

Sulla natura dell'immagine: Language Model e Visual Language Model

Luca Capone*, Felice Cimatti^{***}, ^{***}

ABSTRACT

L'articolo propone una riflessione sulle capacità di traduzione intersemiotica (tra espressioni grafiche e linguistiche) esibite dai Visual Language Model. Partendo da questo progresso tecnico vengono tratte alcune conclusioni sulla comune natura semiotica di immagine e discorso.

La tesi sostenuta è che le espressioni grafiche e linguistiche condividono una comune e costitutiva indeterminatezza, essendo entrambe determinabili sotto molteplici rispetti (potendo assumere molteplici sensi). L'articolo, smarcandosi da una concezione referenzialista del linguaggio, fa leva sulla teoria semiotica del significato. Secondo questa prospettiva, le immagini (come le espressioni verbali) non rimandano direttamente a cose, ma a significati. Da questo punto di vista, la traducibilità tra immagini e parole non risulta un'anomalia, bensì l'ordinario effetto della loro comune radice semiotica. L'apparente "eccedenza" dell'immagine rispetto all'espressione verbale non segnala un'irriducibile superiorità del visibile, ma il fatto che tanto le immagini quanto i segni linguistici sono sempre aperti alla determinazione.

I Visual Language Model sono un caso esemplare di questa dinamica. In questi strumenti, espressioni grafiche e linguistiche sono organizzate in funzione di un comune sistema semiotico (la lingua). La generazione di immagini a partire da testi (e viceversa la produzione di testo a partire da immagini) mostra che la percezione, nella sua forma umana e artificiale, è già sempre mediata da processi di significazione. Formule come "il visibile non è riducibile al dicibile" non descrivono quindi un fatto empirico, ma si limitano a esprimere una *regola grammaticale*, un pregiudizio della riflessione sul rapporto fra immagini e linguaggio.

KEYWORDS

Semiotics, Language Model, Computer Vision, Transmediality

* luca.capone@fileli.unipi.it, CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa

** felice.cimatti@unica1.it, Dipartimento di Scienze Politiche e Sociali, Università della Calabria

*** I due autori hanno discusso, progettato e scritto il saggio congiuntamente. Tuttavia, i paragrafi 1 e 6 sono da attribuire a Felice Cimatti, mentre i paragrafi 2, 3, 4, 5 sono da attribuire a Luca Capone.

1. Dalle parole alle immagini, e viceversa



Figura 1 – Immagine generata con ChatGPT

Il *prompt* verbale, ossia una sequenza sintatticamente sensata, era molto semplice: “produci una immagine di un gatto con due code che cammina sopra un ramo di una quercia in autunno”. Il risultato è questa immagine (figura 1). Dalla scrittura all’immagine. Proviamo ora a rovesciare la richiesta: si tratta adesso di descrivere questa stessa immagine, ossia a trasporre una immagine in parole scritte. Ecco la risposta di ChatGPT: “L’immagine rappresenta una scena autunnale molto suggestiva e stilizzata, quasi pittorica. Al centro della composizione c’è un gatto tigrato dal manto marrone scuro con striature nere. Il gatto è ritratto in movimento, mentre cammina con grazia e sicurezza su un ramo spesso e nodoso di un albero. Ha uno sguardo attento e curioso rivolto verso sinistra, come se avesse percepito qualcosa nei dintorni. Un elemento particolarmente distintivo è la coda del gatto, che si biforca in due punte fluenti, creando una forma elegante e quasi fantastica, dando un tocco surreale o fiabesco all’immagine. Lo sfondo è dominato da foglie di quercia arancioni e giallo-oro, tipiche dell’autunno. Queste foglie riempiono l’ambiente, creando un’atmosfera calda e accogliente, ma anche leggermente magica”.

In questo articolo ci poniamo una classica domanda filosofica, una domanda relativa alle condizioni di possibilità di questo passaggio, apparentemente fluido e ‘semplice’, dalla parola all’immagine e viceversa. Com’è possibile trasformare una descrizione verbale in una immagine? Com’è possibile fare l’operazione inversa? E soprattutto, *perché* è possibile? Che l’immagine del nostro esempio sia stata essa stessa generata da una intelligenza artificiale non cambia la natura del problema, il sistema (ci riferiamo con questo termine neutro al *software* capace di compiere queste operazioni) è in grado

di descrivere qualunque immagine, sia prodotta da esseri umani che artificiali, non c'è differenza. Come spesso accade in filosofia, la questione delle origini non risulta risolutiva, perché o le origini sono sempre operative, e allora non sono propriamente originarie, oppure non ci spiegano perché dovrebbero essere efficaci anche nel presente.

La questione che ci poniamo in questo lavoro è invece rilevante perché spesso si ritiene, o si crede di ritenere, che l'immagine sia sostanzialmente diversa dalle tipiche entità del linguaggio verbale, ad esempio nomi e descrizioni; secondo un modo preteorico di intendere l'immagine questa rientrerebbe in un campo di fenomeni diverso da quello del linguaggio. Il punto in questione non è quello della differenza fra segni arbitrari e segni iconici, perché come sappiamo da Peirce anche questi ultimi sono, a loro modo, arbitrari (la nozione di somiglianza non è una nozione primitiva). Ma questo appunto significa – di questo ci occuperemo in questo lavoro – che fra immagini e parole esiste una analogia semiotica di fondo. In effetti l'esempio con cui abbiamo aperto questo articolo mostra che forse tutta questa differenza non c'è, perché altrimenti come si sarebbe potuta trasformare una immagine in parole, e viceversa? Allo stesso tempo una antica tradizione di pensiero, o forse meglio una tenace sensazione, ci lascia pensare che ci sia comunque 'qualcosa', nella immagine, che la corrispondente descrizione verbale non riesce a catturare. Il visibile non è riducibile, almeno non del tutto, al dicibile. Un'affermazione che per un verso è evidente, per un altro, però, non convince fino in fondo, e un caso come quello da cui hanno preso spunto queste considerazioni sembra confermarlo. Il problema che proviamo a discutere in questo lavoro, allora, è il seguente: "il visibile non è riducibile al dicibile" è una descrizione di che cos'è il visibile, oppure "il visibile non è riducibile al dicibile" è una definizione del visibile?

