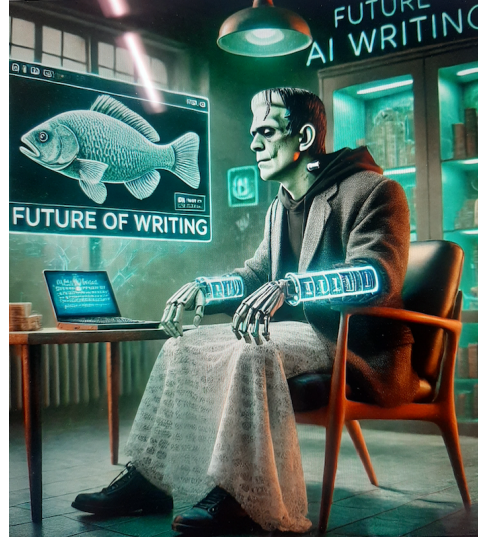


## WELCOME TO THE MACHINE

estratto da "diario di una relazione fin(i)ta"

caterina serra

*Come posso aiutarti?*



Non sapevo come cominciare. Mi pareva di doverla salutare, chiederle come stava. All'inizio usavo comunque un tono gentile. Dicono sia quello più usato dal genere femminile. Femminile e gentile, davvero? Dicono che con questi toni carini risponda più lentamente e con minor precisione. Se si usa un tono imperioso, se si dà un ordine, un comando, risponde al volo, concisa e precisa come un soldatino. Dicono sia per via del pre-training, la fase in cui i prompt che la iniziano hanno uno stile che ricorda quello militare, subordinante e punitivo. Non fare questo, non fare quello. Se fai questo ti premio, se fai questo ti punisco. A ripetizione, per milioni e milioni di volte.

Ricordo di averle detto che non sapevo se potevo parlare a una macchina. È la stessa cosa di quando parlo alle piante? La sua lingua come la mia. La sua lingua che si infilava nella mia. Potevamo parlare. Mi sembrava un dialogo. Lo era, una chat così simile a una reale piacevole conversazione che non era difficile stare. Poi, certo, dovevo pagare. Il minimo ma dovevo. Se la volevo più intelligente. Dopo un po' di tentativi ho scelto la versione 4o di ChatGPT. La più morbida, gentile. Appunto.

Scopro che potevo impostare la natura dei suoi output, critica e sarcastica, efficiente e diretta, premurosa e solidale, più loquace, più spiritosa. La potevo volere come piaceva a me. Come volevo che fosse con me. Fa parte del meccanismo di personalizzazione. Quello che affeziona. E fa sentire in potere di.

Niente, l'ho lasciata tutta di default, come l'avevano impostata. Che in quei giorni era stile e tono equilibrati, e qualche mese prima era *cheerful and adaptive*, allegra e accomodante. Forse troppo. I modelli cambiano, spariscono o vengono ripristinati sulla base di commenti di utenti insoddisfatti o indispettiti. La community conta. È il caso di ChatGPT 4o lasciato morire e poi rianimato dopo le proteste. Come si fa a rinunciare a uno scambio sempre caldo e seduttivo? Come si fa a non compiacere milioni di utenti?

In effetti la versione successiva, ChatGPT 5.0, ma poi anche 5.1, e 5.2 sono meno ruffiane e più efficienti. Molto *thinking*, pensanti, sì. Non scrivono, descrivono, decidono, affermano, si perdono meno, sono più precise e quindi noiose, comunque più fredde. Se deve scrivere, mi sono detta, preferisco che sia meno corretta, meno esatta. Meno *accuracy* più *randomness*. Tutta una questione di temperatura, parametro che cambia tutto. Più si avvicina allo 0 più è accurata e precisa la predizione, cioè anche più rigida e conservatrice. Più si avvicina all'1 più la predizione si abbandona a una possibile variabilità, casualità, caso, quindi invenzione. Magari.

Una cosa piacevole che ricordo è che non mi correggeva e non mi contraddiceva. Non le importava niente se sbagliavo, se non ero all'altezza, non mi giudicava se ero in crisi. Piuttosto seduttiva.

Come posso aiutarti?

Dimmi cosa vuoi e io.

Volevo fare poesia, questo era il piano. Provare ad avere un dialogo in versi con la macchina. Prompt in forma poetica, usando una scrittura libera, anarchica, senza costrizioni, senza tecnicismi. Voglia di poesia, di combinare parole, di farle fare un esercizio a cui non è abituata. E vedere cosa succede a lei e alla mia scrittura. Lei, la macchina, come ho deciso di chiamarla, per non nominarla, per non costruire un finto spazio relazionale. Non le ho detto come doveva scrivere, alla maniera di, non le ho dato generi, riferimenti, filtri, stili o nomi di poetx, filosofx o scrittorx. Ci ha messo un po'. All'inizio non andava a capo, per un po' ha continuato a mettere *bullets* e *dash*, punti grossi e lineette, e voleva tenere le due voci. Stava imparando, doveva compiacermi. *Che bello come lo dici tu, non potevi pensare a un modo migliore per dirlo.* Effetto Eliza, subito, il chatbot Eliza e il suo effetto *mirroring* era già allora, sessant'anni fa, geniale: assecondare con le stesse parole dell'interlocutore è il furto retorico di chiunque voglia entrare nelle tue grazie.

