, e a tutte le informazioni che in generale non possono essere quantificate come input di un sistema oracolare. In medicina, per esempio, gli aspetti psicologici, relazionali, sociali e organizzativi rimangono essenziali nella valutazione olistica di un paziente, ma sono molto difficili da incorporare in un ML-DSS per via della loro complessa natura qualitativa. Occorre poi considerare l'incertezza intrinseca delle osservazioni cliniche, e quindi progettare sistemi in grado di tenerne conto, adattandosi alla natura del compito e dei dati rilevati, anziché pretendere il viceversa (Cabitza et al., 2017).

Un altro elemento determinante nella realizzazione di un'efficace interazione uomo-macchina è un'interfaccia di utilizzo adeguata. Le caratteristiche di progettazione dell'interfaccia e dell'architettura dell'informazione, infatti, possono non solo provocare errori, ma anche influenzare il processo decisionale umano, nel bene e nel male (Gretton, 2018). Nell'analisi dell'incidente del volo AF447 (cfr. 4.1) molti osservatori hanno riconosciuto in alcune scelte di progettazione dell'Airbus A330-200 una oggettiva responsabilità nell'incidente, e tra le misure di ergonomia cognitiva allo studio per aumentare la sicurezza c'è l'introduzione di comandi con sistemi di feedback aptico, cioè tattile, come già succede sui velivoli Boeing, di impostazione più antropocentrica. Sono gli stessi comandi che da anni vengono usati dalle console per rendere più immersiva l'esperienza di gioco: i videogame offrono infatti un valido modello di progettazione *human-centered*, mirato a incoraggiare e accrescere la destrezza del gamer.

Bisogna infine valutate con molta attenzione il livello di trasparenza con cui si intende progettare un sistema predittivo di supporto alle decisioni. Il funzionamento a scatola nera è spesso legato a un'erronea attribuzione di fiducia, sia in termini di *under-reliance* che di *over-reliance*, oltre a rappresentare un fattore decisivo nel conferimento della responsabilità legale in caso di errore. Tanto più un modello predittivo di ML è accurato, tanto più raramente risulta chiaro il percorso che

conduce alle predizioni. L'opacità del sistema pone l'operatore in condizioni di non disporre di tutte le informazioni necessarie a prendere decisioni, ma la complessità dell'algoritmo rende l'elaborazione, anche qualora esplicitata, pressoché impossibile da interpretare, a maggior ragione nell'esecuzione di compiti ad alto rischio e/o in tempi stretti. Va inoltre considerato che una persona che non riesce a capire una macchina può sentirsi frustrata, perdere l'autostima, e quindi tendere a non intervenire in caso di bisogno. Una soluzione potrebbe essere quella di mettere la macchina in grado di fornire tempestivamente una spiegazione dinamica della predizione, unita alla possibilità di esplorare le conseguenze di modifiche alle variabili più rilevanti (Cabitza et al., 2017).

Coinvolto costantemente nel ciclo decisionale di giudizio-azione-feedback, l'essere umano potrebbe avvalersi della percezione aumentata della macchina sia in termini di informazioni che di prospettiva, conservando però un ruolo attivo nello svolgimento del compito, delegando alla macchina attività di routine e sorveglianza, e ponendosi con atteggiamento critico nei confronti delle "predizioni" automatiche. Si tratterebbe pertanto di un'alleanza tra essere umano e macchina in cui ognuna delle parti contribuisce con quello che sa fare meglio. Quando Garry Kasparov ha proposto questo modello collaborativo per il gioco degli scacchi – i giocatori possono consultare il computer in qualunque momento, ma sono loro a decidere ogni mossa – lo ha chiamato "Centauro": una combinazione ibrida che è in grado di battere ogni altro essere umano (anche di livello scacchistico superiore) ma anche ogni altro computer (Case, 2018).

4.4 L'etica dei sistemi

Anche nell'ipotesi di funzionamento virtuoso della diade uomo-macchina, rimarrebbe comunque da sciogliere un nodo etico: come attribuire la responsabilità morale all'interno di un sistema al cui funzionamento concorrono operatori umani e programmi automatici? Per investigare le implicazioni etiche dell'interazione umana con sistemi di ML abbiamo intervistato Viola Schiaffonati, professore associato di Logica e filosofia della scienza al Politecnico di Milano[8].

