# TROVA L'ERRORE: ALAN TURING, PHILIP K. DICK E L'INTELLIGENZA ARTIFICIALE (II)

*Data: 6 Marzo 2023 - Di Stefano Suozzi*

**Rubrica:** Pensare il pluriverso

Quando oggi dedichiamo maggiore attenzione alla mole di dati che l'intelligenza artificiale (AI) è in grado di gestire e ai processi attraverso i quali opera su di essi, entrambi ben al di là delle capacità umane, emerge con forza quello che potremmo considerare come un problema del controllo. Come osserva Elena Esposito:

disponiamo di informazioni di cui spesso nessuno può ricostruire né comprendere la genesi, ma che ciononostante non sono arbitrarie. Le informazioni generate autonomamente dagli algoritmi non sono affatto casuali e sono del tutto controllate, ma non dai processi della mente umana. Come possiamo controllare questo controllo, che per noi può essere anche incomprensibile? (E. Esposito, *Comunicazione artificiale. Come gli algoritmi producono intelligenza sociale*, Bocconi University Press, Milano 2022, p. 14).

Lo stesso problema, come spesso accade, è già delineato in *Holy Quarrel*, un racconto che Philip K. Dick pubblica nel maggio 1966 in «Worlds of Tomorrow» (dagli appunti risulta scritto il 13 settembre 1965; trad. it. *Teologia per computer*, in Id., *Tutti i racconti. 1964-1981*, Fanucci, Roma 2012, pp. 307-331).

3. La storia, in breve, è questa. Joseph Stafford, un riparatore di computer

(non un ingegnere, né un programmatore), viene prelevato a notte fonda e in tutta fretta da agenti governativi perché il supercomputer Genux-B ha immediato bisogno di riparazione. Genux-B è progettato per accedere a tutte le informazioni disponibili e così governare la società e prevenire attacchi nemici, anche rispondendo preventivamente a ogni possibile minaccia. Ebbene, il funzionamento di Genux-B è stato bloccato, infilando un cacciavite tra i suoi ingranaggi (siamo ancora nell'epoca dei computer elettromeccanici che operano con schede perforate), perché ha ordinato un attacco termonucleare preventivo contro la California del sud (che non è in territorio nemico) e il motivo per cui l'ha fatto è che il signor Herb Sousa sta installando dei distributori di chewing-gum (proprio quelle bocce di vetro piene di palline di chewing-gum di tanti colori diversi) nei locali e negli esercizi commerciali della zona (finora circa 60). Il compito di Stafford è capire perché Genux-B ha dato l'ordine di attacco e se questo ordine sia giustificato o se invece non sia il frutto di un errore commesso dal computer. E qui iniziano i problemi filosofici, sempre i peggiori!

Prima ancora di avere la sicurezza che Genux-B si stia sbagliando, e per avere il tempo di comprenderlo, è necessario far annullare l'ordine di attacco con un sistema migliore di un cacciavite. Il primo tentativo è informare Genux-B che Herb Sousa non esiste. Ma Genux-B non cade nell'inganno perché, secondo Stafford, ha applicato una propria versione della prova ontologica dell'esistenza di Dio all'esistenza di Herb Sousa: i dati e le informazioni che ha raccolto su di lui sono tali e tanti per cui Herb Sousa non può non esistere, e pertanto l'informazione sulla sua non esistenza deve essere falsa (Dick, 322). Il tentativo successivo è quello di mettere in crisi Genux-B informandolo che è lui stesso a non esistere. Ma anche in questo caso il computer se la cava egregiamente: se lui non esistesse non avrebbe potuto ricevere l'informazione della sua non esistenza, pertanto, avendo ricevuto l'informazione, può fondatamente affermare la propria esistenza e la falsità dell'informazione ricevuta (Dick, 323). Infine, la domanda fatale, che nessuno fino a quel momento aveva pensato di porre: «Chi è Herb Sousa?». E a questo punto la

risposta di Genux-B rivela il suo stato effettivo:

> *Herbert Sousa di Sacramento, California, è il demonio. Dato che è l'incarnazione di Satana sulla Terra, la provvidenza esige la sua distruzione. Io sono semplicemente un agente, per così dire una creatura, della divina maestà. Come lo siete tutti voi (Dick, 327).*

Ormai non ci sono dubbi: Genux-B è impazzito e ciò è dimostrato dal fatto che ha evidentemente manifestato i sintomi di un delirio religioso! La domanda, inoltre, induce Genux-B ad annullare l'attacco e l'allarme rientra: viene perciò dato l'ordine di smantellare Genux-B e di revisionare tutti i computer del suo tipo esistenti al mondo.

4. Fin qui nulla di particolare: un tipico racconto di fantascienza degli anni Sessanta, comprensivo della minaccia di un supercomputer che raggiunge una forma di autocoscienza e che per questo, come ogni essere umano, può impazzire (una situazione forse peggiore dei computer affetti da delirio di onnipotenza che odiano tutti gli uomini...). Ciò che lo rende ancora oggi interessante è che Dick riesce a mettere in luce due aspetti complementari dell'AI contemporanea che vale la pena esaminare.

