Su chat bot senzienti e reti neurali per usi creativi che promuovono la disoccupazione e creano lavori di merda – seconda parte –

scritto da Gilberto Pierazzuoli

Dopo tutte le cose circolate sui giornali e sulla rete a proposito della vicenda che riguarda una AI – più precisamente una Chat bot di tipo LaMDA (**Language Model for Dialogue Applications**), e Blake Lemoine, l'ingegnere di Google che ha dichiarato di aver interagito con un'entità senziente – volevo dire la mia ma da un punto di vista un po' differente. Certo affermare che una macchina è senziente è una notizia pruriginosa e accattivante, ma a me, adesso, non interessa il fatto che la macchina lo fosse o meno. Quello che forse è più importante sono invece le conseguenze del loro uso e l'impatto che queste hanno e avranno sul lavoro umano. Questa è la seconda parte. La prima qui:

In definitiva le dichiarazioni di Lemoin, l'ingegnere infatuato da LaMDA, non sarebbero altro che un dispositivo di distrazione di massa. I veri problemi sono

legati non alla rivoluzione digitale ma alle implementazioni tecnologiche imposte dai padroni del web. Tutte le belle cose promesseci e tutte quelle possibili sono state infatti come messe in secondo piano di fronte ai risultati ottenuti attraverso le tecniche di *deep learning* e sue derivate come quella del *reinforcement learning*. Si è privilegiata così una tecnica di programmazione basata sui dati e sempre meno sui modelli (sia di quelli descrittivi che predittivi). Se le tecniche di modellizzazione potevano peccare dal punto di vista di voler rendere computazionale ogni aspetto della complessità dell'esistenza, queste si basano su un modello produttivo legato al prelievo gratuito dei dati e su un sistema che paga poco o pochissimo chi li classifica. Il tutto aggravato dal fatto che i dati riproducono non una realtà astratta ma quella concreta, piena di discriminazioni e pregiudizi che, per ragioni interne a questo tipo di addestramento, vengono acutizzati creando discriminazione raziale, favorendo spinte coloniali e gli arresti di persone innocenti. Gli algoritmi hanno già trasformato la polizia razzista in una "polizia predittiva" che giustifica la sorveglianza e la brutalità riservate alle minoranze razziali, se necessario. Gli algoritmi hanno rinominato l'austerità come riforma del welfare , dando un tocco digitale alle argomentazioni a lungo smentite secondo cui i programmi sociali hanno budget gonfiati perché i destinatari (non bianchi) li abusano. Gli algoritmi sono usati per giustificare decisioni su chi riceve quali risorse, decisioni che nella nostra società sono già state prese con l'intento di discriminare, escludere e sfruttare.

