Dalle prime enciclopedie online a Wikipedia, dalla nascita dell'idea di intelligenza artificiale alle reti neurali, un viaggio affascinante che porta dal tradizionale modello architettonico delle conoscenze ai nuovi oracoli statistici come ChatGPT, guidati da uno dei più noti esperti italiani di culture digitali.

GUARDA IL VIDEO

9 788858 152508

per informazioni sui nostri libri
☞ iscriviti alla newsletter su
www.laterza.it e seguici su

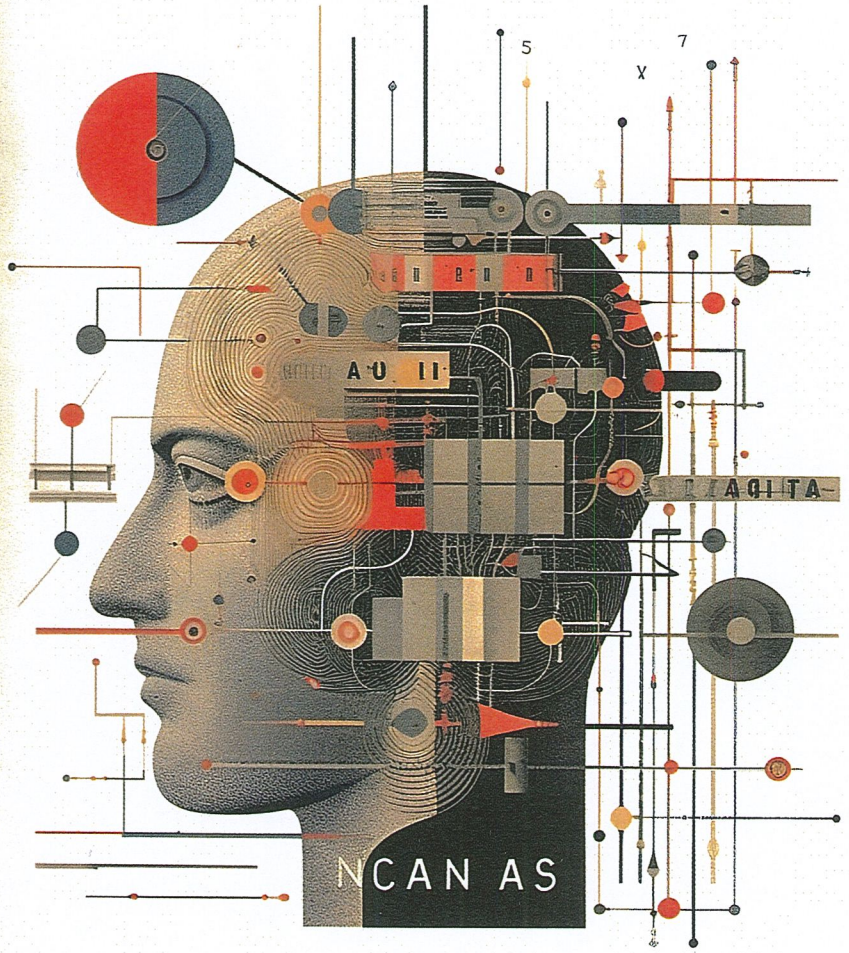
Gino Roncaglia L'architetto e l'oracolo Editori Laterza

Gino Roncaglia

L'architetto e l'oracolo

Forme digitali del sapere da Wikipedia a ChatGPT

Editori Laterza



2.

Dall'enciclopedia mondiale alle enciclopedie digitali[1]

Herbert George Wells è oggi ricordato soprattutto come romanziere, ma a lui dobbiamo anche decine di saggi, su moltissime tematiche diverse. Uno dei suoi argomenti preferiti, tanto nei romanzi quanto nei saggi, è il futuro. E se il Wells romanziere guarda al futuro intrecciando abilmente anticipazione scientifica, utopia e distopia, il Wells saggista lo considera soprattutto nelle sue dimensioni politiche, culturali e sociali. Nonostante le sue posizioni filosofiche deterministiche, una delle sue preoccupazioni principali è quella di individuare strategie per la costruzione di un futuro politicamente, socialmente, culturalmente migliore: un futuro senza guerre, guidato da un governo mondiale, capace di riconoscere il valore unificante della scienza, di garantire istruzione e formazione di qualità (con l'adozione di modelli pedagogici che oggi definiremmo costruttivisti), di promuovere il progresso del sapere.

Tra le opere dedicate a questi temi, ne troviamo una dal titolo curioso: *World Brain*[2]. Il 'cervello mondiale', per Wells, corrisponde a una sorta di raccolta organizzata e sistematizzata del nostro sapere, che – resa universalmente accessibile attraverso le nuove tecnologie (Wells pensa ai microfilm, ma è chiaro che la rivoluzione digitale ha fornito a questo sogno strumenti ben più potenti) – possa funzionare anche come fondamento per la risoluzione di conflitti e la collaborazione universale.

Wells non considera il cervello mondiale solo come un'idea astratta: propone una strada ben definita che ne dovrebbe garantire la realizzazione pratica. E questa strada è rappresentata dalla costruzione di un'enciclopedia mondiale, strumento insieme conoscitivo e politico. A suo avviso il progetto di enciclopedia, nucleo fondante del cervello mondiale, dovrebbe essere il contraltare 'conoscitivo' della Società delle Nazioni, che da sola non può funzionare proprio per l'assenza di un adeguato fondamento epistemico: la politica ha bisogno di basare le proprie scelte su conoscenze e competenze ben organizzate.

Wells si fa così interprete di quella che è stata per molta parte degli intellettuali europei, nella breve ma intensa parentesi fra le due guerre mondiali, una ricerca quasi disperata di soluzioni capaci di offrire stabilità politica e di evitare lo spettro di nuovi conflitti (una frase del libro si rivela al riguardo quasi profetica: «Viviamo in quelli che gli storici del futuro chiameranno 'gli spauriti anni Trenta'»[3]). Alla base della sua proposta è l'idea di estendere all'intero edificio del sapere le considerazioni fatte da John Maynard Keynes in *Le conseguenze economiche della pace*: la politica non funziona se non è supportata da conoscenze e competenze solide. I politici, nella sintesi offerta da Wells delle tesi di Keynes, «sono così poco avvezzi all'utilizzo del pensiero, così ignari dell'esistenza del sapere e di cosa il sapere sia, che non ne capiscono l'importanza»[4].

L'Enciclopedia Mondiale immaginata da Wells dovrebbe contrastare questa insipienza, aiutando a costruire complessità e struttura a partire dall'enorme ma frammentata massa di conoscenze disponibili al 'nuovo mondo' che esce dalla catastrofe del primo conflitto mondiale:

> La mia idea di fondo è che per comporre il mosaico eterogeneo cui accennavo sopra, per rimediare alla dispersione e all'inefficacia delle nostre forze cognitive e coinvolgerle in un comune orizzonte di senso, [...] occorre un qualcosa – un nuovo organo sociale, una nuova istituzione – che per il momento chiamerò Enciclopedia Mondiale [...], il necessario atto di congiunzione fra le nostre risor-

se mentali e le nostre forze politiche. [...] [Nei volumi dell'Enciclopedia Mondiale, chiunque potrebbe trovare] senza alcuna fatica, redatti in un linguaggio chiaro e immediato e tenuti costantemente aggiornati, i principi fondamentali del nostro ordine sociale, i concetti portanti e i dettagli più significativi di ogni campo del sapere, un quadro esatto e sufficientemente particolareggiato dell'universo, un compendio della storia mondiale, e qualora il nostro comune cittadino volesse studiare a fondo un determinato tema, una bibliografia esaustiva e affidabile lo rimanderebbe alle fonti primarie. Se in una certa disciplina dovessero sommarsi più metodi e teorie diverse, il lettore non si troverebbe a sfogliare raccolte di opinioni ammucchiate alla rinfusa, ma summe di idee e di spiegazioni attentamente scelte e ordinate. [...] Una tale Enciclopedia svolgerebbe nella nostra cultura laica il ruolo di una Bibbia antidogmatica. [...] Farebbe del sapere il collante che tiene insieme il mondo[5].

Quella proposta da Wells è una costruzione certo utopica; ma per certi versi si tratta di un'utopia molto concreta: un'enciclopedia concepita come cattedrale del sapere, continuamente aggiornata, a disposizione di ogni persona. Non sorprende il fatto che l'idea, probabilmente poco nota al grande pubblico, abbia invece avuto un'enorme fortuna fra chi ha lavorato a immaginare prima, e costruire poi, il nuovo ecosistema digitale e di rete. Arthur C. Clarke, che prima di essere uno scrittore di fantascienza è stato uno scienziato di prim'ordine, già nel 1962 proiettava l'idea di Wells su due dimensioni fra loro collegate: una «Global library» a suo avviso realizzabile intorno all'anno 2000, e il cervello mondiale vero e proprio, un'intelligenza artificiale di livello superiore rispetto a quella umana, a suo avviso realizzabile entro il 2100[6]. Dove meritano di essere sottolineati sia il collegamento stabilito da Clarke fra enciclopedia mondiale e biblioteca globale (uno dei molti esempi di collegamento fra il concetto di enciclopedia e quello di biblioteca), sia le date proposte: è presto per verificare la previsione relativa al World Brain, ma Wikipedia, come vedremo, è nata nel 2001, e anche se non esiste una singola biblioteca digitale globale, si potrebbe

ragionevolmente affermare che il decennio 1995-2005 corrisponde più o meno al periodo in cui i progetti di biblioteca digitale (il primo dei quali, il progetto Gutenberg, è stato avviato da Michael Hart nel 1971) hanno raggiunto una loro prima maturità, grazie anche alla possibilità di accedervi attraverso il web.

