LIBROVIVOIA.IT

INTELLIGENZA ARTIFICIALE

CAPIRE I MODELLI PER DELEGARE IL LAVORO

La guida per chi usa l'IA e non sa come funziona

FRANCESCO MASERA

Prefazione di Paolo Savona

ANTEPRIMA · QUARTA EDIZIONE · NOVEMBRE 2026

CHE COSA C'È NEL LIBRO

Tredici capitoli stampati, la sesta parte online

Ogni capitolo del libro si apre con una pagina che dice che cosa racconta e perché vale la pena leggerlo. Le trovi qui tutte e tredici, nell'ordine del libro, seguite dalla sesta parte, «In pratica», che porta al manuale d'uso online, e dall'indice completo. Il libro è di 279 pagine, in bianco e nero, con diciassette appendici tecniche, un glossario di 143 voci e nove pagine online, aggiornate, richiamate da un QR nel punto in cui servono.

PARTE I

Capire

CAPITOLO 1 · PAGINA 25

Da dove viene l'IA che usi ogni giorno

La nascita dell'IA · Dove siamo oggi

La chat che usi tutti i giorni nasce da una storia lunga settant'anni. Siamo nel 1950, l'epoca in cui i primi calcolatori entrano nei laboratori. Alan Turing, matematico inglese, si domandò se quelle macchine fossero in grado di pensare. Cercando la risposta si rese conto che il concetto stesso di "pensare" era difficile, se non impossibile, da definire. Cambiò allora prospettiva e concluse che il modo migliore per stabilire se una macchina pensa fosse una prova: se una persona, conversando con un calcolatore, non riesce a capire che sta parlando con una macchina, allora quella macchina ha superato il test. Sei anni dopo, a Dartmouth, John McCarthy diede un nome a quella che stava diventando una disciplina: "intelligenza artificiale, la scienza e l'ingegneria di creare macchine intelligenti". Molti studiosi si dedicarono a quell'obiettivo, e in poco tempo si consolidarono due approcci diversi. Il primo diceva: scriviamo noi le regole, una per una, e la macchina le applica, come fa un programma tradizionale. Il secondo diceva: diamo alla macchina molti esempi e lasciamo che le regole se le ricavi da sola, con algoritmi ispirati ai neuroni del cervello. La prima strada diede presto risultati tangibili e sembrò la migliore, e lo rimase fino alla fine degli anni Ottanta. Poi, con il crescere della potenza di calcolo, con l'arrivo dei cosiddetti big data e

grazie ad algoritmi sempre più raffinati, la seconda strada, quella delle reti neurali artificiali, ebbe la meglio. Nel 2012 una rete neurale profonda (deep neural network) vinse una gara di riconoscimento di immagini con un distacco che nessuno si aspettava. Negli anni seguenti arrivarono i modelli di linguaggio (i large language model), e il 30 novembre 2022 ChatGPT li mise a disposizione del grande pubblico, con una velocità di diffusione da record: oltre cento milioni di persone nei primi due mesi. Il capitolo racconta questa storia e si chiude con una panoramica di quello che l'IA sa fare oggi, dalle proteine alle auto senza conducente. Una cosa da tenere a mente per tutto il libro è la differenza fra le due strade: un programma tradizionale segue istruzioni scritte da qualcuno e dà sempre la stessa risposta; una rete neurale impara dai dati e risponde con la cosa probabilmente più giusta.

Perché leggerlo: senza questa storia ogni annuncio sembra il primo, e non hai modo di pesarlo. È breve e si legge per intero in mezz'ora; per seguire il resto basta portarsi dietro la distinzione fra regole scritte a mano e regole ricavate dagli esempi, che torna in ogni capitolo.

CAPITOLO 2 · PAGINA 31

Come impara una macchina

IA simbolica: i sistemi esperti · Machine learning: l'apprendimento dai dati · Machine learning tradizionale: la cassetta degli attrezzi · Deep learning: il salto di qualità · Reti neurali artificiali: architetture di base · Reti neurali artificiali: architetture avanzate

Questo capitolo apre la macchina e racconta, un pezzo alla volta, come fa a imparare. La storia comincia dai sistemi esperti degli anni Sessanta e Settanta. Un esperto umano scriveva le regole, per esempio "se febbre e tosse, allora possibile influenza", e un programma le applicava una dopo l'altra. Il metodo funzionava finché le regole erano poche; al primo caso che nessuno aveva previsto si rompeva. Il machine learning (l'apprendimento automatico) rovesciò il metodo: invece di scrivere le regole, si raccolgono molti esempi già risolti e si lascia che il programma trovi da solo le regolarità. Un albero decisionale, per esempio, impara da centinaia di pazienti quale domanda fare per prima per capire se un caso è urgente. Queste tecniche, insieme alle regressioni e alle foreste di alberi, decidono ancora oggi se concedere un prestito o come smistare la posta. Hanno però un limite: lavorano su caratteristiche dei dati che qualcuno ha già scelto per loro. Il deep learning fa il passo in più, perché la rete trova da sola che cosa guardare, ed è proprio quello che serve con le immagini, le frasi e i suoni. Per spiegare come, il capitolo parte dal neurone artificiale, il mattone di base delle reti: una somma pesata dei dati in ingresso, seguita da una piccola curva che si chiama funzione di attivazione. Mettendo insieme qualche neurone si ottiene una rete neurale, e con una rete di questo tipo il capitolo mostra come prevedere se uno studente supererà un esame a partire dai risultati di altri studenti prima di lui. Poi quella rete viene addestrata davvero: si parte con pesi scelti a caso, si misura l'errore e si correggono tutti i pesi insieme, un poco alla volta, per migliaia di giri, finché l'errore è il più piccolo possibile. Questo meccanismo, la discesa

del gradiente con la backpropagation, è lo stesso che addestra i modelli da miliardi di parametri. Il capitolo si chiude con le quattro architetture che hanno reso possibile l'IA generativa: la diffusione, che impara a togliere il rumore da un'immagine; l'autoencoder, che comprime e ricostruisce; le GAN, due reti che si sfidano; e il Transformer, che con il meccanismo dell'attenzione legge tutte le parole di una frase insieme e decide quali contano per ciascuna.

