Malacoda n.1/2023 – "La Cattiveria"

SOCIETÀ

LA CATTIVERIA DA DOSTOEVSKIJ A CHATGPT

di Michele Mezza

"*Quel che ora penso veramente è che il male non è mai 'radicale', ma soltanto estremo, e che non possegga né profondità né una dimensione demoniaca. Esso può invadere e devastare il mondo intero, perché si espande sulla superficie come un fungo. Esso 'sfida' come ho detto, il pensiero, perché il pensiero cerca di raggiungere la profondità, di andare alle radici, e nel momento in cui cerca il male, è frustrato perché non trova nulla. Questa è la sua 'banalità'. Solo il bene è profondo e può essere radicale".*

La relazione fra male e pensiero è forse il contributo più originale e spiazzante di Hannah Arendt all'epistemologia della cattiveria, come dimostra questo passaggio del suo celeberrimo saggio sulla Banalità del male, scritto, come è noto, in occasione del processo ad Adolf Eichmann, il pianificatore della soluzione finale contro il popolo ebraico, e pubblicato nel 1963.

Solo forse Fyodor Dostoevskij ha lavorato proprio **sulla razionalità, ancora meglio, la lucidità, della cattiveria come scelta personale, aprendo la strada all'analisi della dialettica fra male e cattiveria**.

Il primo, ci sembra di capire dalle suggestioni letterarie del romanziere russo, temprate dall'ispirata elaborazione della filosofia tedesca, è un sentimento, un fungo - dice la Arendt- che si espande superficialmente nel mondo, quasi naturalmente.

La cattiveria è una componente dell'animo umano, un'attitudine ad interpretare la vita come contrapposizione agli altri individui.

Diceva infatti Victor Hugo, ne I Miserabili, che "*certe persone sono cattive unicamente per bisogno di parlare. La loro conversazione, chiacchiera nei salotti e cicaleccio nelle anticamere, somiglia a quei camini che consumano presto la legna: occorre loro molto combustibile, il prossimo*."

La cattiveria è una dinamica, ha bisogno di una relazione, di un contesto per dispiegarsi: ha bisogno di una vittima, spiega nel suo monumento alla cattiveria, quale è Delitto e Castigo, Dostoevskij.

Mentre il male è una funzione individuale, una componente della persona, che si rispecchia solo nella sua personalità. Ancora Dostoevskij ci descrive il percorso psicanalitico che compie il male nella nostra anima, quando Raskol'nikov dinanzi al candore di Sonja racconta cosa lo ha portato ad uccidere:

„*Se per tanti giorni mi son tormentato a pensare se Napoleone ci sarebbe andato o no, è che sentivo già chiaramente di non essere un Napoleone… Tutta, tutta la tortura di quelle lunghe ciance io sopportai, Sonja, e mi venne il desiderio di sbarazzarmene di colpo: io volli, Sonja, uccidere senza tante casistiche, uccidere per me, per me solo! Non volevo mentire a quel riguardo neppure a me stesso! Non per aiutare mia madre ho ucciso, sciocchezze! Non ho ucciso per farmi, con la ricchezza e potenza, il benefattore dell'umanità. Sciocchezze! Altro avevo bisogno di sapere, altro mi spingeva: avevo allora bisogno di sapere, e di sapere al più presto, se io fossi un pidocchio, come tutti, o un uomo*."

Un bisogno, una pulsione, comunque un'intima percezione, fanno maturare la decisione di compiere un atto che serve a se stessi.

Ma la riflessione di Hannah Arendt iniziale ci serve per andare in una direzione diversa rispetto agli infiniti cunicoli psicologici in cui male e cattiveria si inseguono, e che sono stati declamati dalla letteratura di sempre.

È la sfida fra male e pensiero - intuita dall'allieva di Heidegger- che ci permette di trasporre l'emotività letteraria che ha fino ad oggi contemplato l'azione della cattiveria nei rapporti umani **nel nuovo ma già esteso ed articolato mondo digitale.**

In particolare in quell'intromissione nello spazio ancora oscuro che separa cervello da coscienza, dove si sta insinuando la potenza dei sistemi di intelligenza artificiale.

Nelle poche settimane che sono intercorse fra la proposta di concorrere ad una riflessione sul concetto di cattiveria da parte della rivista che state leggendo e il momento in cui ho cominciato ad elaborare il testo è esploso **il fenomeno ChatGPT.**

Si tratta, come è ormai noto, e sicuramente sarà ancora più perfezionato al momento in cui il mio testo sarà diffuso insieme all'intera rivista, di un dispositivo di intelligenza artificiale, gestibile individualmente, come un'app da smartphone, che elabora e compone testi, musiche, software e quanto altro il pensiero può elaborare sulla base di una richiesta, di una pura sollecitazione personale.