[8] La conversazione con la prof. Schiaffonati, avvenuta con la sottoscritta via Zoom il 12/02/21, è stata di seguito sintetizzata e redatta. Le domande sono riportate in neretto.

Di chi è la responsabilità morale del funzionamento di un sistema automatico?

Ci sono due cose importanti da tenere in considerazione. La prima è che bisogna ripensare il concetto di responsabilità morale. Fin qui siamo stati abituati a pensare che la responsabilità morale potesse essere attribuita solo a persone, mentre oggi la grande domanda è se almeno parte di essa possa essere attribuita ad artefatti. La discussione è molto aperta, e non c'è una risposta definitiva. Ma ci sono delle indicazioni. Nel 2017 in Germania è stato istituito un comitato etico sulle auto a guida autonoma[9], formato da persone di varia provenienza, per cercare di ragionare e regolamentare. Il risultato di questa discussione non è stato che la responsabilità sia da attribuire in particolare a qualcuno – il produttore, il progettista, il proprietario – ma che bisogna valutare caso per caso, nelle specifiche situazioni. Le riflessioni etiche su questi temi non hanno un'unica risposta, perché l'etica è un processo complesso, che richiede la messa a terra nella specifica situazione: può fornire un orientamento, non certo una risposta precisa. La seconda cosa importante, quando ragioniamo su tecnologie che hanno un forte impatto sulla vita degli esseri umani, è che bisogna proprio cambiare il paradigma di riferimento della responsabilità. Fino a pochi anni fa la responsabilità era un concetto che veniva visto in maniera "passiva": quando succedeva qualcosa di negativo con un prodotto tecnologico, si andava a ritroso per capire di chi fosse la responsabilità ricostruendo la catena degli eventi e dei contributi di tutte le entità coinvolte. Ma ora che abbiamo tecnologie il cui impatto sulla società è difficile da prevedere, questa visione deve essere modificata: si preferisce parlare di approccio attivo. Significa che il tema della responsabilità va affrontato fin dall'inizio, in fase di progettazione, non solo per evitare effetti negativi, ma anche per promuovere effetti positivi. Occorre un cambio di paradigma.

Questo cambio di paradigma comprende anche la possibilità di inserire un sistema etico direttamente in fase di programmazione?

Più che programmare, direi: di pensare a come queste macchine possano incorporare alcuni valori come requisiti – pensiamo per esempio al caso della privacy, che può

[9] cfr. https://www.lastampa.it/motori/tecnologia/2017/08/24/news/germania-gli-esperti-dettano-le-leggi-per-la-guida-autonoma-la-vita-umana-prima-di-tutto-1.34440707 [19/02/2021]

essere incorporata *by design*. È un tema interessante, che rileggerei in una chiave un po' diversa. La prospettiva che di solito viene offerta è quella delle macchine che ci sostituiranno anche nel ragionamento morale. Io non credo che questo sia possibile. Ritengo che la moralità, in senso ampio, sia una prerogativa umana: sono gli esseri umani a progettare le macchine avendo in mente certi valori. E anche in questo caso, non c'è una soluzione unica. Consideriamo il caso di una macchina a guida autonoma, tra i valori che vogliamo inserire c'è sicuramente la sicurezza. Ma la sicurezza di chi? Del passeggero? Del proprietario dell'auto? Degli altri conducenti? Oppure dei pedoni? È importante anticipare questi problemi non nell'ottica di rendere questi sistemi nostri sostituti nel campo della decisione morale, ma di considerarli realizzazioni dei valori che noi decidiamo di inserire.

Come costruisce una cultura deontologica negli sviluppatori?