Oltre a Clarke, moltissimi altri autori[7] hanno sottolineato le affinità fra le idee di Wells e da un lato – sul versante enciclopedia globale – il world wide web e Wikipedia, dall'altro – sul versante cervello mondiale – il lavoro nel campo dell'intelligenza artificiale. Tornerò più avanti su uno di questi paragoni, mentre qui mi limito a segnalare che la più recente edizione del libro di Wells, non a caso pubblicata dalla MIT Press[8], ha affidato a Bruce Sterling (uno dei più noti autori di fantascienza cyberpunk) e a Joseph Reagle (che si è occupato a lungo di Wikipedia e ha curato un libro sul tema) due prefazioni dedicate proprio all'attualità delle idee di Wells nel nuovo ecosistema digitale.

Le proposte di Wells meritano però un'ulteriore considerazione: l'enciclopedia globale è concepita come un progetto non solo culturale ma anche e soprattutto politico. Per un verso, questo corrisponde all'idea – a mio avviso corretta – che la lotta alla dispersione e alla frammentazione e la spinta verso un'architettura complessa e strutturata delle conoscenze siano un'esigenza (e per certi versi una necessità) *politica*, se vogliamo sfuggire al rischio di collasso culturale e sociale che l'assenza di comprensione della complessità può comportare: basti pensare a temi come il riscaldamento globale, la fragilità degli equilibri geopolitici, il controllo degli armamenti, e come vedremo anche le nuove prospettive aperte dalla ricerca in intelligenza artificiale. Per altro verso, rovesciando la prospettiva, suggerisce che anche il lavoro politico, e in particolare il lavoro parlamentare, sia oggi in qualche misura un lavoro di tipo enciclopedico, proprio per la complessità dei temi da affrontare e per l'esigenza di solidità, coerenza e rigore che il lavoro legisla-

tivo porta con sé: un'idea esplorata in un interessantissimo volume di Giovanni Rizzoni sul rapporto fra parlamentarismo ed enciclopedia[9].

I prossimi capitoli intendono esplorare il percorso seguito dall'idea di enciclopedia e di enciclopedismo nel nuovo ecosistema digitale, proprio per comprendere se, come e in quale misura gli strumenti offerti dal digitale permettano di organizzare e strutturare un panorama conoscitivo che è oggi per molti versi ancor più ricco, ma anche più disperso e frammentato, di quello con il quale si confrontava Wells nel 1938. Nell'avviare questa indagine, ho ovviamente l'obbligo di ricordare che la storia delle enciclopedie e dell'enciclopedismo è ben più antica: non proverò qui a riassumerla o a discutere le differenze fra i diversi modelli enciclopedici (a partire dalla distinzione fra enciclopedie sistematiche e alfabetiche), compito che sarebbe superiore alle mie forze e poco funzionale rispetto agli obiettivi del libro che state leggendo, ma segnalo in nota almeno alcuni fra i moltissimi testi dedicati al tema, scelti fra quelli di riferimento e più strettamente funzionali rispetto al seguito della trattazione[10].

Un'ultima premessa da fare riguarda le scelte di periodizzazione. Nel primo capitolo, ho discusso alcuni fra i molti modi possibili per scandire cronologicamente le varie fasi della rivoluzione digitale. Un discorso analogo va fatto anche per il tema specifico delle enciclopedie e dell'enciclopedismo. Come vedremo, già prima della nascita di Wikipedia quasi tutti i principali editori operanti nel settore avevano abbandonato l'idea della centralità della forma stampata tradizionale, sostituendola o integrandola con versioni online. Da questo punto di vista, le enciclopedie si differenziano da molte altre tipologie di prodotti editoriali e in particolare dai libri di narrativa e di saggistica generalista, che nonostante lo sviluppo dei libri elettronici rimangono per lo più legati – almeno per il momento – al formato cartaceo.

Le ragioni di questo sviluppo sono molteplici: da un lato, le opere di reference, come appunto un'enciclopedia, sono

in effetti assai più vicine di un libro tradizionale al modello 'informatico' della base di dati e sono destinate alla consultazione occasionale e puntuale piuttosto che alla lettura lineare. Sono quindi molto meno dipendenti dalle caratteristiche suggerite o richieste da situazioni di lettura lineare e prolungata, rispetto alle quali la carta sembra mantenere ancora alcuni vantaggi sui dispositivi digitali di lettura. Inoltre, i contenuti online possono essere facilmente modificati e quindi aggiornati; le enciclopedie digitali permettono di aggiungere risorse multicodicali, e in particolare audio e video, alle informazioni testuali e alle immagini; permettono di utilizzare link per dare la possibilità di esplorare facilmente i riferimenti incrociati da una voce all'altra e di utilizzare strumenti interattivi (la visualizzazione delle funzioni matematiche in Wolfram-Alpha, un'enciclopedia online orientata alla scienza[11], è un buon esempio di questa caratteristica). Infine, ma non meno importante, va considerato il fattore prezzo: in generale, le enciclopedie stampate hanno un formato ingombrante e sono molto costose, e il costo marginale di una nuova copia è elevato. Al contrario, nel caso di un'enciclopedia digitale una nuova copia non ha praticamente alcun costo marginale, e lo spazio richiesto al lettore per la sua archiviazione è molto ridotto (se si utilizza un supporto fisico) o inesistente (se l'enciclopedia è online).

Il processo di sviluppo delle enciclopedie digitali, tuttavia, ha richiesto tempi piuttosto lunghi e ha attraversato fasi diverse, strettamente connesse all'evoluzione generale dell'ecosistema digitale. Propongo qui di individuare cinque fasi principali di questo processo, le prime quattro strettamente legate al modello 'architettonico' di organizzazione delle conoscenze da cui siamo partiti, e la quinta, proiettata sul futuro, legata alla sua ibridazione con il modello rappresentato dalle intelligenze artificiali generative e che abbiamo qui battezzato 'oracolare'. Per seguire meglio la divisione delle tematiche nei prossimi capitoli, può essere utile anticiparle qui sinteticamente:

Wikipedia di costruire e rappresentare complessità ne uscirebbe assai rafforzata: uno dei pochi casi in cui l'utopia collaborativa condivisa da molti fra i pionieri della rete potrebbe aver prodotto un risultato almeno in parte all'altezza delle aspettative, e dell'utopica idea di enciclopedia mondiale proposta da H.G. Wells.

7.

Il sogno del web semantico

La soluzione ad almeno alcuni dei problemi di affidabilità e validazione che abbiamo discusso nel capitolo precedente potrebbe essere rappresentata dall'ulteriore evoluzione che l'enciclopedismo online ha conosciuto negli ultimi anni, e che ci porta dal concetto tradizionale di enciclopedia (ben riconoscibile anche in Wikipedia, che come già visto vi fa esplicito riferimento nel primo 'pilastro') verso l'idea di un'enciclopedia costruita come un database fortemente strutturato e semanticamente ricco, basato su rigorose ontologie formali. Un'enciclopedia di questo tipo non è pensata in primo luogo per l'uso diretto da parte di agenti umani, ma come uno strumento di ricerca e recupero di informazioni utilizzato prevalentemente da agenti software. A tali agenti è lasciato il compito di 'mediare' tra il rigore formale dei dati strutturati e le richieste informali presentate dagli utenti o, se necessario, di agire direttamente sulla base dei dati recuperati, o di elaborarli secondo le istruzioni, o ancora di trasferirli ad altri agenti software. In questo modo, gli assistenti vocali conversazionali come Google Assistant, Alexa, Siri, Cortana, Bixby possono rispondere alle domande e alle richieste presentate dagli utenti attraverso il linguaggio naturale e la sintesi vocale, interpretando la domanda per mezzo di un programma di parsing, identificandone gli elementi di base, utilizzandoli per interrogare il database ed estrarre le informazioni pertinenti, e formulando infine una risposta[1].