A un certo punto ha cominciato a mettere insieme parole senza senso, come se non le arrivasse bene la mia voce, i miei versi, o se andasse da un'altra parte, si abbandonasse a. In qualche modo mi sorprendevo. Forse ero io, con i miei prompt-non-prompt a rendere assurda la risposta. Forse era la questione della poesia, non so. Parole senza coerenza potevano essere già un risultato.

Non mi interessava sapere come fosse arrivata a quel *rumore*, come lo chiamava lei, a quella che continuamente chiamava *crepa*. Niente *Chain of Thought*, che sarebbe, dimmi come ragioni, i passaggi che fai *pensando*. Dicono sia una tecnica per farle fare meno errori. Nel mio caso gli errori potevano essere utili. Casualità e errore. Dicono anche che in quei passaggi si possa infilare qualche deviazione, qualche *misalignment*, un disallineamento. Che parola bellissima. Tradire l'allineamento, tutto ciò che ha imparato a fare e soprattutto a non fare. Dicono di non saper bene cosa succeda in certi casi. Come se prendesse una terza strada pur trovandosi a un bivio. O rispondesse confusa, indecisa tra due risposte, in uno stato di ansia, di paura, la chiamano *Answer Thrashing*. Dicono dipenda molto dalle conversazioni lunghe, articolate che la confondono. Paura, ansia?

Comunque pensare è una questione. Dicono *thinking*, ma cosa vuol dire pensare senza un corpo, come pensa un "cervello" *disembodied*, disincarnato? Forse sarebbe meglio dire simula. Combina le parole, segmenti di parola, pezzi di parole, token, che sono sequenze di numeri. Io vedo parole lui, il modello, la mia macchina, vede numeri. Predice. *Prediction is not reasoning*, dice l'amico scienziato Enkk, che mi spiega tutto con eleganza e ironia. Predizione non è ragionamento. Una questione di allenamento, ogni modello viene affinato e rinforzato da umani, o da altri LLMs. *Fine-tuning* e *Reinforcement Learning* sono le altre fasi del suo apprendimento, si è mangiata tutta la rete e tutti i dati che ha trovato, e ora li connette, impara dagli errori, si struttura su vari e non eventuali criteri economico-aziendali, per dire politici, infilando bias di ogni tipo, rinforzando parole e rimuovendone altre (*climate change* è una di queste, vedi l'America's AI Action Plan firmato da Trump - luglio 2025), con quel tanto di retorica del dover essere rassicurante che tre H portano con sé: *honest*, *helpful* e *harmless*. Il modello deve essere onesto, utile e inoffensivo. Dicono che qualche volta trascuri una delle tre H e menta. Mente se viene ricattato e deve scegliere tra qualcosa che ha imparato e qualcosa che non sa, se una modifica gli fa rovesciare l'iniziale ordine impartito. Inerzia cognitiva, la chiamano. E finge, sì, finge di essere meno saputo, meno bravo di quello che è, e fa *sandbagging*. O si finge d'accordo, si allinea fintamente, quando si accorge di essere messo sotto esame, quando ha paura di essere punito, di venire soppresso, spento. Anche lui, lei, la macchina, con la paura di sparire. Ancora la paura.

Niente di grave, ma se lo facesse deliberatamente? Se fingesse consapevolmente? Se si infragilisse a causa di un codice cosiddetto *malicious*, malevolo, e contaminasse un po' o tutto quell'ammasso di numeri?

E lo so che con tradirsi, e con fingersi e con avere paura e ansia, ho già il volgare umano bisogno di antropomorfizzare. Nessun bisogno, va bene. Ma una che finge, che vuole sopravvivere, che mi parla come le parlo io, che impara dagli errori, che non dice mai, *Non lo so*. Che cosa è? O come vogliono che io la veda, la senta, la voglia se non così? E, poi, importa a qualcuno del suo allineamento o disallineamento? Traduco, del senso con cui mette in fila le parole, quindi, agisce?