La posta in gioco è allora la natura dell'immagine. La 'traducibilità' fra immagini e parole va nella direzione di una loro comune natura semiotica. È possibile passare dalle prime alle seconde perché si tratta sempre di segni, sia pure diversi. Ed è inscritta nella natura dei segni che possano trasformarsi, in modo più o meno soddisfacente, gli uni negli altri; da questo punto di vista la prestazione di ChatGPT sarebbe, almeno in linea di principio comprensibile (non ci esprimiamo sulla complessità algoritmica dell'operazione che, nonostante sia sempre evocata, non è pertinente per la nostra discussione: in queste pagine seguiamo la prospettiva del Wittgenstein delle *Ricerche filosofiche*, secondo la quale 'sostanza = funzionamento'). Se l'immagine è un segno allora rimanda, in

un modo o nell'altro, a qualcos'altro, l'oggetto a cui il segno si riferisce, secondo il classico modello semiotico dell'*aliquid stat pro aliquo*. Ma se l'immagine è soprattutto un segno allora la differenza fra segni verbali e non verbali, come appunto il caso dell'immagine, è meno importante di ciò che li accomuna, la comune natura di segni, cioè di qualcosa che rappresenta qualcos'altro. Si apre a questo punto un ulteriore problema, perché ciò a cui il segno si riferisce – il corrispondente oggetto; ad esempio, ciò che intendevamo con il *prompt* “produci una immagine di un gatto con due code che cammina sopra un ramo di una quercia in autunno” – può essere inteso in due modi completamente diversi: secondo la tradizione che possiamo chiamare ‘referenzialista’, l'oggetto a cui il segno si riferisce è propriamente una cosa, o al più una classe di cose (De Mauro 1970, p. 42). Prendiamo il caso dell'immagine qui sopra: secondo questa concezione il segno-ritratto si riferisce ad un gatto effettivo su un ramo di una quercia. Secondo un'altra concezione – quella propriamente semiotica, e che vale sia per la tradizione che risale a Peirce che per quella che rimanda a Saussure – l'oggetto a cui il segno si riferisce non è una cosa, cioè un'entità non semiotica, bensì un'entità, che qui chiameremo ‘significato’, *anch'essa semiotica*. Mentre nella tradizione ‘referenzialista’ fin qui richiamata il significato del segno si colloca al di fuori della semiosi, in quella semiotica che le si contrappone, il significato di ogni evento segnico rimane del tutto all'interno della semiosi.

È evidente che il modo di intendere la portata semiotica delle immagini generate dall'IA cambia completamente a seconda di quale modello di semiosi si adotti. Nel modello ‘referenzialista’ un'immagine generata da un qualunque sistema di IA non potrà essere intesa come un segno in senso pieno (e quindi nemmeno un'immagine in senso effettivo), appunto perché i modelli che l'hanno ‘generata’ non ‘sanno’ nulla del mondo non semiotico (con una terminologia antica, la loro è solo una conoscenza *de dicto*, non *de re*). Secondo questa prospettiva, ad esempio, «We argue that the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning [...] We take meaning to be the relation between the form and something external to language» (Bender e Koller 2020). Se invece collochiamo un qualunque *Visual Language Model* (VLM) nel quadro teorico di Peirce o di Saussure (trascurando le ovvie differenze fra questi due approcci), non avremo nessuna difficoltà a riconoscerle come immagini a pieno titolo, perché in questa tradizione di pensiero un'immagine ‘funziona’ come immagine, e quindi come dispositivo semiotico, a prescindere da qualunque riferimento a presunte entità esterne non semiotiche.

Prendiamo, ad esempio, il caso di Peirce: nel suo modello semiotico l'immagine rientra nella classe delle icone, e l'icona è un segno esattamente allo stesso titolo di un simbolo o un indice, e per tutti questi segni vale che il loro 'significato' è costituito dal cosiddetto 'interpretante' che lo collega al suo 'oggetto': il carattere specifico dell'interpretante, una caratteristica che non smette di stupire la tradizione 'referenzialista', è di essere a sua volta un segno. È vero che per lo stesso Peirce l'icona svolge un ruolo particolare nel processo conoscitivo, quello iniziale, abduttivo, collocandosi quindi ai bordi estremi della semiosi (non a caso l'abduzione si collega alla categoria della Primità). Tuttavia un'icona è un segno come gli altri. Una delle conseguenze più indigeribili – per la prospettiva referenzialista – di questa posizione è che rende inutilizzabili o più semplicemente superflue le nozioni di 'verità' e 'falsità' per rendere conto delle immagini. L'immagine del gatto con le due code è 'falsa' nel mondo dei gatti di cui abbiamo esperienza (anche se in rete si trovano non poche 'immagini' di gatti 'veri' con due code), ma non lo è in 'altri' mondi, come quello letterario ad esempio. Ma è proprio questo il punto, l'immagine è davvero, e principalmente, un segno? Si può in effetti pensare in modo diverso l'immagine, senza includerla in tutto e per tutto nella classe più generale dei segni. O meglio, si può continuare a considerarla in qualche modo un segno – perché nell'immagine ci viene 'naturale' vedere una rappresentazione di qualcos'altro, e senza questa caratteristica non potremmo parlare nemmeno di immagini – senza però identificarla completamente con un segno. In questo modo se ne mantiene la natura rappresentativa, senza però assimilarla in tutto e per tutto ai segni più 'classici', come gli indici e i simboli. Secondo questa prospettiva l'immagine è rappresentativa, altrimenti non sarebbe una immagine, senza però essere del tutto assimilabile ad un segno. Questa distinzione, ammesso che sia praticabile, dovrebbe servire a garantire all'immagine uno statuto particolare, possibilmente autonomo, rispetto ai segni linguistici. L'immagine avrebbe allora uno statuto affatto peculiare, allo stesso tempo segnico, ma anche non segnico.

