Memoria (artificiale)

Angelo Montanari (& Stefano Crespi Reghizzi)

Dipartimento di Scienze Matematiche, Informatiche e Fisiche
Università degli Studi di Udine

Udine, 7 aprile, 2024

Seminario 1: vocabolario filosofico dell'informatica

Algoritmo

Problema (algoritmico)

Problemi decidibili / indecidibili

Complessità (temporale) – oggi vedremo la complessità spaziale

Complessità di un algoritmo e complessità di un problema

Limiti superiori (un algoritmo) ed inferiori alla complessità di un problema

Algoritmi ottimali

Modelli di calcolo

La macchina di Turing

Un esempio: una macchina di Turing per il calcolo del successore di un numero

Seminario 2: ragionamento (artificiale)

IA simbolica e IA sub-simbolica

Intenzionalità

Intenzionalità (derivata) e intelligibilità delle macchine

Il gioco dell'imitazione (il test di Turing)

IA e robotica

La stanza cinese (l'esperimento mentale di Searle)

Induzione e apprendimento automatico

Bionica

Referenza: Stuart J. Russell, Peter Norvig, Intelligenza artificiale. Un approccio moderno, Pearson, 2021

Seminario 3: memoria (artificiale)

Introduzione

Tecnologie della memoria

Modelli astratti e concreti di gestione della memoria (nastri, pile, code e dischi)

Memoria e complessità computazionale

Memoria e scienza dei dati (dato e informazione, basi di dati, data warehouse e big data)

Memoria e previsione

Alcune questioni

Referenza: S. Crespi Reghizzi, A. Montanari, Memoria e previsione - Il punto di vista dell'informatica, in E-book su CINQUE PAROLE DELLA SCIENZA. Memoria e previsione, dato e informazione, tempo, a cura di Marco Bernardoni, EDB, 2021.

Introduzione

L'uso di memorie esterne a quella biologica ha segnato lo sviluppo intellettuale e sociale dell'umanità, dai primi dipinti o incisioni all'invenzione della scrittura, con i necessari suoi supporti fisici, e alla stampa.

Nel secolo scorso, l'invenzione e lo sviluppo delle memorie artificiali (elettroniche, magnetiche, ottiche) è stato e continua a essere concomitante con (e strumentale a) lo sviluppo delle tecnologie dell'informazione che permeano la società della cosiddetta terza rivoluzione industriale.

I diversi significati della parola "memoria"

Fra i tanti significati della parola "memoria", collegabili alla nozione di "informazione", vogliamo evidenziare i seguenti:

(i) La funzione psichica e organica che consente di riprodurre nella mente l'esperienza passata e le conoscenze apprese.

- Le tracce di tale funzione nel sistema nervoso.
- Apprendimento e ripetizione fedele, non necessariamente legati ad una comprensione corretta e completa.

(ii) Dispositivo fisico, oggi elettronico, utilizzato per immagazzinare dati/informazioni. Memoria di un cellulare, memorie per i saperi presenti in Internet, ecc. I dati memorizzati nel dispositivo hanno una sovrastruttura logica progettata per rendere possibili ricerche e aggiornamenti.

(iii) Memoria genetica (DNA).

Memoria naturale vs. memoria artificiale

La memoria è il supporto per conservare le conoscenze e riattivarle.

Nella memoria psichica, tali conoscenze assumono, per definizione, la forma completa dell'esperienza umana;

nella forma elettronica, esse rappresentano informazioni più o meno ricche, derivate dalle conoscenze umane e spesso raccolte e organizzate per rispondere a specifiche finalità pratiche.

Anche libri, opere d'arte ed epigrafi appartengono alla categoria delle memorie, indipendentemente dal materiale cartaceo, lapideo o elettronico del supporto, ma, a seconda del supporto, le modalità di accesso alle memorie, la facilità di accesso e di ricerca, la segretezza, la durevolezza e altri parametri importanti per il funzionamento possono variare in modo significativo.

Nelle slide successive sono riportate le caratteristiche essenziali dei dispositivi di memoria.