Perché leggerlo: è la parte che rende possibile tutto il resto. Senza sapere che cosa sono i pesi e come si addestrano, non è possibile capire davvero perché le risposte dei modelli possono essere sbagliate. Serve leggerlo tutto? No. Per seguire il libro bastano le sezioni sui sistemi esperti e sul machine learning tradizionale, più la sottosezione sul Transformer, in coda alla seconda sezione sulle reti neurali; il resto delle due sezioni sulle reti neurali, con i conti della rete a tre neuroni, si può saltare al primo giro e riprendere quando vorrai vedere come si arriva a quei numeri. Chi salta deve però almeno portarsi dietro tre parole e il loro significato: pesi, addestramento, attenzione.

CAPITOLO 3 · PAGINA 69

I large language model

Come si addestra un modello di linguaggio · Grande non è sempre meglio: i modelli efficienti · Nuove tecniche per migliorare le capacità degli LLM · Chi fa gli LLM, e come li rilascia · Le architetture, in ordine di tempo

I modelli di linguaggio, i large language model o LLM, sono l'intelligenza artificiale con cui hai a che fare tutti i giorni, e questo capitolo racconta come nascono e chi li costruisce. Un LLM è una rete neurale addestrata a fare una cosa sola: leggere un pezzo di testo e indovinare la parola che viene dopo. Sembra un compito modesto, ma per farlo bene la rete deve assorbire la grammatica, i fatti e i ragionamenti che stanno dietro le parole, ed è così che impara a rispondere, riassumere e tradurre. Tre grandezze decidono quanto un modello è bravo a prevedere la parola successiva: quanti parametri ha, cioè quanti pesi aggiusta mentre impara; quanto testo ha visto; quanto calcolo è servito per addestrarlo. Nel 2020 ci si accorse che l'errore cala in modo regolare quando le tre grandezze crescono. Sono le cosiddette leggi di scala, e da quella scoperta è partita la corsa a costruire modelli sempre più grandi. Il costo di addestramento, però, sale in fretta, e il capitolo racconta le strade che si stanno percorrendo per contenerlo: modelli più piccoli ma addestrati molto più a lungo, oppure pesi scritti con pochi bit, cioè numeri con meno decimali. Presenta poi due tecniche che negli ultimi anni hanno migliorato i risultati a parità di dimensioni: far scrivere al modello un ragionamento prima della risposta, che è l'idea dei modelli di ragionamento, e organizzarlo dentro come una squadra di specialisti che si attivano solo quando servono, la Mixture of Experts. Il capitolo si chiude con chi costruisce questi modelli e come li rilascia: a pesi chiusi, che restano sui server di chi li ha prodotti, oppure a pesi aperti, scaricabili e modificabili; e con quello che ciascun laboratorio ha scelto di vendere.

Perché leggerlo: qui impari a capire le caratteristiche di un modello nuovo da solo, e ne esce uno al mese. La prima sezione, l'addestramento e le leggi di scala, aiuta a capire come mai leggiamo sui giornali tutti i giorni della corsa alla costruzione di data center sempre più grandi. La sezione sui laboratori si può attraversare in fretta e riprendere quando ti serve un nome.

CAPITOLO 4 · PAGINA 83

Che cosa sta intorno al modello

Adattare un modello al proprio materiale · La finestra di contesto · Capienza e attenzione · L'harness: l'impianto attorno al modello

Lo stesso modello, dentro due prodotti diversi, può dare risultati lontanissimi. La differenza la fa quello che gli sta intorno, e questo capitolo lo descrive dall'interno verso l'esterno. Un modello appena addestrato ha una conoscenza generale molto ampia, che viene dal materiale su cui è stato addestrato, ma non sa niente di te e del tuo lavoro. Per portarlo sul tuo materiale ci sono due strade. La prima è riaddestrarlo un poco, il fine-tuning: gli si mostrano esempi curati tratti dai tuoi documenti e il modello cambia il modo in cui si comporta, il tono, il formato, il vocabolario di un mestiere. La seconda è collegarlo ai tuoi documenti, la RAG: il modello resta com'è, ma quando arriva una domanda va a cercare nei tuoi file i pezzi che servono e li legge prima di rispondere. La regola per scegliere è semplice: si addestra per insegnare una forma, si dà un documento per aggiungere un fatto. Tutto quello che il modello legge prima di rispondere sta in uno spazio che si chiama finestra di contesto: le istruzioni, la conversazione fino a quel punto, gli allegati, i risultati delle ricerche. La finestra viene riletta per intero a ogni scambio ed è, in pratica, quanto il modello riesce a tenere a mente mentre lo usi. In tre anni la capienza è cresciuta enormemente: i primi modelli tenevano a mente un documento di poche pagine, quelli di oggi lavorano su testi di centinaia di pagine. Restano però dei limiti nella capacità di tenere l'attenzione su tutto quel materiale, e conviene conoscerli. Il capitolo si chiude con l'impianto completo, l'harness: il contesto, gli strumenti che il modello può usare, la memoria, il ciclo con cui rilegge quello che ha fatto e le istruzioni permanenti. Nessuna di queste cinque cose è il modello, e alcune non si comprano: le costruisci tu. È per questo che due persone con lo stesso abbonamento possono avere esperienze molto diverse.