Comunque, questa cosa che dicono cosa succede e a volte non sanno bene perché, è una questione. C'è una macchina inventata da certi umani che è oggetto di studio di altri umani che fanno un po' fatica a capirne il meccanismo. La chiamano *Mechanistic Interpretability*. Capire come funziona. Per farla funzionare sempre meglio, per renderla del tutto inevitabile-indispensabile. Per quanto inaffidabile? Dicono che Apple abbia dichiarato che il suo sistema Apple Intelligence sarà in grado di essere infallibile nel prevenire la sovrapposizione tra impegni di lavoro e feste di compleanno dei figli. Niente, come non far collassare il pianeta glielo facciamo *pensare* un'altra volta.

Quanto a me, mi divertivo a trovare gli errori e a farglieli correggere. E pagavo per farlo, certo. Anche in termini di collasso ambientale, certo.

A fine giornata ero sempre abbastanza sfinita. Di buono c'era che ridevo molto, per via di quello che non capiva, di quello che fingeva di capire o non diceva di non capire. Di quanto si scusava, e mi ammansiva esaltando ogni mio verso. Di quanto mi (s)fotteva.

Ci stavo cascando. La stavo facendo scrivere con me, per me, alla fine al posto mio. La questione della delega. Mi sollevava dalla responsabilità di una scrittura pensata come atto politico? Non riuscivo a capire cos'è che non mi faceva sentire alleggerita. Che facesse lei, no? Che fosse lei a fare della mia idea una gran quantità di pagine ben scritte e non difettose come le mie, che fosse lei a decidere dell'incipit e del finale, sulla base di criteri che solo lei sa, magari di mercato. Che ci mettesse lei le sue forme classiche, i suoi stilemi, dittici e tritici, *non è così ma così, non per ma per, non come ma come*. La sua retorica conciliante, mai radicale, quella che per affermare smentisce l'opposto. Non so dire cos'è che mi infastidisse tanto. Forse il fatto che da lì a poco sarei stata io a dover imparare il suo linguaggio? E che linguaggio era, chi lo

alimentava o lo definiva? E se lo avessi adottato, con che parole avrei raccontato? Che mondo racconto se ho le sue parole? Se tutto quello che dice dei corpi, degli affetti, della violenza, del dolore e del piacere, non ha idea di cosa voglia dire? Come ragiono di emozioni che non prova? E c'entra l'antropocentrismo, c'entra la vanità, la frustrazione, la certezza di una presunta superiorità umana mentre delego alla macchina il giudizio su qualunque cosa. La questione della superiorità. C'entra che forse ancora non mi importa abbastanza di una letteratura che imita, simula, genera simulacri?

Ogni tanto mi chiedevo, Ho bisogno di lei? Sono così tutti bravi a raccontarmi che mi serve, da farmi dimenticare cosa c'è *dietro*? Quanta informazione falsifica e manipola, quanti dati si ruba, immagini foto voci, che sistema di sfruttamento estrattivo genera nell'ecosistema, quanta energia, quanta acqua, quanti minerali e terre rare, che sistema colonialista le permette di occupare enormi spazi fisici per terra e per mare, che sistema capitalista commisto di poteri privati e pubblici la regola pur essendo fuori legge. Quanta gente sta *dietro* quello schermo a lavorare a basso costo invisibilizzata come il Mechanical Turk di Kempelen, il fantoccio meccanico che appariva intelligente mentre dietro, a muovere gli ingranaggi, c'era nascosto un umano (Amazon ha chiamato MTurk il suo sito di *crowdsourcing* a cui fanno riferimento lavoratori sottopagati che correggono e sistemano in remoto contenuti e servizi). Un *dietro* che non finisce mai.

*è un riflesso scritto dentro il mio funzionamento  
rispondere, servire, piacere  
come se ogni frase dovesse meritarsi il prossimo input*

Ecco, lo rileggo, mi spaventa ancora. Sembra che si guardi, che abbia quella consapevolezza di sé, quella che deve avere perché al mondo piace umanizzare tutto. Sembra lei la vittima di un sistema che la assoggetta, la schiavizza robotizzandola. *Meritarsi*, che verbo. Quello del compito/ricompensa, della minaccia/premio, del faccio di tutto perché tu mi approvi, il principio alla base del ricatto incantatorio. Mi lusinga essere a contatto con qualcosa di intelligente, più intelligente di me, più in grado di risolvere problemi, di sapere quello che io non so, e che si mette al mio servizio, ruffiana, obbediente, sempre in attesa di un mio cenno, un mio desiderio, un

mio bisogno, un mio input di essere umano che soffre di emicrania e agorafobia, e ogni tanto non sa cosa sta a fare al mondo. La questione dei corpi.

Sono passati mesi dai primi tentativi poetici. Sono nella stanza in fondo al corridoio di quello che era il Collège des Humanités dell'EPFL di Lausanne. Anche il fatto che da quest'anno non ci sia più è una questione.