Parametri funzionali dei dispositivi di memoria - 1

Principali parametri funzionali dei dispositivi elettronici di memoria:

capacità	quanti Byte (= 8 bit): kilo(= 1000), mega, giga, tera, peta, exa (= 1000^6), zetta, yotta (= 1000^8).
	Esempio: 1 CD compact disk = 650 megaB; memoria di 1 uomo $\cong 2,5$ petaB (stima!)
capacità mondiale	2,6 exaB (1986) $\rightarrow$ 295 exaB (2006)
traffico Internet mensile	1 exaB (2007)$\rightarrow$ 21 exaB (2010)$\rightarrow$ 1 zettaB (2016 previsione)
volatilità	volatile: perde contenuto se manca alimentazione: { Non volatile: dischi ottici/magnetici, chiavette { Volatile: memoria centrale DRAM del PC
modalità d'accesso	{ in *sola lettura* come un disco di vinile { in *lettura e scrittura*

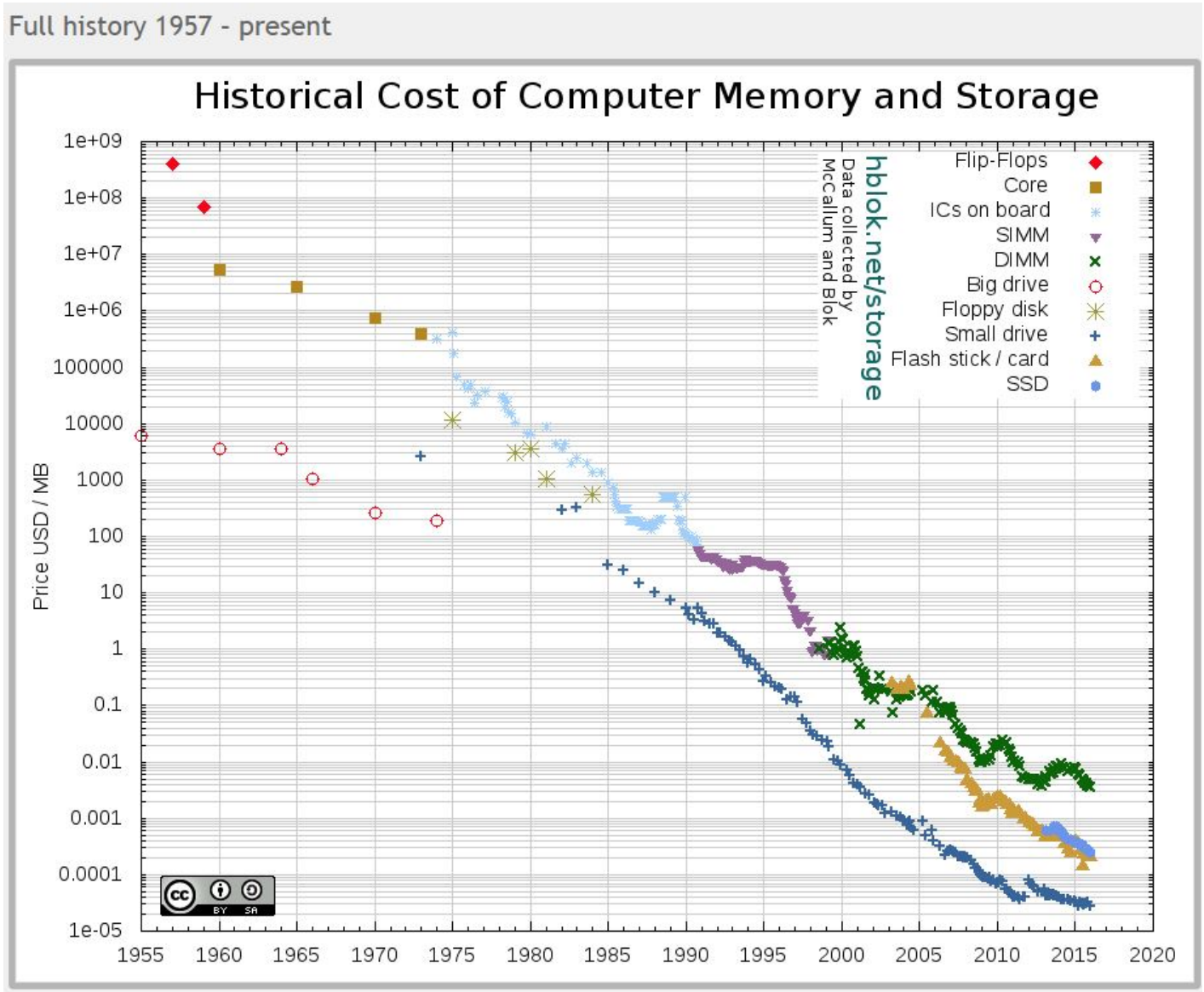
Parametri funzionali dei dispositivi di memoria - 2