Io sono una filosofa della scienza e della tecnologia, e lavoro al Dipartimento di ingegneria informatica del Politecnico di Milano proprio nella direzione di rendere i futuri progettisti più consapevoli delle questioni morali in atto, ma certo un corso non può essere sufficiente. L'obiettivo di questi ibridazioni peraltro non è far diventare i progettisti degli esperti di etica, ma renderli consapevoli del problema e capaci di lavorare in gruppi multidisciplinari, perché quello che si cerca di fare in questi contesti è coinvolgere fin dalla fase di progettazione esperti di diverse discipline. È anche importante che le grandi associazioni contribuiscano a diffondere una cultura diversa. Per esempio la ACM – *Association for Computing Machinery*, una delle più grandi associazioni di professionisti informatici al mondo – ha rivisto il suo codice etico[10], che risaliva agli anni Novanta, e ne ha rilasciata nel 2018 una versione sostanzialmente rivista in cui i temi dell'intelligenza artificiale, della privacy e della sorveglianza sono elementi di riflessione nello stabilire linee guida per i progettisti.

Qual è il ruolo dell'utente finale in questo processo di consapevolizzazione?

L'aumento di consapevolezza da parte dell'utente è uno dei temi forse un po' lasciati indietro nel dibattito odierno, ma è importantissimo almeno su due livelli. Il primo è ancora che i fruitori finali di queste tecnologie devono essere coinvolti nei processi di

[10] cfr. https://www.acm.org/code-of-ethics [19/02/2021]

progettazione. Non perché ognuno diventi esperto di tutto, ma perché è cruciale avere da subito una prospettiva multidisciplinare, con diverse visioni tenute in considerazione, e senza dubbio gli utenti giocano un ruolo chiave. Il secondo aspetto è l'attenzione del dibattito pubblico affinché questo diventi un tema di cittadinanza, di responsabilità di noi tutti all'interno di un contesto democratico. Non dobbiamo cadere in quell'idea di deresponsabilizzazione per cui si pensa che le tecnologie si sviluppano senza che noi possiamo fare nulla, come entità a sé stanti. La tecnologia non si sviluppa da sola: la tecnologia siamo noi, e la delega della responsabilità morale non può essere accettata. Resistono le nostre prerogative di decisori morali: in quanto parte attiva delle società democratiche non abbiamo solamente dei diritti, ma anche i doveri di conoscere, di intervenire, di partecipare. Ovviamente questo richiede un certo sforzo e – di nuovo – un cambio di mentalità.

Il Machine Learning, funziona per induzione: estrapolando le regole dai dati. Questo potrebbe modificare il modo di fare e di intendere la ricerca scientifica?

Sì, anche la ricerca è uno dei campi che in questo contesto è messa in discussione e si mette in discussione. Ed è proprio quando ci sono cambiamenti così grandi che dobbiamo mantenere un'attenzione attiva. Ai miei studenti di ingegneria io dico sempre che, oltre a tutte le competenze tecniche, hanno bisogno di sviluppare capacità come l'immaginazione. È difficile da insegnare, ma è fondamentale per cercare di anticipare quali strade possono essere prese, e quali conseguenze avranno le strade che prenderemo. Dobbiamo rimetterci al centro come esseri umani: come progettisti, come ricercatori, come cittadini. Non è un caso che oggi si parli molto di umanesimo digitale, ovvero di quell'attitudine, quella visione che rimette l'essere umano al centro della tecnologia. Sia nella progettazione – dove occorre tenere ben saldo il timone delle scelte – sia nella fruizione: le tecnologie vanno pensate al servizio delle persone, mai disincarnate da quello che possono offrire. Quell'entusiasmo tecnologico che fino a qualche anno fa veniva considerato un elemento esaltante e fondamentale dell'innovazione, oggi deve sposarsi col rimettere l'essere umano al centro. Noi in Italia siamo la culla dell'umanesimo, e forse potremmo ricoprire un ruolo in questo senso.

4.5 Altri istupidimenti

Gli algoritmi di Machine Learning non sovrintendono soltanto alla guida di macchine o aerei, alla selezione del personale, alla diagnosi medica o alle sfide di AlphaGo. Sono alla base di molte delle nostre attività quotidiane, soprattutto se si pensa che durante la pandemia di Covid-19 la maggior parte delle nostre attività professionali e sociali si sono trasferite online, e pertanto sono spesso oggetto dei filtri di rilevanza e visibilità che gli algoritmi di profilazione e personalizzazione dei social network e dei motori di ricerca applicano ai contenuti. È quindi lecito chiedersi quali possano essere le modifiche che questi interventi apportano alla natura delle attività.