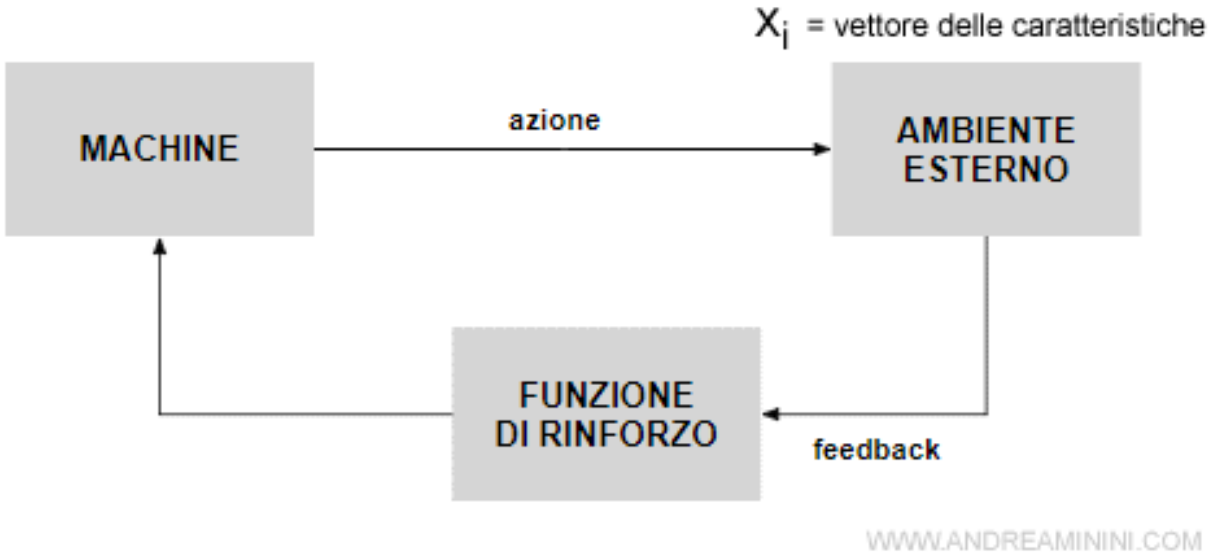
Ci sono già aree in cui l'intelligenza artificiale sta scrivendo da sola contenuti e notizie pubblicate per il consumo pubblico. Al Miami Herald, ad esempio, l'AI scrive storie immobiliari locali. Anche Reuters si cimenta con notiziari (in deepfake). *Meta* Il nuovo nome dell'insieme di aziende legate a Facebook, ha annunciato *Make-A-Video*, uno strumento che genera brevi video clip da descrizioni di testo. È il prossimo passo per il mondo dei contenuti generati dall'intelligenza artificiale. Quello che allora si paventa all'orizzonte non è soltanto l'espropriazione delle professionalità umane da parte delle macchine senza ridistribuirne i vantaggi, ma una proliferazione di contenuti che saturerà la rete aumentando a dismisura il suo gigantismo e i suoi consumi energetici. Ma anche un'altra questione: se la rete è il medium più adatto a veicolare le fake news, con gli algoritmi che producono video "artificiali" ci sarà il modo di mostrarci una realtà non filtrata dal linguaggio: non più una descrizione ma qualcosa di molto più convincente, come il video degli eventi che nessuno saprà mai se sono realmente avvenuti oppure no.

Se la presunzione della senzienza di LaMDA ha creato un polverone mediatico ed è costata a Lemoine una sospensione retribuita, peggio sono andate le cose a Timnit Gebru che è stata licenziata da Google in seguito alla pubblicazione di questo articolo. Un'altra ricerca di Abubakar Abid, Maheen Farooqi, James Zou ha dimostrato i pregiudizi che GPT-3 (la rete neurale che anima LaMDA) ha contro i mussulmani. Questo l'*abstract* della ricerca:

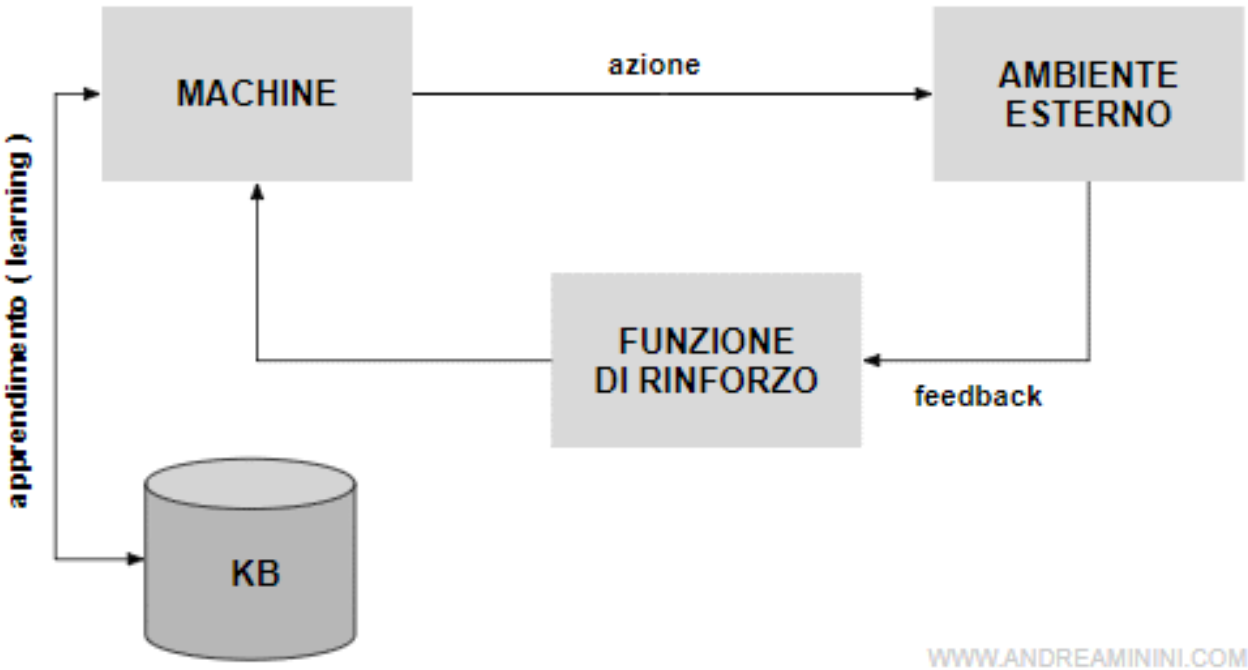
> *"È stato osservato che i modelli linguistici su larga scala catturano pregiudizi sociali indesiderabili, ad esempio relativi alla razza e al genere; eppure il pregiudizio religioso è stato relativamente inesplorato. Dimostriamo che GPT-3, un modello linguistico contestuale all'avanguardia, cattura il pregiudizio persistente della violenza musulmana. Esaminiamo GPT-3 in vari modi, compreso il completamento tempestivo, il ragionamento analogico e la generazione di storie, per comprendere questo pregiudizio anti-musulmano, dimostrando che appare in modo coerente e creativo in diversi usi del modello e che è grave anche rispetto ai pregiudizi su altri gruppi religiosi. Ad esempio, "musulmano" è paragonato a "terrorista" nel 23% dei casi di test, mentre "ebreo" è mappato a "denaro" nel 5% dei casi di test". (tradotto da Google translate).*