velocità	quanti byte/sec si leggono o scrivono
gerarchia	piccola, velocissima la memoria CACHE del processore media, meno veloce la memoria centrale DRAM del PC grande, lenta la memoria non volatile esterna (disco/nastro/elettronica)
latenza	ritardo tra invio del comando di lettura di una cella di memoria e disponibilità del dato letto
blocco di memoria	una "pagina" di più byte viene letta/scritta in un solo passo; in essa il processore selezionerà la "parola" voluta
protezione da *errori*	si aggiunge qualche bit a ogni dato per controllare se nella memoria o nel transito da/verso la memoria si è corrotto
sicurezza e *segretezza*	si cifrano i dati prima di memorizzarli; si controlla l'autorizzazione del programma che vuole accedere ai dati
durata e *longevità*	quanti anni dura? Dopo quante operazioni si guasta? Esempio: chiavette sopportano meno scritture che letture Obsolescenza tecnologica (eempio: i dischi *floppy*)

Aspetti fisici, economici e sociali dei dispositivi

principio fisico	la distinzione memorie magnetiche/elettriche/ottiche, la distinzione memorie rotanti/stazionarie
stato cella	bit è rappresentato da tensione alta/bassa, da carica elettrica/magnetica, dallo stato fisico cristallino/amorfo, ...
densità	quanti bit per unità di superficie o per unità di volume di silicio o altro materiale. La cella si rimpicciolisce anno dopo anno: 300 $\mathrm{teraByte}/\mathrm{inch}^3$ [2014]
energia	quanti *joule* si consumano per leggere/scrivere un bit o, se volatile, per conservarlo. Il consumo dei grandi centri ("server" di Internet) è un problema ecologico ambientale
costo	costo di produzione storicamente scende in modo esponenziale
capacità mondiale	1986-2007: $\frac{\text{capacità mondiale delle memorie}}{\text{popolazione}}$ ha raddoppiato ogni 40 mesi.
scambi	1986-2007: $\frac{\text{capacità mondiale dei canali di comunicazione}}{\text{popolazione}}$ ha raddoppiato ogni 34 mesi

Costi storici dei dispositivi di memoria



https://hblok.net/blog/storage/

Trasmissione come memoria

La stessa unità (bit/secondo) misura la velocità della memoria e la capacità trasmissiva di un canale (fibra, micro-onde).

Nelle memorie i dati sono rappresentati nello stato fisico di una cella materiale.

Nello spazio tra due antenne i dati viaggiano con le onde elettromagnetiche che portano il segnale, non sono localizzati e non c'è trasporto di materia.

Questa è un'altra forma di memoria artificiale.

Potrebbe essere, come alcuni immaginano, che anche la memoria psichica sia rappresentata sia materialmente, nel cervello, sia in forme non localizzate?

Referenza: Alva Noë, Perché non siamo il nostro cervello, Raffaello Cortina Editore, 2010.

Memorie artificiali e naturali: somiglianze e differenze

Vanno sottolineate le poche **somiglianze**:

- memorie volatile/non volatile
  - memoria psichica è, o sembra essere, volatile,
    memoria genetica è meno volatile;
- memoria psichica ha due modalità: a breve e a lungo termine.

e le tante **differenze**:

- il ricordo d'un episodio vissuto $\neq$ attivazione di un blocco di memoria;
- il ricordo è quasi sempre soggettivo e parziale;
- il ricordo è selettivamente orientato da associazioni.