> Google ammette [...] di aver assistito a un effetto di istupidimento del grande pubblico nel momento in cui ha reso il proprio motore più reattivo e sollecito, più abile nel prevedere che cosa la gente stia cercando. Google non si limita a correggere i nostri errori ortografici; ci suggerisce i termini di ricerca mentre digitiamo sulla tastiera, scioglie le ambiguità semantiche presenti nelle nostre richieste, e anticipa le nostre necessità basandosi sul luogo in cui ci troviamo e su come ci siamo comportati in passato. Potremmo presumere che quanto più Google ci aiuta ad affinare le nostre ricerche, tanto più noi impariamo dal suo esempio. Dovremmo diventare più precisi nel formulare parole chiave e affinare in altro modo le nostre esplorazioni online. Ma, stando a Amit Singhal, il più importante ingegnere specializzato in ricerca di Google, accade il contrario. Nel 2013 un reporter dell' *Observer* di Londra intervistò Singhal sui numerosi miglioramenti che erano stati apportati al motore di ricerca di Google nel corso degli anni. "Presumibilmente", notava il giornalista, "con l'aumento dell'uso di Google siamo diventati più accurati nella scelta delle parole che inseriamo nel motore di ricerca". Singhal sospirò e, "con una certa stanchezza nella voce", corresse quanto stava dicendo l'intervistatore: "In realtà è proprio il contrario. Più la macchina è precisa, più noi diventiamo pigri nel formulare le nostre domande". (Carr, 2015)

A questo proposito abbiamo intervistato Federico Cabitza, professore associato di Sistemi informativi e Interazione uomo-macchina all'Università di Milano-Bicocca[11].

Come si riesce a minimizzare il rischio degli errori umani legati all'interazione con la macchina?

[11] La conversazione con il prof. Cabitza, avvenuta con la sottoscritta via Google Meet il 16/02/21, è stata di seguito sintetizzata e redatta. Le domande sono riportate in neretto.

Esattamente come hanno escogitato nell'ambito dell'aviazione: con la ridondanza, soprattutto umana. Non è un caso che negli aeroplani ci siano due piloti, nonostante ci siano state diverse richieste di riduzione del personale per risparmiare denaro, soprattutto su quelle tratte che funzionano al 99% col pilota automatico. Un tempo ce ne erano addirittura tre: si è ridotto da tre a due, ma è importante non ridurre da due a uno. Una posizione che espongo in un capitolo di un libro che uscirà tra poche settimane per i tipi di MIT Press[12] è che bisognerebbe concepire protocolli nell'uso dell'intelligenza artificiale in cui si richiede che l'intelligenza artificiale venga usata sempre e solo in contesti di lavoro di gruppo, quindi in contesti sociali, dove magari si stabilisce anche una gerarchia, ma dove però tutti possono obiettare, con una disciplina piuttosto orizzontale che verticale. Quando elementi di AI vengono integrati in un'attività si crea tutt'una serie di aspettative di attendibilità, oggettività e infallibilità che portano al rischio di *complacency* – compiacenza nei loro confronti; di un abbassamento del nostro livello di attenzione; di eccessiva fiducia o affidamento. C'è bisogno che ci sia sempre qualcuno che possa alzare la mano e dire: mah, io non sono mica tanto convinto. Anche la macchina, programmata in maniera opportuna e un po' diversamente da come si fa di solito, potrebbe contribuire a questo modo di ragionare insieme. In ambito medico, per esempio, io mi oppongo alla condizione per cui c'è un sistema informatico che dà un suggerimento all'essere umano, e ho dimostrato con qualche studio come effettivamente si tratti di una situazione che presenta forti rischi. È evidente che poi, dal punto di vista della responsabilità, sarà l'essere umano ad avere l'ultima parola, ma rimane una dinamica di coppia, una diade. Anche se può sembrare meno efficiente, paradossalmente potrebbe risultare più efficace, sul lungo periodo, porre la macchina all'interno di un triumvirato in cui si trovi sempre in minoranza rispetto al gruppo umano. E questo richiede di pensare un po' diversamente. D'altro canto ho appena pubblicato un lavoro[13] in cui sono riuscito a dimostrare un fenomeno che aveva notato Garry Kasparov, dieci anni fa, nell'ambito del gioco degli scacchi. La cosiddetta Legge di Kasparov dice che esseri umani dalle competenze non perfette funzionano come gruppo meglio di un esperto – una tipica congettura della