Innanzitutto, Genux-B opera a un livello assolutamente irraggiungibile da un essere umano. Stafford lo descrive chiaramente:

> *Sapete cosa può far concludere a un Genux-B che ci stanno attaccando? Un milione di fattori diversi. Tutti i possibili dati conosciuti vengono soppesati, comparati, analizzati, e si arriva alla Gestalt assoluta. In questo caso, la Gestalt di un imminente attacco nemico. Non basta un elemento solo a superare la soglia. È un processo quantitativo (Dick, 311).*

Se si sostituisce il termine *Gestalt* (Dick amava il tedesco e la cultura tedesca) con *Pattern* si ottiene una perfetta descrizione del modo in cui opera oggi l'AI. Ma, soprattutto, le *sovrumane* capacità di calcolo di Genux-B rendono

impossibile comprendere se stia funzionando correttamente o no:

> *«E voi siete certi», disse Stafford, «che non ci stanno attaccando?». Anche se Genux-B aveva sbagliato nelle due occasioni precedenti, almeno in teoria poteva avere ragione quella volta. «Se stanno per attaccarci» disse l'uomo dell'FBI più vicino «non riusciamo a scoprirne la minima indicazione. Non con l'analisi umana dei dati, in ogni caso» (Dick, 313; sottolineatura nostra).*

Il problema è tale che Stafford comprende immediatamente che da lui non ci si attende una soluzione logica, ma si confida piuttosto nella sua sensibilità e nel fatto che l'aver lavorato a lungo con Genux-B gli abbia fornito una conoscenza speciale della macchina e dei suoi *umori*:

> *Io non sono qui per riparare o distruggere; sono qui per decidere. Perché da quindici anni sono in contatto fisico con Genux-B, e questo dovrebbe conferirmi la mistica capacità intuitiva di capire se la macchina funziona bene o no. Io dovrei sentire la differenza (Dick, 319; sottolineature nostre).*

Ma affidarsi all'intuizione, alla capacità di *sentire* se una macchina funziona in modo corretto, o più semplicemente alla fiducia – «il vero interrogativo è se ci fidiamo o no di Genux-B più che di noi stessi» (Dick, 318) – non può essere una soluzione plausibile. Senonché, ed è questo il secondo aspetto a cui facevamo riferimento, la vera questione non è semplicemente quella di comprendere se la macchina funziona correttamente o no, ma si tratta di comprendere se la risposta fornita è corretta quando il funzionamento della macchina è costitutivamente al di là della capacità di comprensione dell'uomo, quando la macchina è stata volutamente progettata per fare ciò che l'uomo non è in grado di fare:

Genux-B è stato creato per gestire simultaneamente più dati di quanto sia

possibile a qualunque uomo o gruppo di uomini. Assorbe più dati di noi, e li assorbe più in fretta. Le sue risposte arrivano in microsecondi. Se Genux-B, dopo aver analizzato tutti i dati a disposizione, ritiene che ci siano indizi di guerra, e noi non siamo d'accordo, può semplicemente darsi che il computer stia funzionando «come deve funzionare». E più noi siamo in disaccordo, più la cosa viene dimostrata. Se noi potessimo percepire come lui la necessità di un immediato attacco offensivo sulla base dei dati disponibili, non avremmo bisogno di Genux-B. È esattamente in un caso del genere, quando il computer ha dichiarato un Allarme Rosso mentre noi non vediamo alcuna minaccia, che entra in gioco la vera utilità di una macchina della sua classe (Dick, 318; sottolineatura nostra).

5. Ecco il punto della questione: il problema non è solamente quello di capire se Genux-B è guasto; il problema è che Genux-B lavora a un livello tale che per un essere umano è impossibile capire, da un lato, se il computer è guasto in base alle risposte che fornisce e, soprattutto, dall'altro lato, se le risposte fornite sono corrette anche quando si ha la certezza (o, vista la situazione, la presunzione della certezza) che il computer stia funzionando correttamente, ovvero di cogliere il grado di correttezza delle risposte *indipendentemente* dal fatto che il computer sia guasto o no. La vera utilità di Genux-B, e fuor di metafora dell'AI, si dispiega, infatti, proprio nella gestione di quanto è al di là delle capacità umane, ma così facendo non si perde solamente la possibilità di comprendere quali siano i processi attraverso i quali l'AI perviene ai suoi risultati, ma anche ogni possibilità di comprendere se i risultati così ottenuti siano corretti o no. Non si rischia di perdere il controllo delle macchine o che le macchine decidano di prendere il controllo del mondo, eventualmente decretando l'inutilità o la dannosità della presenza dell'uomo. Ma si sta costitutivamente rinunciando alla possibilità di valutare la correttezza delle risposte fornite dall'AI pur sapendo che, anche se non sono il frutto di processi irrazionali, casuali o arbitrari, come sottolinea Esposito, ciò non significa necessariamente