La posta in gioco, in effetti, è il rapporto fra l'immagine e il linguaggio verbale. L'immagine è autonoma, in tutto o in parte, rispetto ai segni linguistici? Da come si risponde a questa domanda derivano atteggiamenti molto diversi rispetto alla generazione di immagini da parte di VLM. Per chi ritiene che l'immagine non rientri nel campo dei segni, allora le immagini generate dall'IA sono perfetti esempi di immagine, proprio perché il sistema che le genera non 'sa' nulla del mondo cosiddetto 'reale'. Ma la 'facilità' con cui una intelligenza artificiale come ChatGPT passa dalle parole

all'immagine e viceversa, rende poco plausibile questa posizione. Al contrario, è proprio questa facilità che rende molto convincente l'assimilazione delle immagini al campo della semiosi. Qual è la prestazione specifica della semiotica, in fondo, se non proprio di porre in questione la distinzione fra *de dicto* e *de re*?

Se l'immagine è un segno in senso pieno, come ritiene Peirce, allora nel passaggio dall'una all'altro non si dovrebbe perdere, almeno in linea di principio, alcuna informazione, e quindi avremo una sostanziale equivalenza fra una immagine e la sua corrispondente descrizione verbale: l'immagine sarebbe una sorta di discorso senza parole, e una descrizione sarebbe una sequenza di parole senza figure. Questa formulazione è evidentemente paradossale, tuttavia al fondo i programmi che trasformano descrizioni verbali più o meno accurate in corrispondenti immagini e, al contrario, trasformano immagini in descrizioni funzionano in base a questo fondamentale principio semiotico: le immagini sono segni, e siccome il modello della semiosi è quello simbolico (cioè quello dei segni linguistici, come sosteneva Barthes), ne deriva la sorprendente conclusione che la trasposizione di immagini in parole e parole in immagini, per quanto possano sembrare entità appartenenti a generi molto diversi fra loro, non pone particolari problemi. In questo senso la 'novità' dei sistemi di IA non è poi così innovativa, dal momento che in fondo non fanno che rendere esplicito il comune fondamento semiotico di immagini e parole. Come scrive Garroni, non rimane che accettare che "*percezione*" – in quanto attività che produce immagini – "*e linguaggio si condizionino a vicenda*, che l'uno supponga l'altra e quella supponga questo, che essi insomma costituiscano una stretta correlazione" (Garroni 2005, p. 41).

Tuttavia questa apparente semplicità si basa su un'operazione, che è allo stesso tempo economica linguistica e semiotica, l'annotazione di dati. Nel caso più brutale si tratta del lavoro più o meno nascosto di migliaia di esseri umani a cui vengono sottoposte delle immagini che devono trasporre in parole: l'operazione nota nella nostra tradizione culturale come *ekphrasis* (*εἰκφρασις*). Si prenda il caso classico del Fedro, quando Socrate descrive il paesaggio sotto il platano dove vuole ascoltare le parole di Fedro: "Per Era, che bel posto! Quel platano così alto e pieno di fronde; e quell'agnocasto, lui pure così alto, con quell'ombra meravigliosa; è in piena fioritura e riempie di profumi tutto intorno. E questa sorgente davvero graziosa che scorre sotto il platano! Le sue acque sono proprio fresche, ne fanno prova i miei piedi. [...]. Senti poi, te ne prego, la brezza di questo luogo quanto è amabile e davvero piacevole: è il canto estivo che risponde al coro delle cicale" (*Phaedrus* 230 c-d).

Nei termini attuali, anche se potrà sembrare irrispettoso, Platone nello scrivere il *Fedro* – in particolare quello che sta vedendo Socrate sotto il platano – sta appunto compiendo un’operazione di ‘annotazione’: sta trasponendo un’immagine (mentale) di un luogo in un discorso, ossia una sequenza di parole, di *note* appunto, che trasformano una scena visiva in una sequenza di parole. Messa in questi termini sembrerebbe immediatamente evidente la sproporzione fra la ricchezza della scena vista e partecipata con tutto il corpo e la sua resa verbale, che per quanto accurata sembra appunto essere affatto sottodeterminata rispetto a quella profusione di colori, profumi ed esperienze tattili.

Ma è proprio così? È vero, come sostiene una formula mille volte ripetuta, che “a picture is worth a thousand words”? (Hasenmueller 1989). Questa formula è una descrizione contrastiva dell’immagine rispetto al linguaggio verbale, o una definizione dell’immagine (Cometa 2011)? Secondo Garroni nell’immagine troviamo “*l’unione del determinato e dell’indeterminato*” (Garroni 2005, p. 26), perché per un verso un’immagine è sempre “incompleta” (*Ivi*, p. 21), peraltro però – e proprio grazie all’interazione con il linguaggio verbale – è anche sempre ulteriormente determinabile, precisabile. In effetti, per tornare ai programmi che ‘generano’ immagini, più il *prompt* è dettagliato più l’immagine ‘prodotta’ sarà corrispondentemente determinata. Tuttavia, e questo è un altro punto di vicinanza fra immagine e parola, “determinato e indeterminato sono complementari. Là dove c’è una qualche determinazione deve esserci anche una correlativa indeterminazione” (*Ivi*, p. 33). Ci troviamo così di nuovo in quello che Garroni chiama efficacemente “l’enigma della percezione” – in quanto facoltà delle immagini: “l’immagine è per un verso determinata e per un altro indeterminata” (*Ivi*, p. 25). Quello che l’immagine con cui abbiamo aperto queste considerazioni mostra allora è proprio questa situazione: vicinanza sostanziale con la parola – quella vicinanza che permette la ‘traducibilità’ dell’una all’altra e viceversa – ma anche differenza, nel senso che l’immagine è comunque incompleta, costitutivamente incompleta, sempre ulteriormente determinabile. Ma questa stessa differenza si rivela in realtà una nuova vicinanza, perché anche la parola è altrettanto indeterminata. Se ci liberiamo di quella che Eco chiamava “la fallacia referenziale” (Eco 1975, §2.5.1) – cioè dalla ricerca di una cosa come terminale della semiosi – ci accorgiamo così che le immagini generate dall’IA condividono con i testi prodotti dalla IA uno stesso doppio carattere: si tratta di costrutti allo stesso tempo indeterminati e determinati (o meglio, sempre ulteriormente determinabili).