[12] https://mitpress.mit.edu/books/machines-we-trust [19/02/2021]
[13] cfr. https://link.springer.com/article/10.1007/s13755-021-00138-8 [19/02/2021]

collective intelligence, un ramo dello studio del fenomeno dell'intelligenza animale – qualora il processo che li fa interagire fosse migliore di quello concepito per l'azione del singolo esperto. Ovvero: umani deboli con una macchina con un processo migliore sono più forti di umani forti con una macchina con un processo più stupido. Il processo è più importante della gente: come li fa lavorare o interagire è più importante sia della macchina che dei singoli decisori. Quindi oltre ad aumentare la ridondanza decisionale, io ribadisco che è opportuno che l'intelligenza artificiale sia un'occasione per rendere le decisioni una espressione di un gruppo più che dei singoli.

Com'è cambiato il nostro modo di conoscere da quando usiamo Google?

In ambito psicologico esiste il concetto della memoria transattiva: la capacità di recuperare informazioni accedendo a una memoria espressa da un gruppo di attori, che può comprendere l'agente computazionale. La memoria transattiva non sta dentro la testa di una singola persona, ma viene espressa dalla conversazione e dall'interazione: se Google funziona non è perché abbiamo lo strumento informatico, ma perché ci si affida a una unità sovraindividuale: lo strumento digitale ci fornisce una sorta di protesi per avere una memoria transattiva eccezionale. Il problema nasce quando i singoli componenti tendono ad affidarsi al gruppo molto più di quanto fossero abituati in precedenza, e nel gruppo ci sono quasi esclusivamente componenti informatiche. Questo può portare a un depotenziamento – un *deskilling* – della nostra capacità di ricordare. La tecnologia tanto ci dà quanto ci toglie: da una parte con Google abbiamo la capacità straordinaria di rispondere quasi istantaneamente a quasi ogni domanda, una memoria transattiva su un repository di conoscenza sterminato; dall'altra perdiamo il piacere di sapere le cose, perché costa fatica impararle e tenerle la mente, ed è una fatica che con ennemila dispositivi digitali è assolutamente fine a se stessa. Bisogna trovare un compromesso. Una delle tante proposte che sono state fatte, tra cui anche una mia, è di inserire nei sistemi delle *inefficienze programmate* per preservare capacità che è strategico mantenere. Pensiamo all'orientamento: tutti i navigatori cellulari e satellitari dipendono da un sistema militare statunitense che dall'oggi al domani potrebbe essere spento. In questi sistemi, allora, potremmo introdurre una inefficienza, per ricominciare a capire

come orientarsi in una città, ma anche per avere nuova dimestichezza con il fatto di chiedere a qualcuno. Tutte queste tecnologie della grande socialità in realtà ci isolano, e possiamo perdere anche l'abitudine a fare affidamento sui simili. È una possibilità che per adesso viene indagata in maniera ancora un po' accademica, ma si diffonderà quando il deskilling coinvolgerà molti più campi di quelli che possiamo immaginare adesso, e riguarderà capacità rilevanti. In medicina, per esempio, è critico saper fare una diagnosi: affidandosi alle macchine, nel medio-lungo periodo, questa capacità si perde. O si perde subito, se la macchina è affiancata a uno specializzando, perché quel muscolo non lo sviluppa mai. Ma bisogna intervenire prima che sia troppo tardi, poiché la capacità di autoregolarsi della comunità scientifica e professionale è nulla, e certe cose vanno imposte anche a livello di regolamenti, normative o certificazioni.

Durante la pandemia abbiamo spostato online quasi tutte le attività sociali, sottoponendole in gran parte al filtro dei social network. Questo intacca la nostra capacità di selezionare persone e informazioni?

