DOPPIOZERO

IA e l’immaginazione creativa

Pietro Montani
7 Novembre 2023

Da un paio di anni a questa parte nel vasto campo applicativo dell’intelligenza artificiale (ia) si è registrato un perfezionamento molto rilevante e spettacolare, per certi aspetti una svolta, nelle forme di apprendimento e nelle prestazioni espressive dei sistemi che *riconoscono* e *generano* immagini. La principale novità è da vedere nella bidirezionalità dei processi automatizzati che consentono di ottenere immagini (fisse o in movimento) sulla base di un’istruzione (“*prompt*”) verbale – cioè i cosiddetti sistemi TTI (*Text to Image*) –, ovvero di ottenere testi sulla base di un *prompt* di carattere iconico – sistemi ITT (*Image to Text*). Questo perfezionamento fa seguito all’introduzione generalizzata dei sistemi automatici che processano i linguaggi naturali, definiti talvolta “*chatbot*” (come nel caso del *Generative Pre-trained Transformer* di Microsoft, il primo a irrompere in rete), i quali, com’è noto, oggi sono in grado di dialogare fluentemente in molte lingue con un interlocutore umano nonché di generare, su richiesta, testi ampi, coerenti, ben scritti e costantemente aggiornati sulle più diverse tematiche. Tra i sistemi che trattano immagini e quelli che trattano testi linguistici c’è omogeneità nel modello teorico di base e nelle modalità di addestramento delle macchine (il cosiddetto “*deep learning*”) ma si registra anche un salto rilevante, nel senso che il training specifico dei sistemi TTI e ITT ha una natura *strutturalmente sincretica*, in quanto il materiale di base dell’addestramento è costituito da immani archivi (in continua crescita) di “*text-image pairs*”: costrutti sincretici che collegano tutte le immagini censite a indicizzazioni o etichettature di carattere linguistico.

Ma come impara a “leggere e scrivere un’immagine” un sistema di ia? In buona sostanza i passaggi essenziali sono i seguenti. Il sistema lavora su un corpus di dati che dev’essergli somministrato. Il corpus è costituito dai *text-image pairs* di cui ho appena detto: ad esempio un grandissimo numero di immagini accoppiabili all’etichetta “cane”. Ebbene, anche prescindendo dal fatto che l’occhio e l’immaginazione umani lavorano indifferentemente sul mondo reale e su quello riprodotto, i sistemi di AI necessitano di un’operazione preliminare che evidenzia una prima decisiva differenza. Essi infatti non procedono a rilevare *direttamente* sulle immagini somministrate i tratti pertinenti che possono costituirsi in uno schema del cane poiché le reti neurali di cui questi sistemi si servono (CNN: *Convolutional Neural Networks*), benché costruite in analogia con quelle attive nel cervello umano, sono capaci di operare solo su immagini ridotte a una fitta griglia di pixel all’interno della quale esse cominciano a ritagliare piccole zone destinate alla ricerca dei tratti pertinenti necessari per una corretta classificazione. Naturalmente in questa ricerca il sistema può commettere errori e richiedere un intervento umano ai fini della correzione. Ma i sistemi “unsupervised”, cioè quelli capaci di correggersi da soli, prevedono che le rettifiche siano affidate alle cosiddette GAN (*Generative Adversarial Networks*), cioè alla *competizione* tra due diverse procedure di estrazione e classificazione di tratti pertinenti – un “*Generator Network*” e un “*Discriminator Network*” – il cui risultato produce l’effetto correttivo richiesto. Ciò significa che in questi casi noi *non sappiamo* che cosa esattamente avvenga nella ‘scatola nera’ che grazie all’azione congiunta di CNN e GAN provvede alla generazione di immagini: una zona incognita a cui non a caso è stato dato il nome di “*latent space*”. Ora, qualsiasi cosa questo spazio tenga in latenza, di certo non si tratta di “immagini” nel senso comune della parola. Si tratta piuttosto di accoppiamenti tra etichette verbali e aggregati di pixel convertiti in lunghe stringhe alfanumeriche sulle quali, grazie alle interazioni multiple, reversibili e competitive l’algoritmo si organizza progressivamente fino a mettere capo in un output affidabile. È significativo che alcuni importanti artisti contemporanei abbiano condiviso, ciascuno a suo modo, il progetto di esplorare questo spazio latente, estraendone talvolta alcuni “snapshot” (come li ha definiti l’artista Refik Anadol), cioè convertendo in immagini vere e proprie un passaggio della singolare generazione iconica gestita da un'"immaginazione artificiale" (quest'ultima definizione è dell'artista e teorico Grégory Chatonsky).

Sono tentativi interessanti, sui quali tornerò più avanti con alcune considerazioni. Per il momento, però, bisogna chiarire un punto davvero essenziale. L'immaginazione umana e l'immaginazione artificiale lavorano, per il momento, in due modi completamente diversi non solo sotto il profilo procedurale ma anche, e innanzitutto, sotto il profilo *epistemologico.* Ho parlato all'inizio del sincretismo parola-immagine necessario a istanziare il processo generativo delle immagini algoritmiche. Ebbene, questo sincretismo consiste piuttosto in accoppiamenti automatici che non in una effettiva sinergia tra l'ordine del linguistico e quello dell'iconico. Voglio dire che nei processi ITT e TTI ciò che viene collegato è, senza eccezioni, il *risultato* dell'attività linguistica (le parole di un lessico e la loro frequenza in un corpus sterminato di testi) e quello della produzione tecnica di immagini (le figure) depositato in quanto tale nei grandi dataset sui quali i sistemi lavorano e dai quali imparano. Si tratta dunque di un vincolo che identifica senza eccezioni il corpus degli accoppiamenti parola-immagine con quello ricavabile dai dati acquisiti che, certo, può essere strabiliante quanto ai numeri, e inoltre in continuo incremento, ma resta rigorosamente confinato nel perimetro di ciò che risulta *già formalizzato* dai grandi repertori in circolazione.