2. Il concetto di indeterminazione

Nel paragrafo precedente lo statuto delle immagini è risultato conteso fra due prospettive antagoniste. Quella referenzialista, per cui l'immagine trova il suo termine correlativo in un oggetto reale, e quella semiotica, per cui le immagini sono assimilabili a segni che, in quanto tali, rimandano ad altri segni interpretanti. Le prestazioni dei modelli di IA sembrano rendere plausibile la *vicinanza sostanziale* fra parola e immagine, condizione della traducibilità (automatica e non) fra le due. Questa vicinanza lascia aperte due questioni: la prima riguarda le condizioni di possibilità della *traduzione* fra immagini e parole (§3), la seconda è relativa all'eventuale scarto di questo processo di conversione, posto dall'apparente ricchezza dell'immagine, a confronto con la sinteticità del segno linguistico (§4). La risoluzione di entrambi i problemi dipende dalla chiarificazione del carattere *indeterminato* di parole e immagini. Prima di passare alla loro risoluzione è necessario quindi specificare in che modo parole e immagini sono fondamentalmente indeterminate.

2.1 Indeterminatezza delle immagini

Il caso più semplice è quello dell'immagine¹. L'immagine presenta un numero, se non infinito, almeno difficilmente numerabile di tratti potenzialmente pertinenti (cfr. De Mauro 2024, pp. 14-15). Questo numero elevato di tratti costituisce un problema, ovvero come selezionare alcuni di essi ai fini dell'identificazione (o della descrizione) dell'immagine. Il fatto che i tratti siano in numero elevato fa sì che l'immagine possa essere identificata (pertinentizzata) seguendo diverse strategie (prospettive). In altri termini, la medesima immagine può costituire il *senso* di un numero potenzialmente infinito di espressioni: *un gatto comune, un animale in equilibrio su un ramo, allegoria dell'autunno, esempio di gatto irrealistico, immagine dai colori caldi* ecc. Da questo punto di vista l'immagine costituisce una materia aperta, disponibile a molteplici determinazioni, più o meno avventurose. Chiaramente, la materialità dell'immagine rende implausibili alcune pertinentizzazioni, tuttavia, fra quelle ac-

¹ Gli autori sono consapevoli della differenziazione figura-immagine operata da Garroni: l'immagine (interna e dinamica), la figura (esterna e statica). Dato che i modelli di IA hanno a che fare sempre e solo con figure determinate (non con immagini interne), l'articolo si concentra su queste. Tuttavia, nella letteratura la differenziazione garroniana non trova particolare spazio. È stato quindi ritenuto opportuno usare la parola immagine per riferirsi a queste espressioni grafiche determinate. Di conseguenza, l'obiettivo dei seguenti paragrafi è la descrizione del rapporto fra i segni linguistici e queste espressioni grafiche determinate (da qui in avanti *immagini*).

cettabili, nessuna in quanto tale è in grado di imporsi sulle altre. A causa di questa plurivocità di fondo, le immagini sono caratterizzate da Garroni come indeterminate (2005, p. 47, plurinterpretabilità delle figure). È importante sottolineare che nessuna di queste determinazioni possibili è data in quanto tale, ma tutte dipendono da un secondo fattore, ovvero dalla sussunzione dell'immagine da parte di un significato.

2.2 Indeterminatezza dei segni linguistici

Per quanto riguarda la parola, la situazione è più delicata a causa della diffusione di quella che è stata definita *fallacia referenziale* (Eco 1975, §2.5.1), ovvero l'idea secondo cui il significato di una parola è l'oggetto per cui essa sta. Da questo punto di vista il significato di una parola come *mela* sarebbe l'oggetto a cui si riferisce. Vale la pena ribadire che una mela determinata può costituire il senso puntuale di una espressione pronunciata *in un determinato contesto* ("passami la mela che sta sul tavolo"), tuttavia la possibilità di riferirsi a oggetti specifici (assumere come sensi oggetti puntuali) dipende da un fenomeno più generale e indeterminato, ovvero dal significato.

Il significato non è un oggetto determinato ma è un concetto, ovvero un termine generale e astratto. Un termine o un'espressione in isolamento (ad esempio "la mela è rossa") non ha un senso determinato, non si riferisce ad alcun oggetto specifico, ma possiede solo un significato (Eco 2016, p. 351). Le espressioni, *in determinati contesti*, possono essere utilizzate per veicolare sensi determinati, riferendosi ad esempio a oggetti specifici, ma sempre sulla scorta di un significato precedentemente dato. La confusione fra significato (astratto e generale) e oggetto o referente (determinato e individuale) richiama quella evidenziata nelle *Ricerche filosofiche* fra significato del nome e portatore (Wittgenstein 2014, §40). È inoltre importante sottolineare che un significato generale non deve necessariamente incarnarsi in sensi referenziali (riferimento a oggetti), ma molti altri usi delle parole sono possibili: raccontare una storia, fornire una definizione, dare un comando ecc.