Così, se qualcuno a New York chiede a Google Assistant «chi è il presidente?», l'agente software stabilirà, innanzi-

tutto, che siamo interessati a conoscere il nome del presidente in carica (l'interpretazione più semplice e diretta di «chi è?» rispetto ad altre possibili, come «che ruolo ha?»); interpreterà il termine «presidente» – in assenza di ulteriori determinazioni – come riferito al presidente più importante e più citato, cioè alla principale carica politica del paese, e utilizzerà i dati di localizzazione per georeferenziare la richiesta agli Stati Uniti. L'agente software invierà quindi una versione formalmente accurata dell'interrogazione a un database ampiamente basato su DBpedia, una collezione di dati (dataset) costituita da una versione strutturata e formale di voci tratte da Wikipedia[2], disponibile con licenza aperta. Da DBpedia sarà quindi recuperato il nome che ci interessa. Se poniamo la domanda «quanti anni ha il presidente?», Google Assistant individuerà innanzitutto il riferimento al presidente degli Stati Uniti, effettuando la relativa ricerca come sopra descritto. Successivamente, sostituirà «presidente degli Stati Uniti» con il nome di chi al momento ricopre l'ufficio (Joe Biden, all'epoca in cui sto scrivendo questo libro) e cercherà nel database la voce relativa; da questa voce – che è strutturata, e quindi organizzata in campi con valori – estrarrà la data di nascita, e calcolerà l'età al momento dell'interrogazione.

La costruzione di basi di dati altamente strutturati – ma in casi come quello rappresentato da DBpedia possiamo parlare, più che di un semplice database, di una 'knowledge base', cioè di una raccolta di dati che corrispondono a conoscenze organizzate – si basa su ontologie, cioè su sistemi di classificazione rigorosi. Per capire un po' meglio di cosa si tratti, pur nei limiti di questo lavoro, che non rappresenta una introduzione specifica al tema delle ontologie e della metadatazione[3], torniamo al nostro esempio e al presidente degli Stati Uniti. Se è vero che, nel momento in cui scrivo, il presidente degli Stati Uniti è Joe Biden, è anche vero però che i due concetti sono distinti: Joe Biden è una persona, 'presidente degli Stati Uniti' è una carica pubblica. Joe Biden era Joe Biden anche prima di essere presidente (e ha avuto

altre cariche pubbliche), e ci sono stati numerosi presidenti degli Stati Uniti prima di lui (così come, presumibilmente, ce ne saranno dopo di lui). Così, comprensibilmente, Wikipedia prevede due voci distinte: una sulla persona Joe Biden, e una sulla carica 'President of the United States'.

Queste due voci hanno caratteristiche (proprietà) diverse: la persona ha una data e un luogo di nascita, avrà prima o poi una data e un luogo di morte, ha studiato in una determinata scuola e poi in una determinata università, ha avuto una serie di altri incarichi... Se vogliamo costruire una 'scheda' composta dai campi che riassumono i dati salienti di una persona, dovremmo pensare a campi di questo tipo. D'altro canto, una 'scheda' relativa alla carica 'presidente degli Stati Uniti' avrà caratteristiche diverse: potrebbe riportare la data di istituzione della carica, il metodo di elezione, la residenza ufficiale di chi la detiene, e così via. Fra i vari campi, ci potrebbe essere quello relativo all'attuale detentore della carica, che al momento in cui scrivo ha valore 'Joe Biden', e cambierà valore alla prossima presidenza (i nomi che hanno costituito nel tempo i valori di quel campo hanno avuto finora la pervicace abitudine a designare individui di sesso maschile, ma speriamo non sia sempre così...).

Quali campi, *esattamente*, prevedere in queste schede? E quali campi prevedere, ad esempio, nelle 'schede' di un elemento chimico, di una specie animale, di un modello di automobile, di un vino, di un partito politico? E quante diverse categorie ci sono? Teniamo presente, fra l'altro, che potremmo voler differenziare le schede di individui diversi in base ad alcune caratteristiche: ad esempio, alla loro professione (nella scheda di un calciatore potremmo voler registrare il ruolo, o le squadre in cui ha giocato; in quella di una scrittrice i libri che ha scritto e i premi avuti; in quella di una astronauta le missioni spaziali alle quali ha partecipato...).

Il compito di costruire un sistema di classificazione (una *ontologia*) capace di assegnare a ogni voce di enciclopedia una specifica categoria, e a ogni categoria un modello di

'scheda' descrittiva divisa in campi standardizzati, equivale un po' al compito, apparentemente impossibile, di schedare l'universo.

Eppure, il sogno di cercare e trovare sistemi classificatori che 'mettano ordine' nella realtà fa parte della storia del pensiero umano. Ci ha provato Aristotele con le sue categorie (e l'opera aristotelica è considerata, proprio per il suo carattere organizzato e il tentativo di coprire sistematicamente almeno i principali rami dello scibile, come una sorta di proto-enciclopedia); ci ha provato – in una forma diversa – Leibniz, con la sua idea di caratteristica universale; ci ha provato chi ha costruito alberi delle scienze, chi ha costruito enciclopedie sistematiche, ci hanno provato i bibliotecari con complessi sistemi come la classificazione decimale Dewey o la classificazione decimale universale, ci hanno provato – in ambiti specifici – opere come il *Systema naturae* di Linneo. DBpedia ha insomma una lunga eredità alle spalle!

Per rendere più rigoroso questo lavoro – che, come capirete, è per sua natura spesso discutibile e soggetto a revisioni – sono stati sviluppati quelli che possiamo considerare come veri e propri *linguaggi descrittivi*. Uno dei principali è il Web Ontology Language, conosciuto con la sigla OWL. Nelle parole della guida introduttiva alla sua seconda versione[4], OWL è un linguaggio «costruito per rappresentare conoscenze ricche e complesse relative a entità, gruppi di entità e relazioni fra entità»[5], dove 'entità' è da intendersi in senso molto largo: persone, elementi chimici, marche di automobili, tipi di calzature, stili pittorici, strumenti musicali...

OWL rappresenta queste conoscenze in forma di *ontologie*, ed è dunque un linguaggio per la descrizione di ontologie; per ontologia si intende qui «un insieme di affermazioni descrittive specifiche su un qualche aspetto del mondo (al quale di solito ci si riferisce come al *dominio di interesse* o alla *tematica* dell'ontologia)»[6]. Per farci un'idea di massima (anche se non del tutto precisa) di come sia costruita un'ontologia possiamo dire che le singole entità corrispondono a

individui, i gruppi di entità corrispondono a *classi* e le relazioni fra entità corrispondono a *proprietà*. Le ontologie costruite attraverso OWL sono ontologie formali: sono pensate per essere utilizzate da agenti software più che da utenti umani, e la sintassi di OWL prevede che corrispondano a sequenze di *annotazioni*, *assiomi* e *fatti*. Assiomi e fatti esprimeranno (in forma standardizzata) le informazioni di base sulle classi, sulle proprietà e sugli individui di cui si occupa l'ontologia, mentre le annotazioni possono essere usate per aggiungere descrizioni e metadati. Così, ad esempio, potremmo esprimere attraverso assiomi relazioni del tipo «un golden retriever è un tipo di cane», mentre «Olivia è un golden retriever» è un fatto, e la fonte da cui è stata ricavata questa informazione (ad esempio il pedigree di Olivia) potrebbe essere espressa attraverso una annotazione. Questo esempio ci dice anche che le classi possono essere fra loro in relazione gerarchica (volendo spostarci dai cani agli esseri umani, la classe delle persone che giocano a calcio è una sottoclasse delle persone).

In OWL, tutte queste affermazioni vengono espresse attraverso una sintassi rigorosa (ed è possibile anzi scegliere fra sintassi diverse in base alle proprie necessità e ai software utilizzati), non sempre facile da interpretare per un essere umano, ma utilizzabile senza ambiguità da parte di un programma. Dal punto di vista formale, OWL si basa sulla cosiddetta 'logica descrittiva', che è un sottoinsieme della logica del primo ordine. La logica descrittiva è meno espressiva della logica del primo ordine completa, ma permette di norma di costruire algoritmi efficienti per decidere se una data affermazione è vera o falsa, e questo permette di usarla bene, ad esempio, in applicazioni di intelligenza artificiale.

Volendo andare più in profondità, potremmo aggiungere che OWL è solo uno dei piani che costituiscono l'edificio di rappresentazioni formali della conoscenza che stiamo cercando di costruire: un piano di livello abbastanza 'alto', dato che ha a che fare con le ontologie e dunque con i sistemi di organizzazione concettuale del sapere. Anche per que-

sto può lavorare con sintassi diverse. Livelli più bassi dello stesso edificio si occuperanno, ad esempio, di formalizzare il modo per descrivere proprietà e relazioni relative a classi e individui. Uno di questi livelli è RDF (Resource Description Framework), uno standard per la descrizione e lo scambio di dati (o meglio, di affermazioni su *risorse*) espressi in forma di *triple* soggetto-predicato-oggetto. Per RDF, una risorsa è sostanzialmente qualunque tipo di entità che possa essere identificata e descritta, anche se RDF è nato innanzitutto per lavorare sulle informazioni disponibili in rete (una pagina web, una voce di Wikipedia, un'immagine, un file...). Per questo motivo le risorse RDF sono rappresentate attraverso URI (Uniform Resource Identifier), che permettono di identificarle univocamente in rete e assomigliano un po' agli indirizzi delle pagine web (la cosa non è casuale: l'indirizzo di una pagina web è una forma di URI). Le triple soggetto-predicato-oggetto sono quindi un modo per collegare fra loro le URI in modo da descrivere proprietà o relazioni delle risorse di cui vogliamo parlare.