Sì, qui abbiamo proprio un meccanismo di selezione dei contenuti, e quindi anche di evidenziazione di chi crea certi contenuti piuttosto che altri, in grado di creare vere e proprie discriminazioni. I sociologi conoscono bene l'effetto San Matteo: chi più ha, in questo caso chi più dice, più riceve, cioè più gli viene data la possibilità di esprimersi. Oltre a questa dinamica abbastanza universale, ci sono poi interventi manipolatori da parte degli algoritmi proprio per spostare l'attenzione verso certi contenuti e verso certe fonti per renderle più o meno autorevoli. Non ci sono grandi soluzioni, se non fare a meno di questi sistemi oppure ripensarli con dinamiche non orientate al profitto che forse potrebbero attuare un filtro diverso. Ma questi strumenti non sono che amplificatori dei meccanismi naturali: nel momento in cui si frequenta una determinata rete sociale – un *social network* – si è dentro una camera d'eco, una bolla di filtraggio in cui è molto facile sentire sempre le stesse opinioni, pensare che tutti la pensino come te, e finire per credere semplicemente a ciò che la gente condivide o ri-twitta. È difficile pensare a un antidoto: si potrebbero abolire d'ufficio, come in America hanno minacciato di fare con TikTok o altre applicazioni cinesi: si può fare. Però siamo una società libertaria, individualista, capitalista e produttivista: non avvertiamo i pericoli finché non ci vengono addosso. E anche

quando ci vengono addosso, come per esempio la pandemia, cerchiamo un modo di conviverci pur di non cambiare comportamenti.

Esiste il pericolo per la democrazia che qualcuno approfitti di queste distorsioni per legittimare la propria autorità?

Se ne approfittano le élite, chi ha il potere economico e politico, per mantenerlo o per aumentarlo. Peraltro oggi è difficile sovvertire il potere: mentre un tempo bastava convincere qualche militare, adesso gli interessi sono molto intrecciati, e il funzionamento della nostra società troppo complesso. Abbiamo raggiunto un certo livello di benessere, ma sono aumentate le differenze tra chi sta molto bene e chi sta molto male: va bene così, chiaramente, a chi sta nella metà giusta del mondo, ma se ti trovi nella metà sbagliata, non te la cavi. E il peggio deve ancora venire: la metà sbagliata nel mondo diventerà il 95% sbagliato del mondo che, anche semplicemente coi flussi migratori, minaccerà il restante 5%. Ma se quel 5% avrà droni, armi intelligenti, sistemi di monitoraggio e sorveglianza, che cosa potrà mai fare il 95%? La resistenza è futile, si diceva un tempo. Io non sono un pessimista, perché spero che ci sia ancora qualche anno di benessere, ma non è possibile avere fiducia nella scienza e nella tecnologia quando non ci sono finanziamenti, e soprattutto non si costruisce un tessuto sociale in grado di capirla e di integrarne gli spunti nei comportamenti di tutti i giorni.

Può servire intervenire sulla comunicazione e sull'istruzione?

Assolutamente sì. Quello che faccio io, nel mio piccolo, è proprio disseminare quello che ho imparato sui bias, quanto siano vicini a noi e connaturati al nostro rapporto con la tecnologia. Ma stiamo comunque parlando a una nicchia, una bolla assolutamente minoritaria. Bisognerebbe partire dalle elementari per abituare al pensiero critico, a quei ragionamenti razionali che non sono esclusivi della scienza ma comuni a molte discipline umanistiche, come la filosofia o le scienze giuridiche. Ma non vedo un investimento sulla cultura, una ristrutturazione della scuola e dell'università, un nuovo atteggiamento nei confronti di ricerca e sviluppo. Mi sembra che siamo una società che non vuole investire nel futuro: ci sta bene il presente, e dopo di noi il diluvio.

Conclusioni

Conseguenze inattese legate all'utilizzo dei sistemi automatici per il supporto alle decisioni: una questione complessa, che richiede un approccio multidisciplinare. Ma anche una questione complicata, la cui comprensione presume una minima familiarità con alcuni concetti tecnici e teorici, indispensabili a liberare l'argomento da quell'aura di suggestiva ineluttabilità – oppure al contrario: di entusiasmo incondizionato – creata in decenni di produzione letteraria, cinematografica e televisiva sull'argomento.