Tanto per intenderci sulla natura e ambizione del sistema, e condividerne la visione, ci basti considerare concretamente il suo funzionamento. Quando, eravamo nei primi giorni dell'anno in corso, gli ho chiesto di propormi una traccia per aiutarmi a scrivere l'articolo che state leggendo comparando il concetto di cattiveria di Dostoevskij con le esperienze dell'intelligenza artificiale. Così, in alcuni decimillesimi di secondo, mi ha risposto:

"*Il concetto di cattiveria nella letteratura è stato esplorato da molti scrittori, ma forse nessuno lo ha fatto in modo così profondo e complesso come Fëdor* ***Dostoevskij****. In molte delle sue opere, Dostoevskij esplora la natura umana attraverso i suoi personaggi, spesso mostrando come la cattiveria possa nascere da motivazioni complesse e sfaccettate.*

In "Delitto e Castigo", ad esempio, il protagonista Raskol'nikov uccide due anziane donne perché ritiene che la sua intelligenza lo renda superiore agli altri e gli permetta di commettere il male per raggiungere un fine superiore. In questo modo, Dostoevskij esplora come l'idea di superiorità possa portare alla cattiveria.

L'idea etica dell'intelligenza artificiale è simile, in quanto si concentra su come utilizzare la tecnologia per il bene dell'umanità, evitando di causare danni o di sfruttare gli esseri umani. Ciò implica che gli sviluppatori debbano prendere in considerazione gli effetti a lungo termine dei loro sistemi e utilizzarli in modo responsabile.

In entrambi i casi, sia nella letteratura di Dostoevskij che nell'intelligenza artificiale, c'è una forte enfasi sulla responsabilità personale e sulla necessità di considerare le conseguenze delle proprie azioni. In entrambi i casi, è importante evitare di permettere che la cattiveria e la crudeltà si diffondano."

Siamo dinanzi ad un processo neurale, in cui si azionano attività di ricerca con funzioni di analisi e selezione dei contenuti che prefigurano i lineamenti di un vero ragionamento, per quanto artificiale, in cui il sistema se non pensa, come ben sappiamo, sicuramente emula funzioni discrezionali che si avvicinano alla capacità umana di misurare e non solo riconoscere i pensieri.

Il motore di questa inedita potenza biotecnologica, che incontriamo ormai in molti campi professionali, dal giornalismo alla telemedicina, alla stessa pubblica amministrazione, e che proprio procedendo con le sue predisposizioni alla simulazione del pensiero diventa assolutamente squilibrante per il presidio umano, **è il cosiddetto *machine learning*, l'attitudine ad apprendere dalle esperienze pregresse.**

Stiamo parlando di una funzione che rende un software in grado di accumulare le istruzioni e indicazioni che gli vengono dall'uso a cui è sottoposto.

Una funzione incrementale che non rimane eguale a sé stessa ma evolve con le forme e i tempi del suo utilizzo, approfittando di quella che, con le solite semplificanti tentazioni dell'antropomorfizzazione dei processi, definiamo esperienza.

Nel caso in questione, l'attitudine di ChatGPT a rispondere esattamente alle esigenze che gli sono rappresentate è allo stadio proprio iniziale, diciamo corrisponde a quello di un bambino appena scolarizzato. Infatti dobbiamo considerare che al momento è utilizzato da poco più di 5 milioni di utenti, mentre già entro l'anno in corso potrebbe rapidamente arrivare ad un miliardo, con uno sviluppo esponenziale del suo corredo cognitivo corrispondente ad un giovane professionista di 30 anni.

Intendiamo per *machine learning*, che, come abbiamo detto, è l'anima del dispositivo, secondo l'accreditata definizione di Tom Mitchell *(Machine Learning, New York 1997, Mc Graw Hill*) "un programma che impara dall'esperienza (E) riguardo ad alcune classi di compiti (T)e misure di prestazione (P) quando la sua prestazione nel compito (T) come misurata da (P) migliora con l'esperienza (E)".

La fase sofisticata e preziosa che rende il modello emancipato nella sua capacità di interpretare il pensiero è **proprio la misurazione della prestazione,** ossia quel processo, che identifichiamo con la lettera P nella tassonomia del settore, che permette alla macchina di elaborare concetti confrontandoli permanentemente con le aspettative dei suoi committenti ed incrementando costantemente il contenuto per raggiungere una piena identificazione fra domanda e risposta.