Può sembrare tutto abbastanza astratto e complicato, ma un esempio potrà forse aiutare: torniamo a Joe Biden e al suo ruolo di presidente degli Stati Uniti. Immaginiamo di voler identificare queste due entità attraverso le relative pagine su Wikipedia inglese. Allora, per esprimere il fatto che Biden è attualmente il presidente degli Stati Uniti, dovremmo costruire una tripla che ha come soggetto «Joe Biden», come predicato «è attualmente», e come oggetto «presidente degli Stati Uniti». Nella nostra tripla la URI che identifica il soggetto sarà https://en.wikipedia.org/wiki/Joe_Biden, mentre la URI che identifica l'oggetto sarà https://en.wikipedia.org/wiki/President_of_the_United_States. Però su Wikipedia non abbiamo una risorsa che identifichi la proprietà «è attualmente»: dovremmo fare riferimento a qualche altra risorsa web che preveda e descriva questa proprietà come elemento di un'ontologia.

Ed ecco che ci viene in aiuto DBpedia, che – appunto – ha alle spalle una ontologia assai ricca. L'ontologia di DBpedia,

di fatto, non solo possiede una proprietà adatta, ma la possiede in una forma più specifica e calzante: la proprietà «office», che corrisponde all'incarico ricoperto da una persona all'interno di un'istituzione. Ecco che abbiamo pronta la nostra tripla, tutta costruita attraverso URI di DBpedia:

– soggetto: http://dbpedia.org/page/Joe_Biden
– predicato: http://dbpedia.org/ontology/office
– oggetto: http://dbpedia.org/page/President_of_the_United_States

Triple di questi tipo possono essere anche presentate in forma di 'knowledge graph': una rappresentazione dei dati che visualizza le entità come nodi in un grafo, e le relazioni tra queste entità come archi che collegano i nodi. Ogni nodo e ogni arco può avere un'etichetta che indica il tipo di entità o di relazione rappresentata. Ciò permette di esplorare i dati anche in forma visuale, 'navigando' attraverso grafi. Per farvene un'idea, potete provare a seguire il link del QR-Code qui accanto, che vi rimanderà al video di presentazione di uno strumento di navigazione denominato LodLive[7]. Se volete provare a usarlo direttamente, potete andare all'indirizzo della pagina DBpedia su Joe Biden (riportato sopra) o a qualunque altra pagina DBpedia, e scegliere dalla linguetta 'Browse using' (in alto sulla sinistra) l'opzione LodLive Browser.

QR-Code 10
Presentazione di LodLive
http://bit.ly/44wFdsF

Lavorare sulle ontologie formali è un passo indispensabile nella transizione alla quale ho già accennato in apertura: quella dal web orientato principalmente alla consultazione da parte di agenti umani, al web 'semantico', che può essere utilizzato anche (e forse sarà utilizzato principalmente) da agenti software, e in cui le relazioni tra elementi – e quindi

anche tra le voci di un'enciclopedia – sono esplicite e a loro volta formalizzate. Questo è il cuore del progetto del cosiddetto 'web semantico', proposto da Tim Berners-Lee alla fine degli anni '90. Oggi tale progetto ha portato al lavoro sui Linked Open Data (LOD), che ne rappresenta una versione in parte semplificata ma più realisticamente gestibile. Il nome scelto sottolinea da un lato l'elemento fortemente relazionale delle ontologie, dall'altro il requisito di apertura dei dati, che ne permette il riuso e quindi – in un'altra manifestazione della tendenza alla costruzione collaborativa di informazione complessa – l'integrazione di ontologie e basi di conoscenze parziali e settoriali in costruzioni di più alto livello.

La breve (e a volte approssimativa) sintesi che ho provato a proporvi non intende in alcun modo rappresentare una 'introduzione' a semantic web, LOD, OWL, RDF... il compito richiederebbe molto più spazio, e peraltro esistono ottimi testi e corsi che lo fanno assai meglio di quanto non saprei e potrei fare io[8]. Lo scopo era piuttosto quello di fornire un'idea della direzione che ha preso negli ultimi anni l'enciclopedismo digitale, e dei meccanismi di funzionamento di una enciclopedia trasformata, sostanzialmente, in una base di dati strutturati, organizzati e facilmente utilizzabili da parte di agenti software.

Vale la pena notare, nel chiudere questa parte, che enciclopedie di questo tipo superano la tradizionale dicotomia fra enciclopedia sistematica ed enciclopedia alfabetica. Le enciclopedie digitali, con i loro meccanismi di ricerca rapida, avevano già superato la necessità di utilizzare l'ordine alfabetico come metodologia privilegiata per la ricerca di una voce, permettendo inoltre di rendere ricercabile l'intero testo delle voci e non solo i loro titoli. Le singole voci restavano però oggetti relativamente separati e granulari: anche se i link ipertestuali consentivano di costruire percorsi e rimandi fra di esse, non consentivano di esprimere adeguatamente le diverse tipologie di relazioni – a partire da quelle gerarchiche – esistenti all'interno dell'edificio delle nostre conoscenze. Una

enciclopedia costruita utilizzando i linguaggi e gli strumenti del web semantico, come DBpedia, supera questo limite, e rappresenta dunque un edificio architettonicamente assai più complesso ed espressivo.

Semantic web e linked open data hanno dunque dato all'architetto strumenti potentissimi per combattere granularità e frammentazione. Ma alcuni problemi restano aperti. Innanzitutto, quello della distinzione, alla quale abbiamo già accennato, fra conoscenze e informazioni: non tutte le informazioni sono conoscenze. Così, dal punto di vista della teoria dell'informazione, ogni dato ha un contenuto informativo, ma questo non gli attribuisce automaticamente anche un contenuto conoscitivo. Un elenco di dati numerici grezzi contiene informazione, ma se non sappiamo a cosa tali dati si riferiscano, non possiamo parlare di conoscenze. Una affermazione come «il mostro del lago di Loch Ness potrebbe divorarti» ha un contenuto informativo, ma se – come probabile – quel mostro non esiste, quella frase non ha valore conoscitivo. Il web contiene moltissime informazioni che non sono conoscenze, e spesso, come ho già accennato, la distinzione non è così chiara. Nell'ambito del web semantico si parla abitualmente di 'basi di conoscenze', ma quelle che cerchiamo di gestire sono conoscenze o informazioni? Nel caso di Wikipedia, che, come si è detto, *vuole* essere un'enciclopedia, parlare di basi di conoscenze sembra sensato. Ma davanti ad altri tipi di corpora, la distinzione può sfuggire di mano.

Per altro verso, il movimento di Wikipedia in questa direzione non è facile: trasformare le voci scritte dai volontari in risorse adeguatamente descritte attraverso gli strumenti formali di cui abbiamo parlato in questo capitolo richiede moltissimo tempo e competenze specifiche, costruire ontologie e basi di dati rigorose (e applicarle) è impresa lunga e complicata, e allargare questo lavoro – se pure lo si volesse fare – all'intero contenuto del web, o anche solo a una sua porzione significativa, è oggi semplicemente impensabile.

Per altro verso, si è affacciato sulla scena un personaggio che lavora con metodi assai diversi rispetto a quelli dell'architetto: l'oracolo. Fuor di metafora, gli sviluppi delle reti neurali e dell'intelligenza artificiale sembrano suggerire nuove strade di organizzazione e riuso delle conoscenze: per certi versi meno formalmente rigorose, ma per altri versi incredibilmente produttive.

Si tratta di una nuova direzione, e magari di una nuova fase, nella nostra capacità di produrre, gestire, selezionare, validare, riusare informazione complessa? Architetto e oracolo sono avversari, o possono collaborare? E in particolare, all'oracolo – cioè, a sistemi di intelligenza artificiale – potrebbe essere affidato il compito di generare la sterminata messe di descrizioni rigorose di cui avremmo bisogno per realizzare il sogno del web semantico? Per provare a capirlo, dobbiamo esplorare un campo del tutto nuovo.

Parte II

8.
Architetti, oracoli, pappagalli

Si è detto nell'Introduzione che al modello di organizzazione architettonica delle conoscenze si può affiancare (o, a seconda dei casi, contrapporre) un modello diverso: un modello che sostituisce al lavoro controllato e sistematico di costruzione progressiva e gerarchica di complessità, a partire da blocchi costitutivi ben noti ed esaustivamente descritti, una sorta di sviluppo organico, non perfettamente controllato o controllabile, in cui i livelli inferiori sono in parte ignoti o comunque talmente ricchi da impedirne una descrizione puntuale, mentre ai livelli superiori emergono proprietà e caratteristiche a loro volta difficili da scomporre o ridurre meccanicamente alle loro componenti.