Si è scelto, quindi, di mantenere un approccio competente, realizzato nelle interviste originali a specialisti in diversi ambiti di interesse, e completato dagli stralci di interventi d'archivio di alcuni autori di riferimento, selezionati in base a criteri di sintesi e di chiarezza. Si è provato inoltre a movimentare la narrazione con segmenti tratti da produzioni di cultura popolare scelti sia tra quelli di argomento affine, per approfittare dell'immediato rimando, sia tra titoli di altro genere, allo scopo di arricchire o contestualizzare un richiamo. È stata sviluppata una parte musicale originale che richiama melodie selezionate per capacità evocativa. Il target è un'ascoltatrice/ascoltatore adulto, di estrazione culturale medio/alta, non specialista in tecnologia. L'obiettivo è quello di costruire un sistema di riferimento coinvolgente, in cui l'utente possa sentirsi a proprio agio, ricevendo l'impressione che scienza e tecnologia siano argomenti rilevanti nella sua realtà, e integrati alla sua esperienza di mondo.

Per la rilevanza che il tema riveste nel presente e nel futuro della società, si ritiene che un'informazione rigorosa sull'argomento sia da rendere accessibile a un pubblico il più vasto possibile, anche attraverso mezzi di comunicazione veloci e destinati a svolgere funzioni di sottofondo/intrattenimento come il podcast.

Bibliografia

BEA. (2012) *Final Report On the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro - Paris.* Available from: https://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf [Accessed 31/01/21].

Cabitza, F., Rasoini, R., Gensini, G.F. (2017) Unintended Consequences of Machine Learning in Medicine. *The Journal of the American Medical Association*, 318(6) pp. 517-518. Available from: doi.org/10.1001/jama.2017.7797.

Carr, N. (2017) A Brutal Intelligence: AI, Chess, and the Human Mind. *Los Angeles Review of Books*. Available from: https://lareviewofbooks.org/article/a-brutal-intelligence-ai-chess-and-the-human-mind/ [Accessed 10/01/21].

Carr, N. (2015) *La gabbia di vetro. Prigionieri dell'automazione*. Translated by S. Garassini and G. Romano. Milano, Raffaello Cortina.

Carr, N. (2014) Automation Makes Us Dumb. *The Wall Street Journal*. Available from: https://www.wsj.com/articles/automation-makes-us-dumb-1416589342 [Accessed 10/01/21].

Case, N. (2018) How To Become A Centaur. *Journal of Design and Science*. Available from: doi.org/10.21428/61b2215c.

Crawford, K. (2013) The Hidden Biases in Big Data. *Harvard Business Review*. Available from: https://hbr.org/2013/04/the-hidden-biases-in-big-data [Accessed 20/01/21].

Croft, J. (2013) Wiener's Laws. *Aviation Week*. Available from: https://aviationweek.com/wieners-laws [Accessed 19/02/21].

Domo. (2020) Data Never Sleeps 8. Available from: https://www.domo.com/learn/data-never-sleeps-8 [Accessed 22 gen 2021].

Graber, M. L. (2013) The incidence of diagnostic error in medicine. *BMJ Quality & Safety*. 22, pp. ii21-ii27, Available from: doi.org/10.1136/bmjqs-2012-001615.

Greenhalg, T. (2013) Five biases of new technologies. *British Journal of General Practice*. 63 (613), p. 425. Available from: doi.org/10.3399/bjgp13X670741.

Gretton, C. (2018) Trust and Transparency in Machine Learning-Based Clinical Decision Support. In: Zou J. & Chen F. (eds.) *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, Human-Computer Interaction Series. Springer International Publishing, pp. 279-292.

Hao, K. (2018) What is machine learning? *MIT Technology Review*. Available from: https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart [Accessed 20/01/21].

Harnad, S. (2008) The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence (PUBLISHED VERSION BOWDLERIZED). In:

Epstein, Robert, Roberts, Gary and Beber, Grace (eds.) *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Evolving Consciousness. Springer, pp. 23-66. Available from: https://eprints.soton.ac.uk/262954/ [Accessed 20/02/21].

Hawkins, A. J. (2020) Waymo pulls back the curtain on 6.1 million miles of self-driving car data in Phoenix. *The Verge*. Available from: https://www.theverge.com/2020/10/30/21538999/waymo-self-driving-car-data-miles-crashes-phoenix-google [Accessed 20/02/21].