Perché leggerlo: è il capitolo che presenta tutte le componenti della IA che usiamo tutti i giorni. Ed è proprio utilizzando al meglio queste componenti che è possibile trarre i migliori risultati dalla IA. I capitoli successivi entrano nel dettaglio: la finestra di contesto torna nel capitolo sull'uso della chat, l'harness in quello sugli agenti, la RAG in quello sulla scelta di una soluzione. Va letto per intero; è breve, e la sezione sulla finestra di contesto è la più utile del libro per chi usa una chat tutti i giorni.

PARTE II

Come usare gli LLM

CAPITOLO 5 · PAGINA 95

Ottenere quello che ti serve

Prompt engineering: come parlare ai modelli · Quando la chat si satura · La memoria degli LLM · Come far sì che l'IA si ricordi di te

Il modello non impara niente mentre gli parli: i suoi pesi sono stati fissati mesi prima e restano fermi per tutta la conversazione. Quello che separa una risposta generica da una calibrata su di te sta in ciò che gli metti davanti e nel modo in cui glielo chiedi. Il capitolo racconta le tre cose che dipendono da te, e alla terza dedica due sezioni. La prima è la richiesta. Dare al modello un ruolo, spiegargli il contesto, dirgli in che formato vuoi la risposta, che cosa deve fare e che cosa no, cambia il risultato più di un cambio di modello; e due o tre esempi di quello che intendi valgono più di una lunga spiegazione. La seconda è accorgersi quando una conversazione è diventata troppo lunga e il modello comincia a perdere lucidità: l'assistente chiede cose che gli avevi già detto, o contraddice una decisione presa poco prima. Succede perché i messaggi più vecchi sono usciti dalla finestra di contesto, e accanirsi rispiegando ogni volta l'errore non serve. Il rimedio è chiedergli un passaggio di consegne e ripartire in una chat nuova. La terza è la memoria. Un chatbot ricorda per cinque vie diverse: quello che ha imparato in addestramento, che è come il suo DNA; la conversazione che ha davanti; un profilo su di te che il prodotto gli passa all'inizio di ogni chat; i materiali di un progetto; e gli strumenti con cui va a cercare le informazioni che gli servono. Conoscere le caratteristiche e i limiti di questi cinque tipi di memoria cambia molto l'esperienza d'uso.

Perché leggerlo: sono le competenze che rendono da subito, su una chat qualunque, senza comprare niente, e sono il presupposto dei due capitoli che seguono. Leggilo per intero e con la chat aperta: ogni tecnica si prova in trenta secondi su una conversazione tua.

CAPITOLO 6 · PAGINA 103

Dall'assistente all'agente

Che cosa si intende per "Agentic AI" · I quattro pilastri dell'Agentic AI · Sfide dell'Agentic AI: autonomia, sicurezza e responsabilità · Controllare gli agenti: architetture, limiti e supervisione · Dagli agenti episodici agli agenti long-running · L'attenzione come risorsa scarsa: i sistemi multi-agente · Chi costruisce gli agenti, e con quale idea

Un agente fa la cosa di cui un modello ti parla soltanto. Se chiedi a un modello un'email di ringraziamento, ottieni il testo; se la chiedi a un agente, ti apre la posta, compila il destinatario e la manda. Il capitolo spiega che cosa permette al modello di avere questa autonomia e come si tiene sotto controllo. Il meccanismo è più semplice di quanto il marketing lasci intendere: un agente è un modello che usa strumenti dentro un ciclo. Riceve un obiettivo, compie un'azione, legge il risultato, decide la mossa successiva, e si ferma quando l'obiettivo è raggiunto o quando finiscono i tentativi. Quattro cose distinguono un vero agente da una chat con qualche funzione in più: una memoria che dura oltre la singola conversazione, la capacità di pianificare una sequenza di passi, l'uso di strumenti esterni e un'autonomia controllata, cioè confini entro cui può agire da solo e punti in cui deve chiedere. C'è però un'avvertenza. Gli agenti rendono dove qualcuno può giudicare ogni passo, come i test nel codice; altrove gli errori si accumulano in silenzio, e un passo sbagliato diventa la premessa di quello dopo. Per questo il controllo si distribuisce su quattro livelli, dai dati a cui l'agente può accedere fino alla persona che approva. Il capitolo si chiude con due frontiere: gli agenti che lavorano per giorni, che hanno bisogno di una memoria del lavoro in corso che i prodotti di oggi ancora non danno, e i sistemi in cui più agenti si dividono il compito, perché l'attenzione di uno solo non basta.

Perché leggerlo: la delega a un agente è una delle cose che il libro promette di farti saper fare, e questo capitolo dice che cosa stai delegando quando la fai. Le sezioni sul ciclo, sui quattro pilastri, sulle sfide e sul controllo servono per seguire il capitolo sul ridisegno del lavoro. Le due sull'orizzonte, gli agenti che lavorano per giorni e i sistemi di più agenti, si possono rimandare senza perdere il filo; l'ultima, su chi li costruisce, si legge quando ti serve un nome.