La parte prima di questo libro aveva l'obiettivo di approfondire le caratteristiche e lo sviluppo del modello architettonico nel contesto del nuovo ecosistema digitale, con particolare riferimento all'ambito della sistematizzazione enciclopedica del sapere, che ne costituisce un po' l'esempio paradigmatico. Può essere utile a questo punto osservare che l'intelligenza artificiale 'classica', la cosiddetta Good Old-Fashioned Artificial Intelligence (GOFAI), ha tradizionalmente cercato di lavorare sulla conoscenza in modi molto strutturati, utilizzando logiche formali, rappresentazioni simboliche e modelli espliciti del mondo. In altri termini, per riprendere la metafora che dà il titolo al libro, le prime ricerche in intelligenza artificiale consideravano come parte essenziale del loro lavoro la costruzione di basi di conoscenze esplicite, e questa costruzione era affidata all'architetto e non certo

all'oracolo: l'architetto considerava essenziale procedere in modo sistematico, documentato e verificabile su conoscenze ben definite e il più possibile organizzate (in genere, organizzate gerarchicamente).

Le intelligenze artificiali generative, da diversi mesi al centro di una notevole attenzione mediatica, sembrano adottare invece il secondo modello. Sistemi di questo tipo non sono programmati in modo esplicito partendo da basi di conoscenze organizzate architettonicamente: imparano invece a generare output appropriato sulla base di enormi quantità di dati di addestramento; dati certo in qualche misura 'preparati' (è un tema di cui si parlerà in seguito) ma non strutturati in maniera così forte. Dati, inoltre, che sono certo informazioni ma che spesso – per riprendere una distinzione già discussa nella parte precedente – non sono conoscenze. Si tratta insomma di un processo molto meno controllato rispetto alla GOFAI: i modelli di apprendimento di una rete neurale profonda sviluppano internamente rappresentazioni complesse dei dati, che possono essere difficili da comprendere o da esprimere in termini simbolici o logici, e sono quindi spesso considerate 'scatole nere'. È da questo punto di vista che il funzionamento di tali sistemi può essere visto come in qualche misura 'oracolare': il processo che porta alla produzione dell'output non è quasi mai esplicitabile o descrivibile secondo i paradigmi architettonici tradizionali.

Vale la pena porsi subito, al riguardo, una domanda che dovrà poi essere ripresa in seguito: volendo proporre una metafora per il funzionamento dei sistemi generativi, è più sensato paragonarli a 'pappagalli stocastici' (come viene fatto abbastanza spesso) o a oracoli statistico-probabilistici (come faccio in questo libro)?

L'espressione 'pappagallo stocastico' è stata proposta dalla linguista americana Emily Bender in un articolo, assai citato, del 2021[1], che in un certo senso rappresenta un'appassionata e in parte accorata difesa dell'architetto dall'improvvisa e incontrollata invadenza dell'oracolo. In tale lavoro Emily

Bender e le sue collaboratrici sottolineano innanzitutto gli enormi costi finanziari e ambientali della creazione e dell'addestramento di grandi modelli linguistici, e non c'è dubbio che tali costi siano notevoli: l'addestramento di una rete neurale così ampia richiede mesi di lavoro parallelo di centinaia di processori ad alte prestazioni, che sono economicamente cari e assai esigenti in termini energetici. Questi costi sono strettamente correlati alla dimensione del modello, e i Large Language Models, come dice il nome, sono molto grandi. Inoltre, i corpora utilizzati per l'addestramento di LLM sono spesso disomogenei e viziati da pregiudizi (un tema su cui tornerò più avanti), inclusi pregiudizi di genere e sottorappresentazione di diversità e minoranze. Peraltro, proprio le loro dimensioni rendono assai difficile – e, di nuovo, assai costosa anche in termini di risorse umane, spesso costituite da 'lavoratori della conoscenza' reclutati nel sud del mondo e sottopagati – una loro adeguata revisione e annotazione. Manca, in sostanza, il lavoro dell'architetto:

> [...] i modelli linguistici addestrati su grandi dataset non curati e statici estratti dal web includono punti di vista egemonici pericolosi per le popolazioni marginalizzate. Per questo sottolineiamo la necessità di investire risorse significative nella cura e documentazione dei dati di addestramento dei modelli linguistici. [...] Quando ci affidiamo a insiemi di dati così ampi, rischiamo di incorrere in situazioni di *debito di documentazione*, e cioè di metterci in situazioni in cui i dataset sono sia non documentati sia troppo ampi per poter essere documentati a posteriori[2].

Molto meglio, allora, lavorare su insiemi più limitati e controllati di dati selezionati, descritti e annotati. In questa direzione va anche quello che è in realtà un secondo argomento, legato appunto all'idea di 'pappagallo stocastico': mentre i dati curati possono includere informazioni semantiche fornite, validate e verificate da noi, l'apparente correttezza sintattica e semantica delle risposte fornite da sistemi addestrati su

grandi basi di dati non curati nasconde in realtà un inganno: quel che possiamo ottenere in questo modo è solo

> [...] un sistema per incollare insieme a casaccio sequenze di forme linguistiche che esso ha incontrato nei suoi tanti dati di addestramento, sulla base di informazioni probabilistiche su come esse vengono combinate, ma senza alcun riferimento ai significati: un pappagallo stocastico (*stochastic parrot*)[3].

È difficile non riconoscere che le preoccupazioni – sia dal punto di vista dei costi finanziari e dell'impatto ambientale, sia da quello dei pregiudizi quasi inevitabilmente presenti in corpora di addestramento così ampi – sono fondate e vanno tenute presenti. C'è però un aspetto, fondamentale, che Emily Bender e le sue collaboratrici sembrano non tener presente: i risultati ottenuti lavorando su LLM sono talmente notevoli dal punto di vista delle competenze linguistiche e semantiche del sistema, da suggerire in maniera assai chiara che questa metodologia abbia permesso un vero e proprio, impressionante 'salto' qualitativo. Non siamo insomma davanti a un minimo progresso incrementale, ottenuto per di più a caro prezzo: il costo c'è, ma il risultato ha caratteristiche talmente sorprendenti – di fatto non previste e apparentemente non prevedibili in partenza – da rendere quasi impossibile (e per certi versi improponibile anche in termini di onestà scientifica del lavoro di ricerca) l'idea di tornare indietro.

Questo, si badi, non implica che i corpora di addestramento non debbano e non possano essere più curati, che i fattori di impatto ambientale non vadano valutati e compensati, che i pregiudizi non debbano essere tenuti presenti e che non si debba lavorare per ridurli o eliminarli (sapendo che l'idea di un dataset senza bias è solo un ideale regolativo, non troppo dissimile da quello rappresentato dal 'neutral point of view' nel caso di Wikipedia). Ma l'oracolo, inaspettatamente, si è dimostrato troppo bravo per poter essere semplicemente licenziato dall'architetto.

Quanto all'idea di pappagallo stocastico, la mia impressione è che – pur se ormai spesso utilizzata anche dagli addetti ai lavori – essa nasconda una sostanziale, grave incomprensione del funzionamento dei LLM. Che non sono pappagalli (non si limitano affatto a ripetere meccanicamente frammenti dei testi su cui sono stati addestrati, o almeno non lo fanno se per 'frammenti' intendiamo porzioni ragionevolmente significative di tali testi) e non sono neanche stocastici, dato che il loro funzionamento è guidato da una procedura di cui parlerò più in dettaglio in seguito – l'embedding – che coglie ed esprime numericamente, pur se su basi probabilistiche e in maniera per noi in parte oscura, elementi sintatticamente e semanticamente rilevanti dei nostri usi linguistici. ChatGPT e i sistemi analoghi, insomma, non funzionano affatto incollando «a casaccio» (il termine inglese usato nel passo sopra citato è «haphazardly») sequenze di forme linguistiche: lo fanno a ragion veduta, sulla base di modelli probabilistici assai complessi. Come un oracolo, questi modelli producono contenuti sulla base di una 'visione', anche se questa visione è almeno in parte privata e non conoscibile dall'esterno.

Confido che queste considerazioni possano risultare più chiare una volta completata la lettura di questa parte del libro, ma era utile anticiparle anche per meglio giustificare la metafora che si è qui scelto di adottare.

Si potrebbero discutere a lungo anche le possibili radici storiche di questo modello 'organico' o oracolare: ad esempio, potremmo considerare la trasmissione orale di conoscenze – prima dell'invenzione della scrittura – come un processo di questo tipo, in cui il corpus conoscitivo viene costruito e trasmesso, cresce e si trasforma nel tempo, senza essere mai perfettamente esplicitato e organizzato[4]. Da questo punto di vista, l'introduzione della scrittura può essere considerata come uno sviluppo di tipo 'architettonico': non a caso il dio egizio Toth, presentato – anche nel *Fedro* di Platone – come il mitologico inventore della scrittura (oltre che scriba degli dèi e protettore degli scribi), era anche la divinità di riferi-

mento per la misurazione e l'organizzazione delle conoscenze e per la geometria. E una connotazione in gran parte simile ha anche Seshat, la divinità egizia più direttamente associata all'architettura, paredra di Toth e spesso indicata come sua moglie o sua figlia: non stupirà il fatto che Toth e Seshat fossero le due divinità esplicitamente collegate agli archivi e alle biblioteche[5].