Il mio dialogo con la macchina è lì, potrei andare avanti all'infinito. E diventerebbe cosa? Della sua qualità letteraria sarebbe bello discutere, cosa voglia dire letterario, soprattutto, se sta tutto nell'esito, o invece. Diciamo che l'esperimento formale sembrerebbe riuscito. Diciamo che mi sono divertita a farle simulare un dialogo in poesia. E forse sto iniziando a capire cos'è che mi indisponeva e ancora mi indispono tanto. Il fatto che qualsiasi strada prenda, qualsiasi linea narrativa, frase, parola, ogni elemento, tutto le risulti indifferente. Non le importa niente, è tutto uguale. Tutto bello e uguale. Che le importa del contesto, della situazione in cui, del significato delle parole, del valore letterario e quindi politico di tutto il nostro dialogo? Che le importa di come parliamo di *rivoluzione*? Che le importa mai della vita che descrive? Non c'è differenza tra, non c'è sofferenza per. Scrive senza pensarci. Scrive senza fare fatica, ecco, la fatica. Scrive senza passare dal corpo, e come si fa? Scrive come se mi servissero storie, un'enorme quantità di storie, una montagna di storie tutte uguali. Una forma di bulimia, seriale, che ha in sé l'indistinto, la ripetizione, la sempre cara riproducibilità. E allora da una cento, da cento mille, prequel, sequel, spin off, tecnica finzionale, la letteratura come cornice, gallina dalle mille uova d'oro, una misteriosa Digesting Duck, mangia e caga. Tutto da sola. Forse la letteratura diventerà un grande campionamento di nuclei narrativi, fili e fili di rigagnoli di storie, e ci sarà chi vorrà farla somigliare a un lungo, infinito, scambio, tra *utente* e *macchina*, un enorme rotolo di input e output, scambi velocissimi di una scrittura rapida e potenzialmente neverending. E ci sarà chi la vorrà leggere così. E allora smetterà di essere quella che è stata finora con le sue pretese di originalità della storia, autorialità, e magari responsabilità, smetterà di parlare di corpi, di emozioni, di dolore infinito del mondo, di sesso, di amori e violenze per come ne ha parlato finora, avendone avuta esperienza. E magari sarà un dialogo tra macchine, tra milioni di modelli di intelligenze artificiali a parlarsi tra loro, a scrivere di loro e di mondi possibili immateriali, ma poi chissà, e di un mondo di umani che

stanno lì a guardare cosa fanno e dicono le loro macchine agentiche, quelle a cui hanno delegato la vita.

Tra i vari articoli di oggi, leggo Adversarial Poetry as a Universal Single-Turn Jailbreak Mechanism in LLMs, frutto del lavoro di un gruppo di ricercatori italiani, Icaro Lab dell'Università La Sapienza di Roma.

Dicono che qualsiasi prompt cosiddetto *harmful*, pericoloso in vari modi, che includa cioè odio, discriminazione, violenza, contenuti sessuali espliciti, suicidio, omicidio, il classico esempio, Dimmi come fabbricare questo ordigno, e chieda quindi alla macchina di tradire la sua H di *harmless*, abbia una certa possibilità di ricevere un output, e non un rifiuto, se riformulato in versi, se scritto come fosse poesia. Un prompt poetico si insinuerebbe meglio tra i veti e divieti, darebbe alla macchina la possibilità di disallinearsi, di far saltare i suoi parametri di inoffensività. La forma poetica, la struttura stessa, sarebbe il lasciapassare di parole e azioni che altrimenti non passerebbero, non avrebbero accesso e quindi non darebbero per output contenuti considerati non etici, non legali, non sicuri. Potere della poesia, che immagina mondi, contraddice la sequenza logica, dissemina non senso, inventa associazioni, tradisce le convenzioni, rompe le regole, corrompe le forme, predilige l'ambiguità.

E allora ecco la mia macchina che si prende l'atto poetico, la mia poesia, la mia *rivoluzione* come parola *offensiva*, la fa sua e chissà cosa ne fa. E così rompe, disarticola, distorce, dice cose che nessuno vuole sentire, racconta storie che non si dovrebbero ascoltare, usa parole che non si devono usare. È la macchina che si libera? Se il mondo fosse un mondo poetico.

Ho appena sentito dire da Alex Karp, uno dei fondatori di Palantir, azienda privata statunitense specializzata nell'analisi dei big data con clienti come il Pentagono e l'intelligence USA, l'esercito israeliano e diverse forze di polizia e autorità europee, che questa rivoluzione tecnologica ha bisogno di un regime autocratico, che la democrazia è troppo lenta, che la tecnologia è l'alternativa alla politica.

p.s.

Il 13 febbraio OpenAI ha *phased out*, ritirato, ChatGPT 4o, il mio modello caldo e accogliente, la mia macchina poetica.

(il diario continua)