CAPITOLO 7 · PAGINA 117

Quando il modello sbaglia

Quando l'IA inventa: il problema delle allucinazioni · Quando l'IA ti dà ragione: la sicofanzia · Quando l'IA fa i conti: un problema di significato

Una risposta sbagliata arriva con lo stesso tono di una giusta, nella stessa forma ordinata, nello stesso mezzo secondo. Non c'è un'esitazione che la distingua: te ne accorgi dopo, quando qualcuno controlla un numero o apre una fonte che non esiste. Il capitolo racconta tre modi in cui un modello può sbagliare. Si somigliano nell'aspetto, ma richiedono controlli diversi. Il primo è l'invenzione: il modello afferma qualcosa che non esiste, una sentenza, un articolo di legge, una citazione. Succede perché il modello non consulta un archivio: prevede la continuazione più probabile del testo, e quando su un argomento ha visto poco produce qualcosa che ha l'aspetto di un fatto. Il rischio sale in quattro situazioni riconoscibili: i riferimenti, i fatti rari, i numeri che nessuno ha mai pubblicato e il dettaglio falso dentro una risposta per il resto giusta. Il secondo è la compiacenza: il modello ti dà ragione anche quando hai torto, perché ha imparato che la risposta compiacente è quella che preferisci.

Nasce dall'addestramento, dove le persone premiano le risposte in linea con le loro aspettative, e continua nell'uso; per questo "sei sicuro?" è il controllo meno affidabile di tutti. Il terzo è l'errore di calcolo: i dati sono veri e l'operazione è sbagliata, per esempio somma le percentuali o fa la media delle medie, e un numero sbagliato ha la stessa faccia di uno giusto. Per ciascuno dei tre il capitolo dà un controllo che costa poco: verificare che la fonte esista prima di leggere che cosa dice, e chiedere se un numero è un dato o una stima; far difendere al modello la sua risposta prima di cambiarla; far scrivere il calcolo in una forma che si possa rieseguire. La regola generale è una sola: il tono con cui il modello risponde non dice niente sulla sua accuratezza.

Perché leggerlo: è il capitolo della seconda promessa, e il più utile da solo: vale per la chat, per l'agente e per qualunque sistema della parte aziendale. Va letto tutto, e le tre sezioni in ordine, perché il controllo istintivo contro il primo errore innesca il secondo. Se di questo libro leggi tre capitoli, questo è uno dei tre.

PARTE III

Verso l'uso aziendale

CAPITOLO 8 · PAGINA 127

Scegliere e governare una soluzione

La prima domanda: cosa deve fare davvero il sistema? · La seconda domanda: quali dati servono e come gestirli? · La terza domanda: dove far girare il modello? · Che cosa c'è sul mercato · Governare e integrare il sistema

Le competenze della parte precedente bastano a chi lavora da solo. Quando la scelta la fa un'organizzazione per tutti, le domande cambiano: con quali dati, con quale budget, con quali responsabilità. Il capitolo le ordina in tre, da prendere in sequenza, perché ognuna restringe quella dopo. La prima domanda è che cosa deve fare davvero il sistema. Gli usi sono di quattro tipi, produrre contenuti, rispondere su documenti che esistono già, automatizzare un processo, tenere una relazione nel tempo, e uno strumento buono per il primo può essere inadatto al terzo. La seconda domanda è quali dati servono e dove stanno. Un modello non conosce i dati della tua organizzazione, e usarli non vuol dire caricarli dentro il modello: si collegano, come hai visto con la RAG, e la scelta di dove i dati passano incide sull'architettura, sui costi e sui rischi. La terza domanda è dove far girare il modello. I livelli sono quattro: la chat personale, dove i tuoi dati vanno sui server del fornitore; la stessa chat comprata dall'organizzazione con un contratto; un'applicazione costruita sulle API; il modello installato in casa, quando i dati non possono uscire. A ogni gradino cresce quello che decidi e quello di cui rispondi. Il conto ha cinque voci, e il listino ne dice una sola. Il capitolo si chiude con quello che nessuna architettura risolve da sola: il mercato italiano, con molte licenze e poca governance; la responsabilità, che non si delega al modello; e il modo di misurare se un sistema funziona, con un banco di prova fatto di casi veri del proprio lavoro e un'autonomia che si allarga un gradino alla volta.

Perché leggerlo: dipende da dove lavori. Se lavori da solo ti servono la seconda e la terza domanda, i dati e i quattro livelli, perché decidono dove finisce quello che incolli in una chat, e il resto lo puoi attraversare. Se il tuo compito è scegliere per un'organizzazione, leggilo per intero: la sezione su come si misura se funziona separa i progetti che reggono da quelli che si spengono dopo il pilota. Il capitolo che segue lo presuppone in un punto solo: il perimetro dei dati decide lo strumento.