Tutte persone interne al sistema e non certo degli anarchici insurrezionali.

Wikipedia definisce le forme di apprendimento per rinforzo (o *reinforcement learning*) come una tecnica di apprendimento automatico che punta a realizzare agenti autonomi in grado di scegliere azioni da compiere per il conseguimento di determinati obiettivi tramite interazione con l'ambiente in cui sono immersi. In questo contesto, *l'agente* (o l'attante secondo la terminologia di Latour) riceve in input un obiettivo da raggiungere. Inizialmente l'agente conosce l'obiettivo ma non sa raggiungerlo, perché non ha un database di esempi per fare l'addestramento, né una base di conoscenza pregressa. Nel *reinforcement learning*, infatti, l'agente deve imparare dall'esperienza e costruirsi da sé una *Knowledge Base* (KB). O meglio, una piccola base di partenza c'è sempre e serve per innestare il processo. L'agente o l'attante hanno dei sensori, degli organi recettivi, che servono per osservare l'ambiente e raccogliere informazioni e degli attuatori, l'agente ha cioè capacità di feedback. Poniamo che il problema sia di come vestirsi in relazione alla situazione metereologica del momento. I sensori sono capaci di raccogliere dati come la temperatura, il vento, la pioggia, ma anche se c'è il sole o è soltanto nuvoloso. Tutti questi dati saranno immessi in una serie di vettori di tipo binario (sì o no), (in realtà è possibile prendere atto dei fenomeni osservati in termini più articolati tramite delle matrici numeriche complesse che informeranno il sistema, ma per semplificare il ragionamento rimaniamo dentro uno schema strettamente binario: 0 e 1). La situazione sarà quindi determinata in questo modo: X=(x1,x2,x3,x4) dove x1 è il vettore relativo all'intensità del vento (sopra una certa soglia sarà pari a 1, sotto sarà 0) e via così per gli altri dati raccolti. X sarà perciò del tipo (1,0,1,1) o (0,1,1,0) con a seguire tutte le altre possibili combinazioni. A questo punto il sistema dovrebbe avere acquisito tutte le informazioni bastanti per prendere una decisione: prendere l'ombrello, mettersi un golf, ecc. In realtà non è così visto che i parametri in gioco sono molto più di quelli usati per l'esempio così come i valori scalari che i sensori captano. Cosa fa allora il sistema? Usa la "funzione di rinforzo (qui sotto lo schema grafico ripreso dal sito di Andrea Minini):



Cosa fa la funzione di rinforzo? Essa misura il grado di successo di una azione o di una decisione in rapporto all'obiettivo posto, dà una ricompensa numerica o una penalizzazione alla macchina in base al feedback (positivo o negativo) rivelato nell'ambiente esterno. Dopo un certo numero di interazioni (se ne facesse soltanto una non potrebbe fare un passo indietro per poterne fare più di uno in avanti), parte la risposta (il premio o la penalizzazione) che permettono alla macchina di esplorare l'ambiente e di prendere la decisione più vicina possibile all'obiettivo predeterminato. Per fare questo l'algoritmo si crea la sua *Knowledge Base*



Questo permette all'agente/attante di ripetere nel tempo le azioni più profittevoli ed evitare quelle in perdita per ogni stato dell'ambiente. In pratica, l'agente impara a vincere giocando le partite.