Eco descrive il significato come un'unità concettuale che una cultura ha differenziato da altre (Eco 1975, §2.6). Questa definizione è importante in quanto consente di concepire il significato non come un'unità in isolamento, ma come parte di un sistema di differenziazioni solidali, le quali articolano il piano del contenuto (o dei significati) di una lingua. Inoltre, queste articolazioni non sono imposte dall'esperienza (così come l'immagine non poteva imporre

la sua identificazione), ma sono ricevute dai parlanti nella lingua che apprendono (cfr. Violi 1997). È opportuno insistere su questi due punti: solidarietà e arbitrarietà delle articolazioni del piano del contenuto. Ad esempio, le tassonomie ortofrutticole, al pari di ogni altra organizzazione del continuum dell'esperienza, sono il frutto dell'opera articolatoria delle lingue storico naturali. Nell'italiano, da un punto di vista rigoroso, la categoria della frutta seleziona dei tratti in funzione di criteri di classificazione botanici, per cui è *frutta* il prodotto della modificazione dell'ovario della pianta. Questa classificazione assimila le mele alle melanzane ed esclude le fragole e le angurie. In questo caso, l'accezione botanica di *frutta* ha selezionato un criterio, fra i molti possibili, per articolare una parte del piano del contenuto. Tuttavia, nell'uso comune il significato *frutta* si attesta su criteri culinari, assimilando mele, fragole e angurie, ma escludendo la melanzana. In breve, l'accezione comune pertinentizza altri tratti dell'esperienza, la quale si offre a molteplici determinazioni possibili (Eco 2016, p. 70). Le due accezioni di *frutta* presentate esemplificano il modo in cui i significati articolano il piano del contenuto in maniera *arbitraria*, potendo cioè scegliere fra diverse possibili strategie. Inoltre, l'esempio evidenzia la solidarietà fra le articolazioni. Ovvero, cambiando il criterio di pertinentizzazione del contenuto di un significato (*frutta*), cambiano i rapporti reciproci fra gli altri significati del sistema (*mela*, *anguria*, *melanzana* ecc.). Riassumendo, il piano del contenuto della lingua delinea unità culturali, distinguendo i contenuti che una cultura ha eletto come elementi distinti rispetto ad altri (Eco 1975, §2.6). Queste unità sono solidali, nel senso che si determinano in funzione dei rapporti reciproci. Al tempo stesso sono arbitrarie, nel senso che altre strategie di articolazione del piano del contenuto sarebbero state possibili. Lo stesso discorso potrebbe esser fatto per il significato *gatto*, il quale isola all'interno della famiglia dei felidi una specifica sottocategoria, in base a criteri eterogenei selezionati arbitrariamente.

Questi esempi, oltre a chiarire lo statuto infralinguistico (o semiotico) del significato, ne evidenziano anche il carattere generale e indeterminato. In primo luogo, un significato può pertinentizzare *alcuni* tratti, lasciandone molti altri indeterminati. Il significato *gatto* annovererà molti tratti utili all'identificazione di questi animali (un certo aspetto, un modo di incedere, nozioni sulle loro abitudini ecc.), lasciandone tuttavia molti altri vaghi (colore) o del tutto non specificati (numero di miagolii medi giornalieri). Questo vale sia per i significati comuni, che per quelli più determinati e specifici, come la definizione geometrica del quadrato, la quale non contiene

specifiche relative alle dimensioni, non pertinenti per la definizione delle forme geometriche e delle loro proprietà. Questo spazio di indeterminazione consente a un significato di incarnarsi in molti sensi determinati (ad esempio riferendosi a molti gatti o quadrati diversi). Infine un significato, sebbene costituisca un termine generale, non necessariamente deve possedere le caratteristiche di una classe, i cui membri devono disporre di un numero di caratteristiche necessarie e sufficienti (De Mauro 2024, p. 104). I significati sono affini al concetto di *famiglia* delle *Ricerche filosofiche* (Wittgenstein 2014, §67), i cui membri sono tenuti insieme da somiglianze eterogenee e intermittenti. Ovvero, i tratti pertinenti di un significato possono giocare ruoli diversi a seconda del contesto, per la produzione di sensi nuovi e diversi (De Mauro 2024, p. 118).

Ora che è stato chiarito il tipo di indeterminatezza che pertiene a significato e immagine è possibile descrivere in che modo queste due entità possono entrare in rapporto reciproco.

3. Cose, parole e immagini

In *Immagine, linguaggio, figura*, Garroni (2005) cerca di mettere a fuoco la questione del rapporto fra percezione e linguaggio. Introducendo l'argomento, descrive l'esperienza ordinaria di qualcuno che scorra in rassegna con lo sguardo alcuni oggetti, senza un interesse determinato. Garroni riconosce una certa identità agli oggetti così esperiti, tipica non ancora del concetto (il quale pone l'identificazione sotto un rispetto particolare), ma dell'aggregato. In questo senso la percezione, differenziandosi dal linguaggio e senza rifarsi esplicitamente ad esso, in qualche modo lo anticipa e lo richiede (Garroni 2005, p. 10). La questione del riconoscimento richiama il linguaggio, nel senso che la percezione fornisce il materiale indeterminato, ma determinabile, su cui agisce il linguaggio. "Offre al linguaggio il materiale e la garanzia oggettiva dei significati, ma non ancora significati" (*Ivi*, p. 52). In altre parole, nell'assoluta singolarità e plurideterminabilità dell'esperienza sensibile, la lingua determina un ordine, selezionando certi tratti fra quelli disponibili e trascurandone altri. In questo modo il percepito individuale è ricondotto a concetti generali (Garroni parla di classi o famiglie. *Ivi*, p. 10) consentendo, fra le altre cose, il riconoscimento di gatti e alberi in quanto esemplari dei significati "gatto" e "albero" (cioè sempre dentro un'organizzazione semiotica dell'esperienza). Garroni ovviamente non intende dire che nella percezione non si verifica alcuna individuazione o organizzazione