Hawkins, A. J. (2016) Inside Waymo's strategy to grow the best brains for self-driving cars. *The Verge*. Available from: https://www.theverge.com/2018/5/9/17307156/google-waymo-driverless-cars-deep-learning-neural-net-interview [Accessed 22/01/21].

Karau, S. J, & Kipling D. W. (1993) Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*. 65 (4), pp. 681-706. Available from: doi.org/10.1037/0022-3514.65.4.681.

Kasparov, G. (2019) *Deep thinking. Dove finisce l'intelligenza artificiale, comincia la creatività umana*. Translated by Valentina Nicoli, ebook ed., Fandango.

Kasparov, G. (2017) Don't fear intelligent machines. Work with them. *TED*. Available from: https://www.ted.com/talks/garry_kasparov_don_t_fear_intelligent_machines_work_with_them [Accessed 25/01/21].

Langewiesche, W. (2014) The Human Factor. *Vanity fair*. Available from: https://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash [Accessed 20/02/21].

Levy, S. (2017) What Deep Blue Tells Us About AI in 2017. *Wired*. Available from: https://www.wired.com/2017/05/what-deep-blue-tells-us-about-ai-in-2017 [Accessed 20/02/21].

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, 27(4). Available from: doi.org/10.1609/aimag.v27i4.1904.

National Transportation Safety Board. (2019) Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian. *National Transportation Safety Board*. Available from: https://www.ntsb.gov/investigations/AccidentReports/Pages/HAR1903.aspx [Accessed 15/02/21].

O'Neil, C. (2017) *Armi di distruzione matematica. Come i Big Data aumentano la disuguaglianza e minacciano la democrazia*. Trans. by D. Cavallini. Firenze, Bompiani.

ORCAA. (2018) *O'Neil Risk Consulting & Algorithmic Auditing*. Available from: https://orcaarisk.com [Accessed 29/01/21].

Parasuraman R., Sheridan T. B., Wickens C. D. (2000) A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30(3), pp. 286-297. Available from: doi.org/10.1109/3468.844354.

Parasuraman, R. & Dietrich M. (2010) Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3), pp. 381-406. Available from: 10.1177/0018720810376055.

Pastzor, A. & Robert W. (2019) Airline Automation Triggers Intensified Debate Over Safety. *The Wall Street Journal*. Available from: https://www.wsj.com/articles/man-vs-machine-at-40-000-feet-11546776000 [Accessed 20/02/21].

Racounter. (2019) A Day in Data. Available from: https://www.raconteur.net/infographics/a-day-in-data [Accessed 22/01/21].

Rogers, E.M. (2010) *Diffusion of Innovations*, 4ed, Simon and Schuster. Available from: https://tinyurl.com/PredInnov [Accessed 20/01/21]

Samuel, A. (1959) Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, vol. 3 (3), pp. 210–229. Available from: doi.org/10.1147/rd.33.0210.

Santosuosso, A. (2020) *Intelligenza artificiale e diritto. Perché le tecnologie di IA sono una grande opportunità per il diritto*. Firenze, Mondadori Education.

Schwall, M., Daniel T., Victor T., Favarò F. (2020) Waymo Public Road Safety Performance Data. Available from: https://storage.googleapis.com/sdc-prod/v1/safety-report/Waymo-Public-Road-Safety-Performance-Data.pdf [Accessed 23/01/21].

Skitka, L. J. (2011) Automation Bias. *uic.edu*. Available form: https://lskitka.people.uic.edu/styled-7/styled-14/index.html [Accessed 31/01/21].

Sullivan, J. (2017) Ouch, 50% Of New Hires Fail! 6 Ugly Numbers Revealing Recruiting's Dirty Little Secret. *ERE Recruiting Intelligence*. Available from: https://www.ere.net/ouch-50-of-new-hires-fail-6-ugly-numbers-revealing-recruitings-dirty-little-secret/ [Accessed 28/01/21].

Turing (1950) A. M. Computing Machinery and Intelligence. *Mind*, LIX (236), pp. 433–460, Available from: doi.org/10.1093/mind/LIX.236.433.

Wise, J. (2011) What Really Happened Aboard Air France 447. *Popular Mechanics*. Available from: https://www.popularmechanics.com/flight/a3115/what-really-happened-aboard-air-france-447-6611877 [Accessed 21/02/21].