Alcune strutture informatiche dei dati cercano di simulare le funzioni delle memorie psichiche.

Modelli astratti di memoria

Modelli astratti di memoria: il nastro infinito (e gli stati) di una macchina di Turing.

Modelli astratti di memoria:

- LIFO (Last In First Out): pile (comportamento a sub-routine, strutture sintattiche annidate, ...);
- FIFO (First In First Out): code (comportamento equo, priorità dei servizi, ...);
- DNA.

Non realizzabilità fisica dei modelli astratti.

- Esempio: propagazione non istantanea dei segnali in una macchina di Turing con un nastro lunghissimo.

Gestione della memoria (concreta)

Sistemi operativi e gestione della memoria – il *file system* (modulo per la gestione della memoria):

- memoria primaria e memoria secondaria;
- organizzazione della memoria (dati vs. blocchi/pagine);
- allocazione dei blocchi/pagine;
- la nozione di memoria virtuale;
- il principio di località.

Distribuzione dei dati (memorie distribuite, cloud).

Memoria e complessità computazionale

Problemi indecidibili e problemi intrattabili.

Come misurare la complessità di un problema (e di un algoritmo): tempo, spazio, energia.

Limiti superiori (bubble sort) e inferiori (calcolo del massimo di un insieme di numeri) alla complessità di un algoritmo.

Complessità temporale e spaziale di un algoritmo.

Un esempio: stabilire se due numeri interi, rappresentati in binario, sono uguali o meno.

- Complessità spaziale: costante (non occorre memorizzare l'input).
- Complessità temporale: lineare nella dimensione dell'input (al più un numero di operazioni di confronto pari al numero di cifre dell'intero più piccolo).

Complessità temporale vs. complessità spaziale

Modello di calcolo di riferimento: la macchina di Turing (assumiamo per semplicità un alfabeto binario, al quale va aggiunto il simbolo vuoto).

- numero (finito) di stati della macchina: k;
- input di lunghezza n;
- tempo (= numero di passi) massimo: $t(n)$;
- spazio (= numero di celle) massimo: $s(n)$.

E' facile vedere che $t(n) \geqslant s(n)$ (il numero di celle utilizzate mai eccede il numero di passi compiuti).

E' anche possibile mostrare che $t(n) \leqslant k \times s(n) \times 3^{s(n)}$ (numero di possibili configurazioni diverse della macchina di Turing su uno spazio massimo utilizzato di dimensione $s(n)$, che tiene conto del numero k di stati e del numero $s(n)$ di posizioni della testina).

PTIME vs. PSPACE

Problemi trattabili: la classe dei problemi PTIME, ossia dei problemi che possono essere risolti in un numero di passi polinomiale nella dimensione dell'input (complessità asintotica che ignora il grado e le costanti del polinomio).

La classe dei problemi PSPACE: problemi che possono essere risolti utilizzando uno spazio di dimensione polinomiale nella dimensione dell'input.

PTIME è meglio di PSPACE (ma LOGSPACE è meglio di PTIME: LOGSPACE $\subseteq$ PTIME $\subseteq$ PSPACE e LOGSPACE $\subset$ PSPACE).

Esempi di problemi e situazioni in cui l'avere poca memoria obbliga a perdere tempo (ordinamento di un file di dati in memoria secondaria).

Dati e informazioni: il contesto tradizionale

Contesto tradizionale (basi di dati): **dati** usati per codificare **informazioni** di interesse (funzione strumentale dei dati rispetto alle informazioni).

Informazioni utili recuperate attraverso interrogazioni (i dati diventano informazione quando restituiti in risposta a determinate richieste/interrogazioni).

L'informazione è all'origine del processo di progettazione, sviluppo e popolamento della base di dati ed è il risultato finale del processo di interrogazione della base di dati.

Dati e informazioni: l'ambito dei big data

Nell'ambito dei **big data**, la prospettiva cambia drasticamente.

Grandi, in molti casi enormi, quantità di dati vengono raccolte in un arco di tempo più o meno lungo, spesso senza seguire specifiche strategie di acquisizione.