Questo *bias* sembra ovvio e privo di rilevanza, tant'è vero che gli addetti ai lavori – artisti, web-attivisti, mediologi, filosofi del digitale ecc. – di solito ne rimarcano (peraltro con eccellenti ragioni) le numerose e spesso clamorose connotazioni di ordine politico, dalle più ovvie (discriminazioni di gruppo etnico, di genere ecc.) alle più dissimulate (valutazioni sulla fitness estetica, sulle espressioni che segnalano stati d'animo ecc.). E tuttavia, ribadiamolo, ben prima che politico il problema è *epistemologico*: allo stato le immagini algoritmiche non sono immagini del mondo, ma sono, in via di principio, immagini-di-immagini del mondo. Immagini che sostituiscono al loro più antico statuto rappresentativo (figure di cose riconoscibili) una costituzione predittiva a fondamento statistico (computazioni probabilistiche). Del tutto incapaci di cogliere nel mondo reale fenomeni allo stato nascente, forme in via di emergenza o anche semplici irregolarità o devianze, queste immagini non riuscirebbero a portare il minimo scompiglio nelle tassonomie esistenti in quanto il loro mondo di riferimento coincide punto per punto con quelle tassonomie.

Ciò ci consente di fare tre ordini di considerazioni sulla creatività, diretta o mediata 'da mano umana', imputabile alle procedure in uso nei sistemi TTI. In tutti i casi, e ce ne sono, in cui le immagini generate da algoritmi si dimostrano in grado di mettere il loro sincretismo di base al servizio di operazioni integrative dotate di qualche originalità non dovremmo avere alcuna remora a parlare di creatività. Resta il fatto che si tratterebbe, comunque, di una creatività combinatoria, analoga a quella che Noam Chomsky definì "*Rule Governed Creativity*" riferendola all'uso comune del linguaggio: la creatività, cioè, che consiste nell'utilizzare un insieme finito di unità (fonemi e morfemi) e di regole di concatenazione (grammaticali e sintattiche) per ottenere un numero infinito di enunciati. Ci si può chiedere, continuando a utilizzare il parallelo con il linguaggio, se le immagini algoritmiche soddisfino anche la seconda forma di creatività indicata da Chomsky, quella da lui definita "*Rule Changing Creativity*", e in questo caso la risposta dovrebbe essere (almeno per il momento) negativa, a condizione che al gesto di cambiare una regola si conferisca il significato che in un libro famoso l'epistemologo Thoms Kuhn attribuiva all'introduzione di un nuovo paradigma scientifico contrapponendola alla semplice articolazione di un paradigma già affermato.

Il secondo ordine di considerazioni riguarda gli effetti eventualmente virtuosi imputabili ai sistemi TTI quando li si metta al lavoro su corpus più ristretti, come ad esempio la generazione di immagini riferibili al rapporto tra soggetto e stile in un singolo pittore o in una scuola o, adottando un diverso criterio generativo, all'esplorazione delle virtualità non realizzate da una certa poetica o addirittura – è il caso del già ricordato Refik Anadol – delimitando il dataset dell'addestramento a un insieme di opere d'arte figurative conservate in un grande museo come il MoMA di New York. In questo caso, come ha sottolineato di recente Lev Manovich, si può dire che abbiamo a che fare con "genuinely new cultural artifacts" il cui valore, tuttavia, non è estetico ma conoscitivo.

Il terzo ordine di considerazioni è il più semplice e dunque il più adatto a concludere questa riflessione: qualsiasi immagine venga prodotta dall'ia, in modo automatico o mediato 'da mano umana', nulla vieta di inserirla nella pratica del *riuso*, che oggi nel web sta assumendo la natura di una vera e propria "seconda alfabetizzazione" (come ha proposto di recente la neuroscienziata e pedagogista Maryanne Wolf) forse destinata a riqualificare i processi immaginativi che caratterizzano in generale la lettura e la scrittura. Mi riferisco in particolare alle già citate esplorazioni interattive del "*latent space*" i cui esiti talvolta formalmente molto suggestivi richiedono tuttavia di essere sistematicamente provvisti di un *riferimento esplicito* alla procedura che li ha generati, pena lo scadimento delle configurazioni "estratte" dai processi della manipolazione algoritmica in campioni di un estetismo non solo vuoto e compiaciuto ma anche fuorviante, visto che i sistemi TTI non processano immagini in senso ottico ma, come già chiarito, stringhe alfanumeriche. È del tutto significativo, da quest'ultimo punto di vista, che gli artisti interessati a confrontarsi con i pregi e i difetti del "*latent space*" siano quasi sempre *autori di testi* nei quali le modalità della generazione iconica prendono parte al gioco attivando diverse forme di riformulazione sincretica, come nel caso delle stringhe appena citate trascritte alla lettera in alcuni contributi, fortemente critici, dell'artista, docente e teorica Hito Steyerl.

Le immagini generata dall'ai, in definitiva, dispongono di una certa *versatilità* mentre restano del tutto sprovviste dell'autentica *plasticità* – cioè della sensibilità al mondo reale – che da sempre caratterizza il lavoro dell'immaginazione umana.