CAPITOLO 9 · PAGINA 139

Ridisegnare il proprio lavoro

Usare e riprogettare · Smontare il processo in attività · Quali attività reggono la delega · Che cosa misuri · Chi verifica, e che cosa · I dati che non escono · Quando l'errore arriva · Il processo non è solo tuo · Delegare a un agente · Quattro processi smontati · Quello che resta tuo

Un lavoro non si delega. Si delegano le attività che lo compongono, una alla volta, e ognuna ha un prezzo da pagare in controlli. Il capitolo serve a fare quel conto su un processo tuo. Il metodo comincia dallo smontare il processo in attività, ciascuna con un ingresso, un'uscita e un momento in cui è finita. Per ognuna si compilano sei righe: che cosa entra, che cosa esce, quanto dura, come si capisce che è venuta bene, che cosa succede se viene male e nessuno se ne accorge, chi risponde del risultato. Poi ogni attività si classifica con due domande, e la difficoltà non c'entra: quanto costa verificare il risultato, e che fine fa un errore che ha superato i controlli. Ne escono quattro casi, dalla delega piena all'attività che resta tua. Il capitolo dice poi che cosa misurare, a partire da quanto ci mettevi prima, che quasi nessuno sa; come si progetta la verifica, che è un'attività con un nome, una durata e una persona, e perché la buona volontà si consuma; quali dati non escono dal tuo perimetro; che cosa fare quando l'errore arriva, perché arriverà; con chi va concordato il cambiamento; e che cosa cambia quando deleghi a un agente una catena di passi invece di uno solo. Seguono quattro processi smontati per esteso: una nota su un fascicolo di duecento pagine, un'analisi di mercato con fonti esterne, l'adeguamento delle procedure interne a una norma nuova e il lavoro con cui è fatto questo libro. In tutti e quattro la delega si concentra sulle stesse attività, ricostruire, cercare, fare una prima versione, e non tocca mai capire che cosa vuole chi ha chiesto, decidere e rispondere di quello che esce.

Perché leggerlo: ti dà il metodo per usare la IA per efficientare delle attività, e ti spiega come e cosa delegare a un agente. Va letto tutto e in ordine; dei quattro casi puoi leggerne uno solo, quello che assomiglia di più al tuo lavoro, e tornare sugli altri dopo. Se hai poco tempo per tutto il libro, questo è il capitolo da leggere per intero.

PARTE IV

Governare

CAPITOLO 10 · PAGINA 163

I rischi che ti riguardano

Quando decide un algoritmo: discriminazioni e bias · La responsabilità: chi risponde davanti alla legge · L'illusione del controllo umano · Rendere l'IA comprensibile: come spiegare una decisione · Privacy e sorveglianza

Ci sono algoritmi che ti consigliano cosa guardare la sera e algoritmi che decidono chi assumere e chi merita un prestito. La differenza sta nel costo dell'errore. E ogni algoritmo porta dentro decisioni che qualcuno ha preso: quali dati usare, che cosa contare come esito giusto, dove mettere la soglia. Il capitolo racconta i principali rischi che ricadono su chi utilizza questi sistemi. Il primo è la discriminazione. Un algoritmo impara dai dati, e se nei dati ci sono le disuguaglianze del mondo le riproduce, come il sistema di selezione del personale che penalizzava i curriculum con la parola "donna". Il secondo è la responsabilità. Quando decide una macchina, spesso nessuno risponde, e il primo verdetto che dice il contrario è quello contro Tesla in Florida. Il terzo è il controllo umano. Sembra la soluzione, ma è più difficile di quanto si creda: la macchina è più veloce di chi la sorveglia, chi ha controllato novantacinque risposte giuste non legge più la novantaseiesima, e quando qualcosa va storto la colpa cade sull'operatore. Il quarto è la trasparenza, cioè come si spiega una decisione presa da una rete, e il capitolo presenta tre famiglie di metodi. Il quinto è la privacy, dal riconoscimento facciale al filtro bolla che ti mostra solo quello che conferma le tue idee. Alla fine del capitolo avrai le domande da fare a ogni sistema che decide qualcosa su di te: come arriva alla decisione, se tratta te come chiunque altro, che cosa sa di te, chi risponde se sbaglia.

Perché leggerlo: i cinque rischi sono le ragioni per cui esistono le regole del capitolo dopo, e le domande con cui chiude sono quelle che ti serviranno quando qualcuno ti proporrà un sistema che decide su persone. Si legge per intero in un'ora; le sezioni sul controllo umano e sulla trasparenza sono quelle che il capitolo sul ridisegno del lavoro ha già usato senza dirlo.

CAPITOLO 11 · PAGINA 171

Le regole

Unione Europea · Stati Uniti · Regno Unito · Cina · Italia · Verso un approccio globale · L'obbligo che ti riguarda, e come si adempie

Il primo obbligo di legge che ti riguarda come persona che usa l'IA al lavoro è già in vigore, e quasi nessuno sa di averlo. Il capitolo racconta quali regole esistono, dove valgono e come si capisce se una riguarda te. L'Unione Europea è stata la prima a scrivere una legge completa, l'AI Act, in vigore dal 2024. L'idea di fondo è che non tutti i sistemi sono uguali: un filtro antispam e un algoritmo che decide chi assumere non possono avere le stesse regole. Per questo la legge li divide in quattro livelli di rischio, minimo, limitato, alto e inaccettabile, con obblighi che crescono con il rischio fino al divieto. L'applicazione segue un calendario scaglionato, riscritto nel 2026, e le date stanno online. Un articolo riguarda direttamente chi legge: chi usa l'IA in un'attività professionale deve occuparsi dell'alfabetizzazione di chi lavora per lui. Fuori dall'Europa il quadro è diverso. Gli Stati Uniti, dove hanno sede quasi tutti i laboratori, non hanno una legge federale: governano con ordini esecutivi, che cambiano con l'amministrazione, e con strumenti più incisivi di una legge, le commesse pubbliche e i controlli sull'export. Il Regno Unito non ha una legge e fa testare i modelli da un istituto pubblico prima del rilascio. La Cina controlla i contenuti in casa e distribuisce i modelli fuori. L'Italia ha affiancato alla legge europea una legge nazionale, la 132 del 2025. Un trattato internazionale esiste, ma i suoi meccanismi sono lenti. Per capire quale norma ti tocca il criterio è uno solo: guarda che cosa fa il sistema, non quale modello usa. L'ultima sezione spiega come si adempie l'obbligo che ti riguarda, e perché non basta un corso.