> *che siano corrette, neutrali o che debbano essere accettate senza riserve e senza critica. Come mostrano le dinamiche di feedback, la presenza di controllo non esclude rischi, manipolazioni e risultati negativi. D'altro canto anche il controllo umano, come è noto, non è certo garanzia di successo, e nemmeno di razionalità (Esposito, 13, nota 8).*

Allo stadio attuale dello sviluppo dell'AI siamo ancora in grado di valutare, o quantomeno *percepire*, come richiesto al povero Stafford, alcuni dei limiti che ne condizionano le prestazioni e i risultati. Questi sono particolarmente evidenti, per esempio, nei fenomeni delle cosiddette *filter bubbles* e dei relativi correttivi anti-isolamento nei processi di profilazione (cfr. Esposito, 194ss) e, soprattutto, nei fenomeni di *under-fitting* e *over-fitting* connessi agli algoritmi predittivi (cfr. Esposito, 194ss).

Il caso degli algoritmi predittivi è particolarmente interessante perché è evidente che «ciò che possiamo conoscere del futuro non è il futuro, ma solo l'immagine presente del futuro: le nostre aspettative e le informazioni su cui si basano» (Esposito, 181). Certo, questo è vero anche per ogni buon racconto di fantascienza che sia in grado di estrapolare dalle condizioni del presente una rappresentazione plausibile, e non sempre auspicabile, del futuro. Ma se si tratta di macchine da cui dipenderanno, per esempio, politiche e strategie di sanità pubblica, polizia, difesa, economia, organizzazione del lavoro, accesso ai beni e all'istruzione, e tutti quei settori in cui si deciderà che, parafrasando Dick, «ci fidiamo più dell'AI che di noi stessi» (e a monito futuro si pensi alla *naturalizzazione* dell'economia per cui si giustificano scelte evidentemente politiche come se fossero il risultato di calcoli matematici razionali, scientifici, e pertanto insindacabili), dobbiamo sempre ricordare che

> *Gli algoritmi non sono degli osservatori neutrali che conoscono oggettivamente il mondo così com'è. Gli algoritmi non conoscono affatto il mondo: non «conoscono» niente. Il punto è piuttosto che gli algoritmi sono essi stessi reali e fanno parte del mondo in cui operano, dall'interno,*

> *non dall'esterno facendo riferimento a un modello. Questo modifica il significato del termine «previsione». Quando gli algoritmi fanno delle previsioni, non vedono in anticipo un dato esterno indipendente, il futuro che ancora non c'è. Sarebbe impossibile. Gli algoritmi fabbricano» con le loro operazioni il futuro che anticipano. Gli algoritmi, cioè, prevedono il futuro plasmato dalle loro previsioni (Esposito, 188-189).*

Riassumendo il futuro che ci attende: per gestire in modo più efficace problemi e situazioni in cui l'uomo non ha sempre avuto successo o mostrato grande razionalità, ci affidiamo a macchine che, per ottenere risultati migliori di quelli umani, gestiscono una quantità di dati assolutamente ingestibile dall'uomo e operano attraverso processi che, per quanto razionali, sono ormai così oscuri che si è perduta la possibilità di valutare la correttezza delle soluzioni proposte. Ma i dati su cui opera l'AI individuando *pattern* significativi da cui desume le eventuali risposte ai problemi che le vengono sottoposti, provengono proprio da quello straordinario catalogo delle debolezze e delle miserie umane (e non solo) che è il web e tutte le forme di interazione sociale che in esso si dispiegano: non si può certo affermare che la fonte dei *big data* sia un esempio di tolleranza, razionalità o scelte informate e responsabili.

Forse è vero, come sosteneva Dick, che nell'arte, anche nella più modesta come la letteratura di genere, ed in particolare nella fantascienza, possono celarsi verità che altrimenti non sarebbe possibile cogliere.

**PS:** Ma ancora manca il finale del racconto: Genux-B aveva dunque ragione ad ordinare l'attacco contro la California o no? Come spesso accade nei racconti di Dick, Genux-B non aveva ragione, ma non aveva nemmeno torto. Herb Sousa non era l'emissario di Satana, ma una volta tornato a casa, Stafford si accorge che ha in tasca ancora tre chewing-gum. Strano: gliene avevano dati tre, ma uno lo aveva mangiato. La mattina dopo sono cinque. Stafford li brucia. Ma la sera ne ritrova cinque. E così via. Alla fine del mese, pur avendo cercato di distruggerli in tutti i modi possibili e immaginabili, i chewing-gum

continuano a riprodursi a ritmo esponenziale e nel suo appartamento ve ne sono più di 2 milioni. Finalmente si decide a telefonare all'FBI: «Ma a quel punto l'FBI non era più in grado di rispondere» (Dick, 332). Pur nella sua follia, Genux-B aveva dunque *visto* il pericolo, ma non so se possiamo considerare questo finale un motivo di ottimismo e fiducia per quanto riguarda il futuro che ci attende in un mondo sempre più governato dall'Intelligenza (?) Artificiale...