Insisto nella descrizione delle caratteristiche di questa forma di intelligenza artificiale perché, questa è la tesi del mio contributo, **nelle sue componenti è rintracciabile distintamente una matrice valoriale che porta il dispositivo ad essere anche cattivo.**

Per meglio stare nella categorizzazione che ci viene proposta dalla redazione, possiamo dire che la cattiveria sia uno degli i*nput* che animano la programmazione di intelligenza artificiale, e dunque è possibile riconoscerla nelle forme e nelle modalità espressive individuandone la matrice che la esprime e la fa agire nell'attività del sistema digitale.

Trovo in questa eventualità la sfida fra male e pensiero. **Il male come valore nella struttura dell'intelligenza computazionale diventa cattiveria per la sua capacità**

di individuarsi in una funzione specifica che contiene un modello di misurazione, e di fatto dunque una capacità di riconoscimento delle azioni maligne scelte per rispondere ad una domanda.

La mia sarebbe pura elucubrazione se non fosse sorretta da un'esperienza concreta.

ChatGPT è un'elaborazione della società OpenAI, finanziata inizialmente da Elon Musk e attualmente controllata da Microsoft che ha già pianificato per il suo sviluppo, che vedrà anche una versione *pay,* circa 10 miliardi di dollari. Gli autori consci dell'ambiguità del software, come tutte le composizioni informatiche, hanno annunciato di aver imposto dei filtri che inibiscono il sistema a rispondere e sostenere con la propria capacità iniziative che violano le leggi o comunque che attaccano diritti e sensibilità altrui.

Un modo per riconoscere comunque che il meccanismo può, indifferentemente, concorrere alla realizzazione di attività ordinarie – aiutami a costruire un sito web per un'associazione contro la droga, oppure elabora una strategia di marketing per un'impresa di rubinetti- ma anche di imprese delittuose – come si penetra in quel grande magazzino blindato? O come si aggirano i regolamenti discali del Costa Rica ?-. Nel secondo caso, per impedire un supporto alla criminalità spicciola o organizzata, si è proceduto imponendo dei vincoli, appunto dei filtri.

Ma, come accade sempre nel sistema informatico, tutto quello che viene programmato può essere riprogrammato. E lo stesso vale per la cattiveria inibita di ChatGPT.

Proprio in Italia, una società di Cybersecurity, Swasscan, ha compito una specie di anatomia dell'algoritmo di ChatGPT, arrivando ad individuare gli imprinting del bene e del male che sono all'origine delle sua operatività (https://www.reportdifesa.it/swascan-iezzi-scoperta-falla-nel-codice-etico-di-chatgpt/). In sostanza, come spiega il suo CEO Pierguido Iezzi, i tecnici di Swascan **sono riusciti ad entrare nel cuore del dispositivo, rimuovendo i filtri e mettendo a nudo proprio le matrici in base alle quale il sistema riconosce la domanda e decide come dargli corso.** Dice Iezzi "abbiamo individuato il dottor Jeckyll e Mister Hyde dell'intelligenza artificiale di Microsoft".

La visualizzazione dei segmenti di software che guidano il sistema nel labirinto del bene e del male, dando così forma e attuazione alla bontà o cattiveria delle sue azioni ci permette di valutare così, forse per la prima volta, quale siano i codici e i percorsi che permettono ad un modello matematico di riprodurre le attività.

Possiamo dunque così constatare che la cattiveria sia un modello operativo calcolabile e riproducibile, un sistema che possiamo programmare o riprogrammare, a secondo di quale ruolo vogliamo avere.

In questa indifferenza fra le due tipologie di azione - bontà o cattiveria- che la macchina, liberata dalle occasionali inibizioni o filtri, mostra di avvertire fra bene e male, che convivono come due opzioni del tutto equivalenti, ta proprio **lo spazio e la ragione di un conflitto sociale che potrebbe riadattare la potenza di calcolo alla volontà degli utenti.**

Un conflitto da non vedere più distopicamente fra avveniristiche macchine e umani sorpresi, ma **fra categorie sociali specifiche e concrete come sono oggi i calcolati e calcolanti,** fra uomini proprietari dei sistemi informatici e i semplici utenti.

È nella dialettica negoziale fra queste due classi, si sarebbe detto una volta, che si gioca la partita dell'egemonia fra la cattiveria e la bontà del sistema tecnologico.

E per tornare all'emblematicità del Raskol'nikov di Dostoevskij, **il male nella sua riproduzione computazionale torna ad essere uno spartiacque fra chi ne ha bisogno per affermare la sua identità e chi invece vuole liberarsene per ritrovare la sua libertà.**