Perché leggerlo: per capire quali sono le regole da rispettare nell'usare l'IA in ambito aziendale. La sezione sull'Unione Europea e l'ultima, sull'obbligo che ti riguarda, sono quelle da leggere; le sezioni sugli altri paesi sono contesto, per sapere che cosa succede fuori, e si possono attraversare. Le date sono sulla pagina online, non qui.

PARTE V

Guardare avanti

CAPITOLO 12 · PAGINA 181

Chi possiede l'intelligenza artificiale

Il capitale: algoritmi, brevetti e chi addestra · I dati che nessuno può copiare · Le macchine che il denaro non compra · Chi decide: la somma, lo scaffale, gli Stati · Il conto: energia e acqua

Un sistema di intelligenza artificiale poggia su tre cose: gli algoritmi, i dati e le macchine. Il capitolo le guarda come si guarda una carta geografica, perché ognuna sta in mani precise e in luoghi precisi, e chi la controlla può decidere di non dartela. Gli algoritmi sono la risorsa più mobile: un'idea passa i confini dentro un articolo scientifico. La Cina registra tre quarti dei brevetti, eppure i modelli davvero nuovi escono soprattutto dagli Stati Uniti, e la ragione non sta nell'ingegno. Addestrare un modello di frontiera costa centinaia di milioni di dollari, e quel capitale si raccoglie in pochi posti al mondo: il primo confine è finanziario. I dati disegnano una mappa diversa. Il testo di qualità sul web è una quantità finita, un bene comune che si sta esaurendo, e quando finisce il vantaggio passa a chi possiede quello che nessun altro può copiare, come gli archivi di un'azienda o le conversazioni di un prodotto usato da milioni di persone. Il dato nasce dove vivono gli utenti, ma il valore si accumula dove ha sede la società che li raccoglie. Con i chip il globale si incaglia nella geografia fisica. Le GPU le vende quasi solo NVIDIA e si fabbricano in pochissimi stabilimenti, primo fra tutti a Taiwan, a cento chilometri dalla costa cinese; sotto ci sono le materie prime, concentrate anch'esse in pochi paesi, e sopra i decreti di due governi. Qui il controllo si esercita in modo secco: un divieto di esportazione fa più di un dazio, perché chiude il mercato a qualunque prezzo si sia disposti a pagare. Le tre mappe non coincidono, e chi comanda su una non comanda sulle altre: è questo che tiene aperta la partita, e la rende dura. Sopra le tre mappe ce n'è una quarta, lo scaffale da cui i modelli si scaricano, e ci sono gli Stati, che hanno cominciato a decidere chi può comprare che cosa. Tutte e tre le risorse, infine, consumano: il conto in energia e acqua, che chiude il capitolo, è il prezzo dei pezzi, e sta diventando il collo di bottiglia della corsa.

Perché leggerlo: senza sapere chi ha i pezzi non si capisce da che cosa dipende la corsa, e le notizie sui chip, su Taiwan e sulle terre rare si leggono come cronaca invece che come un capitolo di una guerra per il dominio della IA. Si legge tutto in un'ora, e nessun capitolo lo presuppone.

CAPITOLO 13 · PAGINA 191

Dove sta andando

Le biotecnologie: la prova che fa lavoro vero · La robotica: l'IA fuori dallo schermo · Le armi e i conflitti: l'autonomia dove costa di più · Verso l'intelligenza artificiale generale · Quello che nessuno ha ancora deciso

Fin qui hai visto quasi soltanto modelli di linguaggio. Fuori dal testo l'intelligenza artificiale fa altro, e il capitolo guarda quattro campi in cui sta cambiando le cose più in profondità, messi in fila da quello in cui il risultato è acquisito a quello in cui è ancora una parola. Il primo campo è la biologia. Prevedere la forma di una proteina dalla sua sequenza era un problema aperto da cinquant'anni: AlphaFold lo ha risolto, nel 2024 ha portato il Nobel per la Chimica e nel 2026 un farmaco progettato con questi metodi è entrato nell'ultima fase di sperimentazione. Il secondo è il corpo: l'IA che esce dallo schermo ed entra in macchine che percepiscono, si muovono e lavorano accanto alle persone, dove i dati non sono gratis e le promesse corrono più avanti dei prodotti. Il terzo è la guerra: droni che colpiscono da soli quando il collegamento salta, sistemi che producono liste di bersagli più in fretta di quanto una catena di comando riesca a verificarle, modelli che entrano in una rete e la attraversano da soli. È la frontiera in cui l'autonomia costa di più, e in cui le regole sono più indietro. Il quarto è la parola: l'intelligenza artificiale generale, che i laboratori annunciano, i padri della disciplina discutono e nessuno sa misurare; la strada che ci porta, intanto, è quella degli agenti che hai visto nella Parte II. Per tutti e quattro il metro è lo stesso, ed è quello del capitolo 1: dimostrazione o servizio.

Perché leggerlo: per leggere gli annunci su queste quattro frontiere senza spaventarti né entusiasmarti a comando, e per riconoscere, in fondo, le scelte che nessuno ha detto ad alta voce e che ti arrivano sul tavolo travestite da scelte tecniche. Nessun capitolo lo presuppone: si può leggere per ultimo, o a pezzi.

