# Spiegare il male a una macchina

Volevo scrivere un post sulla neo-censura che sorge dai goffi tentativi di rilevazione automatica dell'odio nel linguaggio online e nei dataset della IA. Ma mi sono reso conto che c'è un passaggio fondamentale da fare a monte.

Il discorso che ingiuria, calunnia, umilia, è detestabile in quanto *fa male*. Allora risaliamo alla domanda alla quale ogni "etica" computazionale deve rispondere: *si può spiegare a una macchina che cosa vuol dire "fare del male"*?

Credo che questa domanda contenga il nòcciolo problematico del rapporto tra umani e macchine. Altro che intelligenza e test di Turing.

Subito viene alla mente la celebre prima "legge della robotica" di Isaac Asimov: «Un robot non farà del male a un essere umano né permetterà che, per la propria inerzia, un essere umano si faccia del male». Il fascino senza tempo di questo comandamento che l'uomo-dio impone alla sua creatura

dimostra la maestria dell'Autore. Tuttavia, appena usciamo dall'incantamento della fantascienza e pensiamo sul serio a come metterlo in pratica, la sbornia passa. È dura essere dèi.

Qualunque sia la sua natura, un agente non può obbedire a un comando se non sa o può interpretarlo. Nella fattispecie, una macchina non potrà rispettare quella legge se non capisce che vuol dire "fare del male". Come spiegarglielo?

Fare del male. Ma a chi, come? "Fare del male" è una possibilità inesauribile, sfaccettata e personale quanto la vita. Ha senso dire "fare del male" così in astratto? Il male è sempre concreto e particolare. La solenne universalità della legge di Asimov non ha riferimento. Ciò che sembrava bello e solido a parole si sbriciola fra le dita.

Eppure noi *sappiamo* che significa "fare del male". Il male è il nostro pane quotidiano sin dal primo giorno di vita: la fame, la sete, il distacco. Il nostro corpo soffre se si sbilancia anche solo di poco rispetto alle sue, diciamo così, condizioni normali di esercizio. E i possibili sballi di un organismo tanto complesso non si contano.

Come vive il male in prima persona, così il corpo sa riconoscerne i segni quando affiorano sui corpi degli altri. Basta una smorfia, una lacrima, un lamento, e il loro male-essere si fa strada dentro di noi. Qualcosa dentro si agita, si allarma, si muove in soccorso, o scappa, o talvolta incattivisce.

Questa letterale *com-prensione* del male altrui è istintiva e automatica. Avviene con intensità che vanno da zero a insopportabile, e variano con gli individui e le circostanze. È una naturale capacità di rispecchiamento del cervello, non solo umano: in noi si attivano schemi motori e stimoli viscerali corrispondenti a quelli attivati nell'altro, mentre lo osserviamo, lo ascoltiamo, o lo immaginiamo soltanto a leggerne le storie ben narrate.

Per spiegare *a parole* il male che soffriamo c'è da attraversare due oceani in verticale: uno per portare alla coscienza sentimenti reconditi, distinguendone provenienze e sapori come sommelier; un altro per scegliere e combinare le parole adatte a raccontarli. Imprese estreme se non impossibili per la maggioranza di noi.

Eppure al nostro specchio interiore questi sforzi sovrumani non servono. Un cervello normale fa tutto quel lavoro di simulazione senza coinvolgere la coscienza e tanto meno le parole. È così, nell'esperienza da corpo a corpo sedimentata in vissuto, che sappiamo se abbiamo fatto o potremmo "fare del male" a qualcuno.

Ora, facciamo entrare le macchine in questo gioco che va avanti da chissà quante centinaia di migliaia di anni. Niente corpo vivente, niente specchi modellati dall'evoluzione per sintonizzarsi con gli umani. Come imporre a queste macchine la legge di Asimov? Non ci resta che tentare di descrivere esplicitamente, a parole, cosa significa "farci del male".

Non basterà qualche esempio generico. Bisognerà informarle in dettaglio su *ciascuno di noi*. E in queste informative soggettive bisognerà annoverare tutte le nostre fragilità fisiche, tutte le nostre paure più profonde, tutti i nostri disturbi e imbarazzi, tutti quei minimi oltraggi che feriscono, tutti i traumi e le frustrazioni, le contrarietà, le delusioni e i sacrilegi. Una confessione *totale*. Il sogno dell'inquisitore.

Questi inventari del male, frutto di una consapevolezza assoluta e di una verbalizzazione che fa arrossire Dostoevskij, ovviamente non sono che un'altra invenzione letteraria, come le leggi della robotica. Ricordano le "Vendicazioni" di Borges, quelle disperse nella sconfinata Biblioteca di Babele. Sono resoconti ideali, ma nascerebbero già vecchi, presto superati dalle nuove vicende che ci attendono e ci cambieranno ancora.

Il cuore del problema comunque è un altro. Se esistessero questi ritratti testuali, solo *altri umani*, non tutti, e non certo delle macchine, potrebbero *com-prenderne* il senso per immedesimazione e approssimazione. Perché senza il terreno comune di un corpo simile, e meglio ancora se c'è memoria condivisa degli eventi che lo hanno plasmato, emozioni e sentimenti sono solo parole fra altre parole. Correlazioni statistiche, operazioni algebriche, e nulla più.

Attraverso questo modello esteriore e matematico del linguaggio c'è ben poca speranza che una macchina informatica possa *com-prendere* i sentimenti, cioè accoglierli e integrarli dentro i suoi modelli di azione, così da poter obbedire al primo comandamento di Asimov.

Comandamento che di fatto, prima che per i robot, vale per gli umani. È notevole che gli umani non lo rispettino proprio quando manca loro la facoltà dello specchio emotivo, come manca alle macchine. Ma nemmeno agli umani sani si può insegnare il rispetto degli altri con la mera imposizione di *parole politicamente corrette*. Figuriamoci a una macchina.

Rimane la solita possibile soluzione, quella che ci attende sempre nel rapporto con le macchine: adeguarsi a loro per futili motivi e per volontà di incoscienti, accettare versioni distorte della realtà, e sopportare tutto il male non calcolabile che ce ne verrà.