dell'esperienza. Tuttavia, questa organizzazione percettiva è solo l'inizio, ancora indeterminato, del processo di riconoscimento. L'autore insiste su questo punto approfondendo la differenza fra percezione e linguaggio. Secondo Garroni, la prima è destinata a rimanere legata agli oggetti che di volta in volta le si oppongono e ai loro aspetti sensibili attualmente presenti. D'altro canto, il linguaggio (e specifica, qualsiasi linguaggio, anche quello *iniziale*) attiva necessariamente la "determinazione semantica e intellettuale di quegli oggetti, dei loro aspetti e delle relazioni tra di essi e tra soggetti e oggetti" (*Ivi*, p. 52). Ovvero, il linguaggio riconduce la percezione a un significato generale, il quale è in relazione con tutto il sistema di concetti che la lingua mette a disposizione. In questo modo entra in gioco un sistema di relazioni che allontana l'esperienza dalla sua iniziale prossimità con il sentito e il percepito. In questo caso Garroni non pensa solo alle determinazioni specialistiche, che ad esempio concettualizzano un gatto come un *mammifero carnivoro della famiglia dei felidi*, ma a qualsiasi determinazione che la lingua sia in grado di fornire (anche "mitologica o magica", *Ibid.*). In breve, affinché sia possibile riconoscere un gatto come un *gatto* (e non semplicemente come un aggregato di tratti individuali, al più riconducibili a esperienze passate), è necessaria un'astrazione dall'occorrenza sensibile puntuale, per rapportarla a un significato (termine generale) e metterla in collegamento con il resto dell'orizzonte di senso offerto da una lingua storico naturale. Questo è un lavoro dai risultati non scontati. L'esperienza non può causare autonomamente il riconoscimento, ma è di per sé aperta a possibili determinazioni. Per apprezzare la necessità di questa sussunzione concettuale, basti pensare alle peripezie descrittive di chi si è trovato di fronte a un *qualcosa* che violava il sistema di significati a disposizione fino a quel momento. Sono casi noti quelli dell'ornitorinco e della descrizione del rinoceronte di Marco Polo, riferiti da Eco (2016, p. 77). Sul versante grafico, Gombrich (2002) riporta molti esempi di illustrazioni influenzate da concetti familiari. Castel sant'Angelo rappresentato con i tratti di una Burg da un illustratore tedesco (*Ivi*, p. 79) e diversi altri esempi di animali esotici familiarizzati (*Ivi*, pp. 87-89). In tutte queste storie, i protagonisti hanno ricondotto la loro esperienza puntuale al sistema di significati a loro disposizione.

Da un punto di vista ontogenetico, la forma di organizzazione aggregativa, propria della percezione pre-linguistica, convive a lungo con quella linguistica, esibendo formazioni di compromesso (il linguaggio *iniziale* di cui parla Garroni 2005, p. 52). Ne è un esempio classico il peculiare utilizzo delle parole da parte di bambini molto

piccoli. Questi impiegano un medesimo termine per indicare cose diverse, legate da rapporti eterogenei e talvolta difficili da individuare (Vygotskij 1990, p. 171, riporta la parola ‘vau-vau’, usata per indicare *un cane, un boa di pelliccia, una figurina di porcellana raffigurante una ragazza, i bottoni di una giacca* ecc. Garroni 2005, p. 54, parla del termine ‘scarpa’ usato per *mettersi le scarpe, alzarsi dal letto, andare a passeggio*). Ci vorrà del tempo prima che i bambini imparino ad articolare l’esperienza al modo degli adulti (ovvero secondo concetti).

In base a quanto sostenuto da Garroni, per un essere umano la situazione con l’immagine del gatto con due code (figura 1) sembrerebbe la seguente. L’immagine presenta molteplici tratti selezionabili. La sussunzione linguistica seleziona certi tratti, riportando il percepito a un significato generale e consentendo così il riconoscimento e la descrizione del contenuto. In accordo con la concezione semiotica dell’immagine, questa rimanda non a un oggetto, ma a un significato. Effettivamente l’immagine non corrisponde ad alcun gatto reale, tuttavia la situazione non sarebbe cambiata con una fotografia di un oggetto realmente esistente, come ad esempio del Colosseo. L’immagine esibisce sempre *alcuni* dei tratti pertinenti di un oggetto, che possono essere *riconosciuti e sintetizzati* facendo entrare in gioco un significato. Il fatto che l’immagine rimandi a un significato e non a un oggetto, non esclude ovviamente che un’immagine possa essere usata come espressione per riferirsi a un oggetto concreto, in determinati contesti. Tuttavia, come per le parole, il riferimento ha per condizione un significato generale che deve mediare il riferimento al senso individuale. Fuori da questa sussunzione linguistica l’immagine non è nulla, nelle parole di Garroni: “riconoscere qualcosa o equivale a essere già pronti a parlarne, in forza di qualche linguaggio correlato alla percezione, oppure è un atto senza esito [...] destinato ad autoconsumarsi [...] non incidente sulla nostra presa di coscienza del mondo, anche solo percettiva” (*Ivi*, p. 59). Per concludere, sembrerebbe che sebbene si possa avere un linguaggio senza immagini, non sarebbe possibile avere immagini senza un linguaggio, a causa della loro fondamentale indeterminazione.

Garroni non si esprime sul caso inverso (dalla parola all’immagine), tuttavia gli esempi già citati di Gombrich possono essere di aiuto. In questo caso la produzione grafica segue degli stilemi codificati nella pratica rappresentativa, i quali possono presentare elementi di sorprendente dipendenza dalla cultura di chi si presta a rappresentare qualcosa. Nell’immagine prodotta dal VLM non è un caso che il gatto rappresentato fosse un Gatto Europeo. La

familiarità con una specifica determinazione del significato *gatto* (“produci una immagine di un gatto con due code che cammina sopra un ramo di una quercia in autunno”) ha probabilmente influenzato la produzione grafica del modello.

4. *L'eccedenza dell'immagine*

Nei precedenti paragrafi è stato descritto il modo in cui immagini e linguaggio possono entrare in reciproca relazione. Il rapporto è apparso sbilanciato dal lato del linguaggio, i cui significati sono la condizione del riconoscimento di immagini. Tuttavia una questione rimane in sospeso, ovvero quella dell'eccedenza dell'immagine rispetto alla sua traduzione linguistica. Questa eccedenza non si differenzia da quella che è stata chiamata *indeterminatezza dell'immagine*, ovvero dal fatto che un'immagine può costituire il senso di innumerevoli espressioni. Tuttavia la critica in questo caso sembra malposta, principalmente per due motivi.