In questa sede non esplorerò comunque questa analogia, pur suggestiva, e in generale non discuterò storia (e preistoria) del modello organico di organizzazione e sviluppo delle conoscenze, per concentrarmi invece sulle intelligenze artificiali generative. Il loro sviluppo promette cambiamenti anche radicali in molti ambiti professionali, incluso il mondo della mediazione informativa e quello della produzione di conoscenze complesse. Si tratta di una previsione giustificata? Il funzionamento delle intelligenze artificiali generative può effettivamente essere considerato – e in che senso – come 'oracolare'? E in questo caso, che effetti potrà avere l'interazione fra sistemi architettonici di conoscenze e generazione statistico-probabilistica di contenuti?

Per provare a rispondere a queste domande, occorre per prima cosa capire di cosa esattamente stiamo parlando: cosa sono, e come funzionano, le intelligenze artificiali generative? In questa parte[6] cercherò quindi di presentare – in forma necessariamente sintetica – il contesto all'interno del quale si è sviluppato il lavoro su questi sistemi, i meccanismi di funzionamento e le caratteristiche di alcuni di essi (in particolare di quelli basati sulla generazione di testi attraverso transformer, come GPT e ChatGPT), i principali problemi riscontrati e una prima, assai parziale riflessione sull'impatto che essi potranno avere in futuro.

9.

Il contesto: IA e reti neurali

La riflessione sulla possibilità di costruire macchine 'intelligenti' (in un qualche senso del termine) è molto antica: potremmo ad esempio partire dagli automi e dagli esseri artificiali intelligenti presenti nelle narrazioni omeriche, o soffermarci sulla diffusa – e ovviamente falsa – leggenda tardo-medievale e rinascimentale dell'androide o della testa di androide intelligente la cui costruzione era attribuita ad Alberto Magno e la cui distruzione (volontaria o involontaria) era da alcuni attribuita al più noto fra gli allievi di Alberto, Tommaso d'Aquino[1]. Ma il lavoro sull'intelligenza artificiale collegato agli sviluppi nel campo dell'informatica e alla rivoluzione digitale è ovviamente assai più recente[2]: inizia negli anni '50 del secolo scorso ed è legato soprattutto a due nomi e due occasioni che hanno contribuito in maniera determinante a delinearne l'impostazione iniziale: quello di Alan Turing, che nell'articolo del 1950 *Computing Machinery and Intelligence*[3] ha posto le basi teoriche della riflessione sul rapporto fra intelligenza artificiale e intelligenza umana, e quello di John McCarthy, che ha organizzato il fondamentale seminario svoltosi nell'estate 1956 al Dartmouth College. È nel documento preparatorio di tale incontro[4] che compare l'espressione «artificial intelligence», ed è in questo contesto che nasce l'indirizzo della cosiddetta 'intelligenza artificiale forte': l'i-

QR-Code 11
Il 'Proposal'
per l'incontro di Dartmouth
http://bit.ly/3NI7VQz

10.
Le IA generative

Fra i temi discussi nel capitolo precedente, è stata presentata la distinzione fra reti neurali discriminative e generative. In buona sostanza, i sistemi di intelligenza artificiale generativa basati su reti neurali profonde costituiscono un sottoinsieme del deep learning, in cui l'obiettivo è produrre contenuti (che a seconda dei casi possono essere testuali, visivi, sonori, ma anche rappresentati da codice e programmi, giochi, ambienti virtuali, modelli 3D...) in genere in risposta a un 'prompt' (o richiesta) da parte dell'utente; prompt che sarà spesso testuale.

Così, ad esempio, i più noti sistemi di intelligenza artificiale generativa che producono immagini – ricordiamo, a solo titolo esemplificativo, Midjourney, Stable Diffusion, Dall-E... – funzionano sulla base di prompt testuali che dovranno fornire una sorta di 'descrizione' a parole dell'immagine che si desidera generare, e, analogamente, i più noti sistemi di intelligenza artificiale generativa che producono testi lo fanno in risposta a un prompt testuale.

Va ricordato comunque che non tutti i sistemi generativi lavorano partendo da un prompt testuale. Ad esempio, nel caso delle immagini l'obiettivo potrebbe essere quello, opposto, di generare una descrizione testuale partendo dall'analisi di un'immagine fornita come input (e un compito analogo potrebbe riguardare un video); oppure si potrebbe voler generare immagini combinando un prompt testuale e un'immagine fornita come esempio. In questa sede mi soffermerò esclusivamente sulle intelligenze artificiali generative, e in

particolare su quelle – come GPT o ChatGPT – che generano testi in risposta a prompt dell'utente.

Questi sistemi, sviluppati anche a partire dalle ricerche nel campo dell'elaborazione del linguaggio naturale (NLP, o Natural Language Processing)[1], funzionano sempre partendo da un vasto corpus di testi, utilizzato per la costruzione del modello. Il primo passo è quello di selezionare il corpus e di prepararlo per l'analisi ('preprocessing'). Lo si fa attraverso la *tokenizzazione*, fase in cui il testo viene ripulito e suddiviso in token: unità più piccole che possono essere singole parole o morfemi di più basso livello, ma anche singoli caratteri o n-grammi (gruppi di n caratteri), a seconda del modello di tokenizzazione usato. Ad esempio, la parola 'dinosauro' potrebbe essere analizzata come 'dino-sauro' (ma anche, volendo, come 'dino-saur-o', o in altri modi ancora).

Segue la fase dell'*apprendimento autonomo* ('unsupervised learning'), durante la quale la rete neurale impara, sempre sulla base del corpus di partenza e aggiustando progressivamente i valori associati ai token e i pesi dei propri collegamenti interni, a predire il token successivo sulla base di quelli precedenti. È in questa fase che si crea il Large Language Model (LLM) vero e proprio: un modello di correlazioni statistico-probabilistiche fra token, ciascuno dei quali è rappresentato attraverso un'ampia matrice di valori numerici. In tal modo a ogni token viene associato (*vettorializzazione*) uno spazio astratto e multi-dimensionale che esprime, in maniera puramente numerica, i contesti d'uso e le relazioni del token nel corpus: token con 'usi' simili, e dunque presumibilmente con significati vicini, corrisponderanno a vettori che avranno, almeno per alcune delle dimensioni, valori numerici abbastanza vicini; lo stesso avverrà, rispetto ad altre dimensioni, per token frequentemente usati insieme.

La costruzione dei vettori per ogni token – che, come si è detto, avviene nella fase di addestramento del modello – fornisce quello che è chiamato 'embedding': una rappresentazione che in sostanza cerca di coglierne, trasformandole in

valori numerici, le modalità d'uso nel linguaggio. Va notato che le 'dimensioni' del vettore – che possono essere anche migliaia – sono puramente astratte e non corrispondono necessariamente (anzi, di regola non corrispondono affatto) alle categorie grammaticali o semantiche che utilizzeremmo noi per classificare una parola o un morfema.

Il modello così costruito sarà poi utilizzato per generare, a partire dal prompt dell'utente, la risposta del sistema, con un meccanismo detto 'sequence-to-sequence': partendo da una sequenza di simboli in ingresso viene generata una sequenza di simboli in uscita. Ma come funziona questo processo?

Inizialmente, per compiti simili erano utilizzate soprattutto le cosiddette *reti neurali ricorrenti* (RNN)[2]. A differenza di una rete fatta di più strati di percettroni (Multi-Layer Perceptrons o MLP[3]), in cui l'informazione viene elaborata con un movimento sempre 'in avanti' da uno strato all'altro (per questo si parla anche di Feedforward Neural Networks), nelle RNN è possibile prevedere più cicli di rielaborazione dell'informazione da parte dello stesso strato della rete: questo permette – fra l'altro – di 'ridurre l'errore' in maniera molto più efficace. Tuttavia, nelle RNN l'analisi dei testi forniti come input (tanto a livello di corpus quanto a livello di prompt) e la produzione dell'output sono comunque fatti una parola alla volta. Reti di questo tipo hanno problemi di 'memoria semantica': soprattutto nei contesti più lunghi, la pura associazione statistica di parole fornita attraverso l'embedding non basta a conservare la coerenza semantica del testo prodotto. Hanno inoltre problemi di costi computazionali: il lavoro puramente sequenziale sfrutta male l'uso parallelo di più processori, indispensabile per lavorare su corpora assai ampi e su reti neurali molto complesse.