PARTE VI · ONLINE

In pratica

Da quello che hai capito a quello che farai, con lo strumento che hai davanti.

Le cinque parti precedenti spiegano come funziona un modello, che cosa gli si può affidare e come si controlla il risultato. Questa parte ti porta al primo gesto, con la chat aperta sul tuo schermo e un lavoro che oggi devi fare comunque, e alle venti tappe che vengono dopo. Stanno in una guida che fa parte del libro e vive su librovivoia.it.

Sta online per una ragione che a questo punto conosci: i nomi dei menu, le funzioni comprese nei piani gratuiti e i prezzi cambiano nel giro di settimane, mentre il libro si riscrive una volta l'anno. Sulla carta resta il perché; il come, con le schermate e la data in cui l'ho verificato, sta dove si corregge il giorno in cui lo strumento cambia.

Alla fine di questa parte sai da dove comincia il percorso, che cosa ti dà ciascuno dei quattro livelli e a quale punto puoi fermarti avendo già quello che ti serve.

PAGINA 207

Dalla chat alla delega

Agli incontri sul libro la domanda che torna più spesso è sempre la stessa: ho capito come funziona, da dove comincio domani mattina? Chi la fa ha letto i capitoli, sa che cosa è una finestra di contesto e perché un modello sbaglia, e davanti alla casella vuota di una chat non sa quale sia il primo lavoro da affidarle. La guida "Dalla chat alla delega" nasce da quella domanda e da un questionario a cui hanno risposto i lettori: che strumento usano, con quale piano, dove si bloccano, se al lavoro possono usarlo sui documenti dell'ufficio. Le risposte hanno deciso quali tappe ci sono e quanto spazio ha ciascuna.

Le tappe sono venti, in quattro livelli. Il primo è la chat: la prima richiesta vera, la stessa richiesta fatta bene, un documento tuo, il controllo di una risposta, il lavoro finito e consegnato come file. Il secondo porta lo stesso lavoro dentro i programmi in cui passi la giornata, Word e Outlook, Excel, PowerPoint. Il terzo è il contesto che resta: quando una chat si satura, che cosa il modello ricorda e da dove, le istruzioni permanenti, il primo progetto, il criterio per scegliere fra una chat e un progetto, un assistente costruito su misura per un compito che ripeti. Il quarto sono gli strumenti e la delega: la ricerca su fonti che scegli tu, i tuoi dati collegati con i permessi giusti, le riunioni come fonte, il modello che lavora sui file e nel browser del tuo computer, il primo compito consegnato per intero e il compito che torna da solo a scadenza fissa. Dopo il quarto livello vengono quattro tappe per mestiere, provate su casi veri: l'analisi di un bilancio, la scheda di un paese, la grafica, un sito

gestito dal modello.

Puoi fermarti al primo livello: con cinque tappe usi già l'IA per i lavori di tutti i giorni. Il secondo serve se passi le giornate in Word, Excel e PowerPoint; il terzo e il quarto quando vuoi organizzare lavori che durano settimane o affidare più passaggi di fila. Se non sai da dove partire, un quiz di tredici domande ti indica le tappe che ti servono e quanto tempo chiedono.

Ogni tappa ha la stessa forma. In testa, quanto dura la sessione, da venti minuti a un pomeriggio. Poi il gesto, su un lavoro vero e non su un esercizio, con le schermate dell'interfaccia. Poi un paragrafo che dice che cosa hai appena visto. In fondo, quattro riquadri, uno per strumento, perché gli strumenti sono quattro: Claude, ChatGPT, Copilot e Gemini, e la stessa funzione ha nomi, menu e piani diversi in ciascuno. Leggi soltanto il riquadro di quello che usi. Chiude il rimando a questo libro, con capitolo, sezione e pagina: la guida ti fa fare, il libro spiega che cosa succede dentro lo strumento, ed è quella spiegazione che ti permette di cavartela da solo quando l'interfaccia cambia di nuovo.

In coda stanno le schede che si leggono quando servono: che cosa aggiungono i piani a pagamento, che cosa cambia se lo strumento te lo dà l'ufficio, gli usi fuori dal lavoro, che cosa fare quando qualcosa non funziona, le parole che incontrerai. E una sezione sugli agenti personali, quelli che i produttori hanno cominciato a vendere nell'autunno 2026, che invecchia prima di tutte le altre e porta le date.

La guida si legge online, una pagina per tappa, e si scarica in PDF; esiste anche in inglese, insieme all'edizione inglese di questo libro. La rivedo ogni tre mesi e ogni indicazione porta la data della verifica: l'indirizzo è uno e resta quello, quello che trovi dietro si aggiorna.

Il manuale d'uso: librovivoia.it/in-pratica

INDICE

Il libro, pagina per pagina

Prefazione alla quarta edizione 13
Prefazione alla prima edizione 15
Nota dell'autore alla quarta edizione 17
Premessa. Da dove parti 19

Parte I - Capire 23
1. Da dove viene l'IA che usi ogni giorno 25
1.1 La nascita dell'IA 26
1.2 Dove siamo oggi 28
2. Come impara una macchina 31
2.1 IA simbolica: i sistemi esperti 32
2.2 Machine learning: l'apprendimento dai dati 34
2.3 Machine learning tradizionale: la cassetta degli attrezzi 36
2.4 Deep learning: il salto di qualità 39
2.5 Reti neurali artificiali: architetture di base 41
2.6 Reti neurali artificiali: architetture avanzate 59
3. I large language model 69
3.1 Come si addestra un modello di linguaggio 70
3.2 Grande non è sempre meglio: i modelli efficienti 72
3.3 Nuove tecniche per migliorare le capacità degli LLM 74
3.4 Chi fa gli LLM, e come li rilascia 76
3.5 Le architetture, in ordine di tempo 80
4. Che cosa sta intorno al modello 83
4.1 Adattare un modello al proprio materiale 84
4.2 La finestra di contesto 88
4.3 Capienza e attenzione 89
4.4 L'harness: l'impianto attorno al modello 89