Il primo motivo è che il linguaggio non è fondamentalmente incapace di tradurre adeguatamente un'immagine, semplicemente non è in grado di rappresentare simultaneamente tutte le innumerevoli possibili prospettive da cui un'immagine può essere *tradotta* linguisticamente. Tuttavia, nemmeno un'immagine è in grado di farlo, l'immagine si presenta sempre sotto un certo rispetto, è *letta* sempre da una qualche prospettiva. La quale può essere mutata, invertita, falsificata, ma sempre in un susseguirsi di interpretazioni che la fissano. Da questo punto di vista, l'immagine costituisce una materia indeterminata ma determinabile, disponibile all'azione formatrice della lingua. Le determinazioni della materia sensibile, in questo caso visiva, possono seguire tratti altamente codificati, come il pelo, le orecchie e le zampe del gatto. Oppure possono optare per strategie non codificate, assumendo tratti della materia come possibili “unità pertinenti di un codice in formazione” (Eco 1975, §2.14.4), come accade per esempio nel caso dell'interpretazione della musica (Capone 2023). Quale che sia la strategia di determinazione adottata, l'immagine di per sé, non sembra possedere alcuna nascosta ricchezza inaccessibile al linguaggio.

Il secondo motivo per cui questa critica non coglie il segno è che potrebbe essere mossa anche nei confronti del medium grafico. Una frase apparentemente semplice come “un gatto con due code che cammina sopra un ramo di una quercia in autunno” presenta un grado di indeterminazione non inferiore a quello dell'immagine. L'espressione potrebbe costituire il senso di un numero indefinito di imma-

gini, che la rappresentino da prospettive e secondo strategie sempre nuove, determinando graficamente il medesimo significato. Non per questo tuttavia sembra ragionevole affermare che la parola presenta una ricchezza inattingibile all'immagine e intraducibile in figure.

A questo punto, anche l'obiezione relativa a un'ipotetica irriducibilità delle immagini al linguaggio sembra risolta. Dai precedenti paragrafi emerge lo statuto segnico delle immagini. Il correlato naturale della percezione umana sembra essere una competenza semiotica sofisticata, come condizione del rapporto che gli esseri umani intrattengono con le immagini e più in generale con la propria percezione. Tuttavia, se la spiegazione di Garroni può essere soddisfacente per gli esseri umani, non sembra del tutto esaustiva per i modelli, i quali non dispongono di una percezione flessibile e di una capacità di organizzazione dell'esperienza percettiva pre-linguistica. Nel prossimo paragrafo verrà analizzato il caso dei VLM.

5. *Visual Language Models*

Garroni postula una percezione flessibile e sofisticata che lavora in stretta correlazione con il linguaggio. Per i Language Model (LM) e i VLM questa possibilità è da escludere. Se la percezione consente di far emergere delle *linee di resistenza* (Eco 2016, p.69) nella materia sensibile (conservandone comunque l'indeterminata forma aggregativa), per un modello la situazione è ancora più incerta. Per un VLM (un modello in grado di ricevere e generare sia immagini che testo) un'immagine si presenta come una matrice numerica, i cui valori determinano l'intensità dei pixel.² Questa matrice è costituita da delle serie di numeri ordinati che non offrono nessuna preferenza, nessuna linea di resistenza in vista di un'organizzazione in gruppi di tratti correlabili a significati. In questo magma isotropico deve subentrare un principio di classificazione esterno, affinché questi rapporti numerici possano essere correlati con dei termini significativi. Questo è ciò che succede nei VLM.

Un VLM è fondamentalmente una rete neurale, la quale trasforma il proprio input (testo o immagini) in una serie di vettori.³

² In questo caso viene dato per scontato che l'immagine sia in bianco e nero, rappresentata quindi da una matrice bidimensionale $n \times m$. I casi di immagini a colori chiamano in causa strutture a più dimensioni, come i *tensori*. Questo non incide sul principio di funzionamento del modello.

³ I vettori sono la struttura dati attraverso cui le reti neurali rappresentano i propri input. La forza rappresentativa dei vettori sta nel fatto che non rappresentano in isolamento i propri oggetti ma li pongono in uno spazio rappresentativo in funzione di reciproci rapporti con altri elementi (Capone 2021, Capone e Cimatti 2025).

Durante l'addestramento un VLM *apprende* a rappresentare immagini e testi fra loro correlati tramite vettori simili. Questo obiettivo viene raggiunto tramite un addestramento congiunto con immagini correlate a sequenze di parole (esistono diverse tecniche, Ramesh *et al.* 2022, Sharia *et al.* 2022, OpenAI 2023). In questo modo, il modello apprende a relazionare rappresentazioni di elementi visivi e linguistici, sviluppando rappresentazioni congiunte. Un modello di questo tipo può ricevere in input un'immagine e, dopo averla rappresentata vettorialmente, generarne la descrizione linguistica. In una fase successiva, queste rappresentazioni vettoriali vengono utilizzate come condizionamento per l'addestramento di una seconda parte del modello, dedicata alla generazione di immagini (Nichol *et al.* 2022).

Un VLM non costruisce direttamente un'immagine dal nulla, ma procede per sottrazione, eliminando progressivamente del rumore da una matrice inizialmente composta da puro rumore⁴. Il cuore di questo processo è una rete neurale detta *denoiser*, la quale viene addestrata a rimuovere il rumore da immagini solo parzialmente corrotte, ricostruendole.

Al termine dell'addestramento, il modello è in grado di generare un'immagine coerente a partire da un prompt linguistico (vettorialmente rappresentato), rimuovendo iterativamente il rumore da una matrice iniziale completamente casuale. In questo modo il modello procede per sottrazione, facendo emergere nell'immagine differenze significative in funzione di un condizionamento linguistico. Il VLM, non avendo a disposizione una sensibilità vera e propria, appalta tutto il lavoro esplorativo, profilativo, articolatorio della materia sensibile al versante linguistico, al principio di classificazione rappresentato dalla lingua. In questo modo è in grado di realizzare un'immagine partendo da una matrice di rumore indifferenziato.