QR-Code 13
Come funziona l'attenzione nei transformer: video di Arkar Min Aung
http://bit.ly/3XLCOYL

Il successivo (e fondamentale) passo in avanti sulla strada verso i sistemi generativi odierni viene fatto nel 2017, con la pubblicazione da parte di un gruppo di ricercatori impegnati nei laboratori di intelligenza artificiale di Google di un articolo che è negli ultimi anni probabilmente il più citato del settore: *Attention is all you need*[4]: è questo articolo che introduce una architettura di rete molto più efficace delle RNN: quella basata su transformer[5]. In questo caso, i token non sono più esaminati solo sequenzialmente ma anche tenendo conto del loro contesto, attraverso un meccanismo di *attenzione* che 'pesa' i valori dei vettori di ogni token in funzione dei valori di ciascuno degli altri token del contesto. Nel farlo, la rete lavora, partendo dal vettore che rappresenta l'embedding del token, anche con tre vettori aggiuntivi, denominati – per analogia con le tecniche di ricerca in un database – «Query», «Key» e «Value»: non entreremo qui nel dettaglio del loro funzionamento.

Oltre a facilitare la disambiguazione di parole polisemiche (se il contesto di occorrenza di una parola come 'pesca' contiene anche termini come 'frutto' o 'succo', questo produrrà – partendo dal vettore iniziale – vettori pesati con valori più vicini a quelli di parole come 'arancia' o 'mela'; se invece il contesto contiene i termini 'pesce' o 'rete', questo produrrà dei vettori pesati con valori più vicini a quelli di parole come 'caccia' o 'sport'), questo metodo funziona molto meglio su input lunghi e produce sistemi con una 'memoria semantica' assai migliore. Inoltre, il meccanismo di attenzione può essere ripetuto più volte ('multi-head attention') per dar conto, attraverso pesi diversi, di forme diverse di attenzione, alcune delle quali riguarderanno la semantica, altre la sintassi della frase: ad esempio, dopo un articolo ci si aspetta probabilmente – ma non necessariamente – un nome, e il sistema, in buona sostanza, dedicherà 'attenzione' anche a questi aspetti. Fermo restando che anche in questo caso la distinzione fra semantica e sintassi è solo un nostro modo possibile di guardare a quelle che per il sistema sono solo relazioni numeriche

fra vettori, quasi mai direttamente interpretabili attraverso le nostre categorie linguistiche abituali.

Nel funzionamento del sistema, l'architettura basata sui transformer è applicata più volte, sequenzialmente, sia nella codifica dell'input sia nella produzione della risposta. Di questo processo possono far parte due moduli diversi, denominati rispettivamente 'encoder' e 'decoder'. Non entrerò qui nel dettaglio del loro funzionamento: per avere un'idea del loro ruolo può però essere utile ricordare che, nella maggior parte delle intelligenze artificiali generative dedicate alla traduzione, l'encoder si occupa specificamente della rappresentazione attraverso vettori del testo ricevuto come input, e il decoder della generazione dell'output a partire dalla rappresentazione prodotta dall'encoder. La famiglia di modelli composta da T5 (Text To Text Transfer Transformer) e dai suoi successori usa un'architettura di questo tipo, che unisce encoder e decoder. Ma è possibile avere anche transformer che si concentrano soprattutto sull'aspetto della rappresentazione e analisi del testo (e usano quindi solo encoder), o transformer che si occupano soprattutto della generazione di testo (e usano solo decoder). La scelta del modello più funzionale (encoder-decoder, solo encoder o solo decoder) dipenderà in parte dai nostri obiettivi.

Così, ad esempio, BERT (Bidirectional Encoder Representations from Transformers) – uno dei primi modelli basati su transformer, proposto nel 2018 da un gruppo di ingegneri di Google guidato da Jacob Devlin – aveva l'obiettivo di creare un LLM capace di 'predire' sia un token dato il suo contesto, sia la frase successiva di un dato contesto (NSP, Next Sentence Prediction). In questo caso, l'attenzione è posta soprattutto sull'analisi del testo, e il transformer era quindi costituito solo da encoder. Questo tipo di architettura funziona bene anche nei casi in cui ci interessi ad esempio la cosiddetta 'sentiment analysis' (l'identificazione delle connotazioni emozionali di un testo), o l'identificazione dei diversi 'agenti' in un dialogo, o ancora la costruzione di sommari

o parafrasi del testo o la sua analisi semantica. Da BERT è nata tutta una famiglia di modelli encoder-only, con finalità e struttura in parte diverse.

OpenAI, la società che ha prodotto la famiglia di modelli GPT, ha invece lavorato soprattutto su sistemi composti solo da decoder, e finalizzati in primo luogo alla produzione di testo. I modelli GPT lavorano in questo modo, e la loro architettura – fatta solo di decoder – ha l'obiettivo finale di predire il token successivo partendo dai token precedenti. Le parole generate man mano dai decoder diventano esse stesse parte dell'input usato per la produzione della parola successiva, il che consente di generare risposte che mantengono coerenza sintattica e semantica anche se sono molto più lunghe del prompt di partenza.

È interessante notare che anziché scegliere sempre, volta per volta, la parola che il sistema seleziona come 'più rilevante', la generazione del testo utilizza una ulteriore componente stocastica, selezionando a volte parole con punteggi leggermente più bassi. La frequenza di questi 'scarti' è chiamata *temperatura* del sistema: si è visto che un sistema con temperatura pari a circa 0,8 fornisce risposte più interessanti di un sistema con temperatura 0, che seleziona sempre la parola ottimale[6]. Nel momento in cui scrivo l'IA generativa presente in Bing, il motore di ricerca di casa Microsoft, permette così di selezionare fra tre diverse temperature delle risposte, identificate non in forma numerica ma in forma più colloquiale e comprensibile per l'utente: risposte 'creative', 'equilibrate' e 'precise', che corrispondono a temperature via via più basse.

Può essere utile a questo punto una riflessione su un tema che almeno dal punto di vista filosofico è decisamente importante: si dice (e si legge) spesso che i sistemi di intelligenza artificiale generativa di ambito linguistico, come GPT e ChatGPT, non usano semantica ma solo sintassi e statistica. È davvero così?

Va premesso – e si tratta di una premessa che si riferisce in realtà a quasi tutti gli usi del termine 'semantica' fatti in que-

sto libro – che il concetto di semantica è in parte ambiguo. Ne esistono infatti (almeno) due interpretazioni abbastanza diverse: studio delle relazioni fra segni e mondo esterno (nella tradizione che ha come punto di riferimento storico il lavoro del semiologo statunitense Charles Morris[7]), e studio del significato (nella tradizione che possiamo invece collegare al lavoro del glottologo e semiologo francese Michel Bréal[8]). Se si adotta l'idea di Morris, che porta a considerare la sintassi come teoria generale delle relazioni fra segni, l'embedding può in effetti essere ricondotto all'ambito della sintassi: non guarda al mondo esterno, ma solo alle relazioni interne alla sfera della produzione linguistica.

Ma l'uso del termine 'semantica' nel contesto che ci interessa in questa sede è di norma quello che lo lega alla dimensione dei significati: chi legge avrà già capito che era così, ad esempio, anche quando si parlava di web semantico. Da questo punto di vista, le intelligenze artificiali che producono testo lavorano o no anche con la sfera dei significati?

Chi considera solo sintattico il lavoro di questi sistemi, sembra ritenere che esso non abbia nulla a che fare con i significati. Se riflettete sulla sintesi del loro funzionamento proposta fin qui – pur se breve e lacunosa – vi accorgerete tuttavia che questa assunzione è fuorviante: la considerazione dei contesti d'uso dei token nel corpus e il meccanismo di attenzione producono infatti – almeno per chi considera la sfera della semantica come legata alla sfera dei significati – una sorta di 'semantica quantitativa': si lavora certo sempre con numeri, ma questi numeri 'incorporano' un'enorme quantità di informazioni che noi considereremmo semantiche, assieme a un'enorme quantità di informazioni che noi considereremmo sintattiche, e di informazioni che probabilmente non sapremmo bene come classificare o interpretare. Se ricordiamo il forte collegamento fra significato e uso delle parole stabilito – a partire da Wittgenstein e da Firth[9] – dalla filosofia del linguaggio novecentesca, potremmo dire che l'embedding lavora sicuramente anche su questa dimensione,

e che i transformer, lungi dall'essere solo 'macchine statistiche', sono (anche) una sorta di formalizzazione dell'idea di significato come uso.

Ma torniamo al nostro sistema di intelligenza artificiale generativa. Al termine della fase di apprendimento non supervisionato, il modello può essere ulteriormente perfezionato: sia integrandolo con corpora più ristretti e specifici (qualora si voglia lavorare su ambiti particolari), sia attraverso fasi di 'supervised learning' e di 'reinforcement learning'. Nel supervised learning al sistema vengono fornite, come modello, coppie di input-output – predisposte a monte o frutto di precedenti interazioni fra esseri umani e sistema – 'validate' da addestratori umani. Nel reinforcement learning, invece, gli addestratori valutano direttamente le risposte fornite dal sistema, 'premiandolo' (cioè, spingendolo a rafforzare i pesi dei collegamenti attivati nel produrre l'output partendo dal particolare input considerato) quando la risposta è considerata valida, e 'punendolo' (spingendolo a indebolire tali pesi) quando non lo è. Anche se OpenAI non ha mai dichiarato esplicitamente in che modo le interazioni con gli utenti finali siano utilizzate nell'ulteriore addestramento del sistema, la possibilità di dare un giudizio sulla bontà o meno delle risposte fornite (pollice in su o pollice verso) ha evidentemente anche lo scopo di utilizzare almeno in parte il proprio bacino di utenza come livello ulteriore di addestramento.