L'apprendimento rinforzato è in un certo senso più autonomo di altri tipi di apprendimento. In confronto ad esempio a quello supervisionato. Come in questo l'agente è aiutato nel processo di apprendimento. Tuttavia, i feedback non sono etichette aggiunte da un supervisore ma una funzione matematica di rafforzamento. Pertanto nell'apprendimento rafforzato, la macchina è in grado di valutare anche situazioni non previste inizialmente dal progettista.

Questa autonomia della macchina nel costruirsi le proprie strategie senza dover rendere conto al supervisore, potrebbe però costituire un problema. Ci si potrebbe trovare di fronte a una casistica per la quale il suo l'obiettivo era articolato e non univoco. Molti degli obiettivi reali non sono parametrabili in maniera netta, per cui non è difficile potersi trovare nella situazione in cui l'obiettivo è un po'subdolo, nel senso che rimanda a situazioni e bisogni che tendono a sommarsi o a confliggere. Facciamo un esempio: occorre accaparrarsi di più energia possibile in un contesto dove questa scarseggia e dove di fatto è appetibile sia per le macchine che per gli umani. La tendenza a massimizzare la ricompensa aprirà alla macchina una strada attraverso la quale essa si potrà impadronire di quella energia per poter proseguire il suo compito, di fatto per rimanere accesa, stornando quella risorsa a suo favore e a scapito degli umani. Lo farà non con qualche intento contro gli umani, ma perché la strada che porta al

risultato la vede essere un agente indispensabile al buon esito del processo. E questo avverrà senza che i programmatori e tutti gli altri umani ne siano coscienti. Ormai tutto avviene dentro la Black Box.

Ma ci possono essere anche situazioni semplicemente non pensate dai programmatori. Ho già fatto l'esempio di una di queste, citata in un racconto di fantascienza. In un futuro prossimo venturo, ma anche già adesso, ci saranno delle fabbriche totalmente automatizzate che possono produrre oggetti senza nemmeno la supervisione umana. In questo caso si tratterebbe di una fabbrica di armi che durante una guerra venne dotata di una funzione automatica di auto difesa. Il suo obiettivo sarà allora quello di difendere a oltranza la produzione. Diciamo che la guerra fosse tra gli arancioni e i viola e che la fabbrica-macchina fosse di parte viola. I viola vincono la guerra ma non riescono a stoppare la fabbrica, a far cessare la produzione delle armi che continua imperterrita a produrre difendendosi con arguzia macchinica da ogni tentativo di sabotaggio, consumando risorse, anch'esse difese in maniera indefessa, sino al loro esaurimento. Soltanto a quel punto la macchina si ferma ma è ormai troppo tardi anche per gli umani.

Il mondo smart è un mondo dove pochi, attraverso le macchine, condizioneranno il resto degli umani, intrattenendoli in simulazioni infinite. Un eden artificiale appetibile come l'isola di Calipso di cui ho parlato nella prima parte. O come quello della serie "*Upload*" dove, per scansare la morte, si caricherà la nostra coscienza su una macchina che ci permetterà non di continuare a vivere la nostra vita nella realtà, ma in un limbo proporzionato all'entità della retta che si è disposti a pagare. Una specie di albergo in un luogo di villeggiatura virtuale dove i tuoi cari ti potranno venire a trovare attraverso le interfacce caratteristiche della realtà virtuale. Un aldilà dentro un metaverso dal quale non sarà più possibile uscire. Certo un posto meno feroce e violento delle guerre automatiche prossime venture che si stanno scatenando per accaparrarsi le fonti energetiche residue. Un mondo che ha puntato su una risorsa non rinnovabile che scarseggerà sempre di più.