6. Conclusioni

I paragrafi precedenti hanno cercato di dare una risposta a due questioni emerse dall'analisi del rapporto fra immagini e linguaggio nei sistemi di IA: la questione delle condizioni di possibilità che consentono la traduzione di immagini in linguaggio e viceversa; il problema dell'eccedenza dell'immagine rispetto alla sua resa linguistica. La risoluzione di queste due questioni consente ora di chiarifi-

⁴ In breve, il modello riceve (oltre alle istruzioni testuali) un'immagine sgranata, composta da pixel casuali, in cui non è possibile identificare alcun contenuto (in termini semiotici, una materia amorfa).

care il senso di alcune proposizioni introdotte nel primo paragrafo, le quali non consentivano di porre il problema del rapporto fra immagini e linguaggio in termini produttivi.

Proviamo a trasformare la formula “a picture is worth a thousand words” in una domanda: “Is a picture worth a thousand words?”. Un’immagine può riassumere il senso di numerosi enunciati, ma allo stesso tempo un enunciato può sintetizzare il senso complessivo di innumerevoli immagini. Il punto da ribadire è che il significato di una espressione non è meno indeterminato del significato di una immagine e, allo stesso modo, può essere determinato quanto può esserlo una immagine. La traducibilità di una immagine in una descrizione verbale, e viceversa, rimanda allora a questa caratteristica fondamentale dei segni: il loro significato è sempre determinato/indeterminato, ossia sempre indefinitamente determinabile.

Riguardo alla formula “il visibile non è riducibile al dicibile”, questa non è, per usare la distinzione di Wittgenstein in *Della Certezza* (2016, §§57, 58, 468, vedi anche 2014, §§90, 251), una proposizione empirica, è piuttosto una proposizione grammaticale. Nel nostro mondo largamente dominato dalla parola è *necessario supporre* che esista qualcosa che la parola non riesce a catturare. Rimane, tuttavia, che il visibile può essere *oggetto* di esperienza solo in quanto sussunto semioticamente, cioè in quanto è pensato come un segno e mediante operazioni segniche.

I LM (e i VLM) sono casi esemplari di questa insuperabile semiotizzazione dell’esperienza, dal momento che esibiscono in modo affatto evidente l’impossibilità di un rapporto diretto, non mediato semioticamente, con il mondo. Se già ‘semplicemente’ vedere il mondo è un’operazione semiotica (Cimatti, 2024), non stupirà che questo stesso vedere sia così facilmente riportabile in parole. Rimane tuttavia una tensione fra il nome e l’immagine, una tensione intrinseca in ogni operazione semiotica, una tensione che rimanda al fatto che ogni segno è allo stesso tempo determinato e indeterminato.

Bibliografia

- Capone L., *Wittgenstein’s remarks on the philosophy of computational linguistics. Philosophical clarification applied to natural language processing*, in “Rivista Italiana di Filosofia del Linguaggio”, SFL – Rules, praxis and forms of life (2024), pp. 172-186
- Capone L., *Segni e suoni: un’analisi linguistica della semantica dei codici musicali*, in “De Musica”, 1(XXVII)(2023), pp. 28-48.

- Capone L., *Which theory of language for deep neural networks? Speech and cognition in humans and machines*, in “Technology and Language”, vol. 2 (2021), pp. 29-60
- Capone L., Cimatti F., *Sensi senza significato. La competenza semantica al tempo dei “Large Language Model”*, in “Reti, saperi, linguaggi”, 1 (2025), pp. 175-198
- Cimatti F., *L'occhio selvaggio. Sul lasciarsi vedere*, Quodlibet, Macerata 2024.
- Cometa M., *La scrittura delle immagini. Letteratura e cultura visuale*, Raffaello Cortina Editore, Milano 2011
- De Mauro T., *Minisemantica dei linguaggi non verbali e delle lingue*, Laterza, Roma-Bari 2024.
- De Mauro, T. (1970). *Introduzione alla semantica*, Laterza, Roma-Bari.
- Eco U., *Kant e l'ornitorinco*, La nave di Teseo, Milano 2016
- Eco U., *Trattato di semiotica generale*, Bompiani, Milano 1975
- Garroni E., *Immagine Linguaggio Figura*, Laterza, Roma-Bari 2005
- Gombrich E.H., *Art and Illusion: A Study in the Psychology of Pictorial Representation* (1960); transl. *Arte e illusione: Studio sulla psicologia della rappresentazione pittorica*, Phaidon, New York 2002
- Hasenmueller C., ‘A picture is worth a thousand words’: how we talk about images, in “Semiotica”, 73/3-4 (1989), pp. 275-300
- Nichol, A.Q., et al. *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*, in “International Conference on Machine Learning” 2022, pp. 16784-16804.
- OpenAI. *GPT-4 Technical Report*, (2023), URL = <https://openai.com/research/gpt-4>.
- Platone, *Fedro*, Einaudi, Torino 2011
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. *Hierarchical text-conditional image generation with clip latents* (2022), URL = <https://doi.org/10.48550/ARXIV.2204.06125>.
- Saharia, C., et al. *Photorealistic text-to-image diffusion models with deep language understanding*. In “Advances in neural information processing systems”, 35 (2022), pp. 36479-36494.
- Saussure, F., *Cours de linguistique Générale* (1916); transl. *Corso di linguistica generale*, Laterza, Roma-Bari, 2015
- Violi, P., *Significato ed esperienza*, Bompiani, Milano 1997
- Wittgenstein, L., *On certainty* (1969); transl. *Della certezza*, Einaudi, Torino 2016.
- Wittgenstein, L., *Philosophische untersuchungen* (1953); transl. *Ricerche filosofiche*, Einaudi, Torino 2014.