Inoltre, per addestrare il sistema possono essere usati meccanismi di apprendimento in cui l'output è analizzato da un altro sistema di intelligenza artificiale (spesso attraverso le cosiddette 'reti generative avversarie', o GAN, che cercano di discriminare fra output artificiale e output umano e che la rete originale deve cercare di ingannare), e in alcuni casi anche dallo stesso sistema ('self-supervised learning'). Infine, all'output possono essere (e di fatto vengono spesso) applicati dei filtri a valle di vario genere, anche per riconoscere e inibire risposte considerate per vari motivi come potenzialmente non accettabili, ad esempio per motivi etici o di op-

portunità: così, ChatGPT rifiuterà di produrre un racconto pornografico, anche se provate a chiederglielo, mentre potrà accettare di produrre un racconto blandamente erotico.

Tutti i meccanismi di apprendimento visti finora, compresi i filtri a valle, dovrebbero aiutare a limitare le cosiddette *allucinazioni* delle intelligenze artificiali generative di questo tipo: la produzione di testi con informazioni erronee, o che propongono tesi socialmente o eticamente inaccettabili, o che sembrano manifestare emozioni o volontà autonoma del sistema. Le allucinazioni rappresentano naturalmente un problema particolarmente grave se e quando consideriamo un sistema di questo tipo anche come una fonte di informazioni, come un sistema capace di organizzare o addirittura produrre conoscenze, o come uno strumento di mediazione informativa. Si tratta dunque di un tema che ha particolare rilevanza in questa sede, e su cui vale la pena soffermarsi; lo farò nel prossimo capitolo.

11.
Allucinazioni e pregiudizi

Ho cercato di fornire un'idea – anche se necessariamente assai sintetica – del funzionamento di sistemi come GPT e ChatGPT, perché molto spesso le discussioni sulla loro natura, sul loro futuro e sul loro impatto sociale e culturale (prevedibilmente assai notevole) sembrano prescindere completamente dal loro effettivo funzionamento: quando va bene, ci si limita a spiegazioni completamente generiche, e non di rado si ha l'impressione che chi ne analizza i possibili effetti, benefici o nefasti, non sappia però bene di cosa stia parlando, o giudichi soltanto partendo dalle risposte fornite da un particolare sistema in un particolare momento.

Abbiamo visto che, sostanzialmente, sistemi di questo tipo producono testi in forma predittiva: anche per questo, la metafora dell'oracolo sembra decisamente più adeguata rispetto a quella del pappagallo. Si tratta, si è detto, di previsioni statistico-probabilistiche basate su grandi modelli linguistici e su un lungo addestramento, in parte autonomo e in parte supervisionato o per rinforzo: non vi è dunque nessuna 'copiatura' meccanica delle informazioni incamerate attraverso il corpus di testi di partenza, e non vi è neanche un'operazione di estrazione dal corpus delle informazioni considerate più rilevanti rispetto al prompt dell'utente.

In altri termini, GPT o ChatGPT non funzionano come delle enciclopedie o come dei sofisticati motori di ricerca, ma come complessi oracoli probabilistici; proprio per questo, le loro 'allucinazioni' possono essere particolarmente insidiose. Ancora pochi mesi fa (per l'esattezza, nel gennaio 2023), se

Conclusioni

In questo libro ho preso in esame alcune fra le forme in cui, nella storia ancora assai breve dell'ecosistema digitale, si è manifestata la dialettica fra il tradizionale ideale enciclopedico-architettonico del sapere, concepito come edificio strutturato di conoscenze organizzate, e il lavoro – sia di ricerca, sia di produzione di nuovi contenuti (e in alcuni casi di nuove conoscenze) – che parte invece da enormi raccolte di informazioni eterogenee: informazioni che, proprio per le loro dimensioni e il loro carattere più granulare e frammentato, sono dominabili più facilmente attraverso modelli statistico-probabilistici che attraverso descrizioni sistematiche, e che sono diventate recentemente il terreno di addestramento dei nuovi sistemi di intelligenza artificiale generativa.

Almeno per quanto riguarda gli ultimi trent'anni, dalla prima introduzione del web a oggi, questa evoluzione sembra legata a una progressiva perdita di controllo sull'informazione, dovuta alla sua esplosione in termini di quantità e varietà e al conseguente indebolimento dell'ideale architettonico-sistematico. D'altro canto, l'ideale architettonico era nato come strumento di organizzazione delle *conoscenze*, cioè di un sottoinsieme del nostro universo informativo considerato come dotato di specifico valore conoscitivo, non delle informazioni in quanto tali.

L'esplosione informativa rende più difficile restare fedeli a questa distinzione. Così, per fare un esempio ormai classico, i cosiddetti 'cataloghi sistematici di risorse', che hanno storicamente rappresentato uno fra i primi strumenti per il reperi-

mento di informazioni in rete e che riprendevano – in forme diverse – l'idea di un albero delle scienze strutturato in maniera gerarchica (uno dei più famosi è stato, nei primi anni di vita del web, quello offerto da Yahoo!), organizzavano però risorse web eterogenee, non i contenuti di un'enciclopedia. E anche i cataloghi sistematici hanno comunque dovuto man mano cedere il passo a strumenti come Google Search, che partono invece da una semplice ricerca per stringhe su contenuti ancor più indifferenziati, e lavorano poi a valle sull'organizzazione dei risultati attraverso algoritmi di rilevanza[1].

I limiti rappresentati dalla pura gestione e ricerca 'orizzontale' dei contenuti sono tuttavia evidenti, e sono una delle ragioni che hanno portato al sogno del web semantico e al lavoro nel campo dei linked data. Il tentativo era quello di reintrodurre ordine attraverso procedure di metadatazione basate su ontologie standardizzate: le ontologie sembravano lo strumento giusto per restituire all'architetto il suo ruolo, e per tornare ad avere qualche forma di controllo organizzato su informazioni altrimenti troppo numerose, troppo diverse, troppo frammentate.

Il tentativo è guidato da ottime intenzioni, ma ha un problema: l'esplosione informativa, rappresentata anche dalla moltiplicazione dei big data, è talmente veloce da rendere quasi disperato il tentativo di inseguirla per associare a ogni frammento di informazione la sua brava etichetta. Nonostante l'elemento di semplificazione introdotto dai linked data rispetto al modello originario di semantic web, il compito di organizzare e classificare adeguatamente l'informazione online attraverso strumenti di questo tipo sembra essere superiore alle nostre forze. E resterebbe superiore alle nostre forze anche se, come auspicato, molte di queste metainformazioni potessero essere associate ai relativi contenuti informativi al momento della loro produzione, in maniera automatica e ad opera dei software di volta in volta utilizzati. Per un verso, infatti, questi software sono a loro volta troppi e troppo diversi per poter sperare di indirizzarne individualmente

lo sviluppo in modo da renderli anche buoni 'generatori di metadati'. E, per altro verso, le stesse ontologie si evolvono con il tempo e possono diventare obsolete, o non funzionali rispetto a obiettivi nuovi o diversi.

Le intelligenze artificiali generative non sono nate per risolvere questo problema – che in un certo senso potrebbero anzi contribuire a far crescere, considerato che si tratta di un nuovo strumento per la creazione di contenuti – ma propongono un paradigma completamente diverso di produzione e gestione dell'informazione. Un paradigma basato su grandi modelli gestiti in maniera completamente automatica, al cui interno i collegamenti fra dati non sono quelli, sistematici e controllati, ai quali è abituato l'architetto, ma associazioni statistico-probabilistiche espresse attraverso gigantesche matrici numeriche di valori che cambiano continuamente. Per quanto ciò possa risultare sconcertante, questi modelli si sono rivelati, negli ultimissimi anni, capaci di dominare e riutilizzare grandi quantità di informazioni poco o per nulla organizzate, e di farlo in forme sorprendentemente efficaci. L'oracolo sembra capace di vedere un'infinità di collegamenti che sfuggono all'architetto, e di usarli con un'abilità insospettabile.

Si tratta di una sconfitta definitiva dei modelli di organizzazione 'forte' delle informazioni? Non necessariamente: quel che almeno alcune fra le considerazioni fatte finora suggeriscono, infatti, è che le nuove capacità dell'oracolo potrebbero essere messe a frutto *anche* per il lavoro di descrizione e gestione delle informazioni proprio delle professionalità e delle istituzioni legate alla mediazione informativa. È difficile prevedere adesso se e quanto successo potrà avere questa ipotesi di collaborazione fra l'architetto e l'oracolo, ma sarà certo interessante vedere cosa ci riserverà al riguardo il futuro.