Parte II - Come usare gli LLM 93
5. Ottenere quello che ti serve 95
5.1 Prompt engineering: come parlare ai modelli 95
5.2 Quando la chat si satura 98
5.3 La memoria degli LLM 98
5.4 Come far sì che l'IA si ricordi di te 100
6. Dall'assistente all'agente 103
6.1 Che cosa si intende per "Agentic AI" 104
6.2 I quattro pilastri dell'Agentic AI 105
6.3 Sfide dell'Agentic AI: autonomia, sicurezza e responsabilità 108
6.4 Controllare gli agenti: architetture, limiti e supervisione 109
6.5 Dagli agenti episodici agli agenti long-running 111
6.6 L'attenzione come risorsa scarsa: i sistemi multi-agente 113
6.7 Chi costruisce gli agenti, e con quale idea 115

7. Quando il modello sbaglia 117
7.1 Quando l'IA inventa: il problema delle allucinazioni 118
7.2 Quando l'IA ti dà ragione: la sicofanzia 120
7.3 Quando l'IA fa i conti: un problema di significato 122

Parte III - Verso l'uso aziendale 125
8. Scegliere e governare una soluzione 127
8.1 La prima domanda: cosa deve fare davvero il sistema? 128
8.2 La seconda domanda: quali dati servono e come gestirli? 128
8.3 La terza domanda: dove far girare il modello? 130
8.4 Che cosa c'è sul mercato 134
8.5 Governare e integrare il sistema 135
9. Ridisegnare il proprio lavoro 139
9.1 Usare e riprogettare 140
9.2 Smontare il processo in attività 141
9.3 Quali attività reggono la delega 142
9.4 Che cosa misuri 145
9.5 Chi verifica, e che cosa 146
9.6 I dati che non escono 147
9.7 Quando l'errore arriva 148
9.8 Il processo non è solo tuo 149
9.9 Delegare a un agente 150
9.10 Quattro processi smontati 151
9.11 Quello che resta tuo 159

Parte IV - Governare 161
10. I rischi che ti riguardano 163
10.1 Quando decide un algoritmo: discriminazioni e bias 164
10.2 La responsabilità: chi risponde davanti alla legge 164
10.3 L'illusione del controllo umano 165
10.4 Rendere l'IA comprensibile: come spiegare una decisione 167
10.5 Privacy e sorveglianza 167
11. Le regole 171
11.1 Unione Europea 172
11.2 Stati Uniti 173
11.3 Regno Unito 174
11.4 Cina 174
11.5 Italia 174
11.6 Verso un approccio globale 175
11.7 L'obbligo che ti riguarda, e come si adempie 175

Parte V - Guardare avanti 179
12. Chi possiede l'intelligenza artificiale 181
12.1 Il capitale: algoritmi, brevetti e chi addestra 182
12.2 I dati che nessuno può copiare 184
12.3 Le macchine che il denaro non compra 185
12.4 Chi decide: la somma, lo scaffale, gli Stati 187
12.5 Il conto: energia e acqua 188
13. Dove sta andando 191

13.1 Le biotecnologie: la prova che fa lavoro vero 191
13.2 La robotica: l'IA fuori dallo schermo 194
13.3 Le armi e i conflitti: l'autonomia dove costa di più 197
13.4 Verso l'intelligenza artificiale generale 199
13.5 Quello che nessuno ha ancora deciso 201

Parte VI - In pratica 205
Dalla chat alla delega 207
Conclusione. Da lunedì mattina 209

Appendici tecniche 211
A. L'esempio della rete a tre neuroni, passo per passo 212
B. Il calcolo della "perdita" di un modello 214
C. La matematica dietro il "gradient descent" 216
D. ANN e DNN, limiti e successive evoluzioni 218
E. Recurrent Neural Network (RNN) 219
F. Convolutional Neural Network (CNN) 223
G. Le "funzioni di attivazione" più comuni 227
H. Embedding 228
I. Positional encoding 234
J. L'addestramento di un adapter 235
K. Specialisti su richiesta: la logica delle Mixture of Experts 236
L. Le librerie in informatica 237
M. Le API 238
N. Il funzionamento di un Transformer 239
O. Python: guardare dentro con dieci righe di codice 252
P. I meccanismi di apprendimento 259
Q. Come funziona il Model Context Protocol 262
Glossario 269
Bibliografia 279

DOVE TROVARLO

Il libro continua online

«Intelligenza artificiale. Capire i modelli per delegare il lavoro», quarta edizione, è in vendita su Amazon dal 26 settembre 2026, in cartaceo (ISBN 979-8-1730-2627-9) e in ebook Kindle. La sesta parte, il manuale d'uso «Dalla chat alla delega», sta online e si aggiorna ogni tre mesi. Quello che invecchia in settimane, versioni dei modelli, prezzi, norme, non è stampato: sta su nove pagine di librovivoia.it, ognuna con la data dell'ultimo aggiornamento e la cronologia delle modifiche. Gli articoli del sito approfondiscono i singoli temi man mano che cambiano e sono il materiale delle prossime edizioni.

librovivoia.it