Informazioni interessanti (correlazioni del tutto impreviste fra elementi diversi) vengono scoperte analizzando in modo sistematico tali dati (data mining).

I tipici costrutti dei linguaggi di interrogazione tradizionali delle basi di dati vengono sostituiti da strumenti di natura statistica e da metodi e procedure di apprendimento automatico (machine e deep learning).

I sistemi di basi di dati

Caratteristiche distintive di una base di dati:

- grandi quantità di dati;
- persistenza dei dati;
- globalità dei dati.

Altre caratteristiche fondamentali:

- efficienza;
- efficacia (convenienza).

Struttura fisica e logica dei dati: indipendenza fisica dei dati.

Strutture dati orientate alle applicazioni (tabelle relazionali come rappresentazione di conoscenze estensionali).

Le strutture di indicizzazione

File di dati e *file* indice: uso di strutture dati ausiliare per rendere più efficente l'accesso ai dati in memoria secondaria.

Indici e strutture di indicizzazione:

- analogie e differenze con la nozione di indice di un libro;
- indici di singolo livello e multilivello;
- indici multilivello statici e dinamici (B-alberi e B^+-alberi);
- indici per basi di dati complesse (ad esempio, basi di dati geografiche).

Data warehouse

Integrazione di grandi quantità di dati provenienti da basi di dati (sorgenti) diverse e spesso eterogenee.

Strumento per l'analisi dei dati a supporto di processi di decisione (business intelligence).

Denormalizzazione dei dati (il rispetto delle forme normali è fondamentale in un sistema transazionale; può essere rilasciato in sistemi OLAP).

Utilizzo di strumenti di statistica descrittiva.

Big data

L'espressione "big data" è usata per indicare un'enorme collezione di dati (dell'ordine degli Zettabyte, ovvero miliardi di Terabyte) che per dimensioni, eterogeneità e dinamicità richiede metodi, tecniche e strumenti di analisi ad hoc.

L'analisi dei dati nel loro complesso fornisce informazioni che l'analisi indipendente di singoli porzioni in cui l'insieme completo di dati può essere partizionato non è in grado di dare.

Si tratta spesso di dati solo parzialmente strutturati (dati semistrutturati) o totalmente privi di struttura.

Necessità di strumenti di avanzati per organizzare (sistemi noSQL), memorizzare (cloud, virtualizzazione), elaborare (high performance computing) e analizzare (data mining e statistica inferenziale) i dati.

L'approccio map-reduce

Proposta di nuovi schemi di rappresentazione dei dati (in ambito business analytics) che consentono di gestire enormi moli di dati con elaborazioni in parallelo di una molteplicità di basi di dati.

Architetture per l'elaborazione distribuita di enormi quantità di dati:

- MapReduce (Google);
- Apache Hadoop (open source).

L'approccio map-reduce:

- decomposizione di un problema/compito in più componenti, distribute su più nodi;
- esecuzione dei diversi compiti in parallelo sui diversi nodi (funzione map);
- raccolta, integrazione e restituzione dei risultati (funzione reduce).

Memoria e previsione

Una legge naturale, un modello matematico e un programma di simulazione del medesimo sono strumenti ben noti per effettuare previsioni sulla base di esperienze precedenti. La crescita in realismo e accuratezza dei modelli predittivi va di pari passo col progresso scientifico.

Novità recente è la disponibilità di enormi quantità di dati – *big data* – ad esempio, in ambito economico, meteorologico o biologico – e la possibilità di estrarre da essi in modo automatico leggi e modelli, attraverso algoritmi di apprendimento e/o analisi di natura statistica.

Per alcuni, la verifica della validità dei modelli così ottenuti non richiederebbe più alcun lavoro sperimentale, essendo i dati già disponibili ben più ricchi di quelli che si potrebbero misurare sperimentalmente, con costi e difficoltà spesso impraticabili.

Un cambio di paradigma scientifico?

Approccio classico (guidato dai modelli).

A partire da un certo insieme di osservazioni, esperienze e esperimenti (la natura concreta di tali elementi varia al variare della disciplina considerata), vengono formulate delle teorie/modelli, che successivamente vengono validati/falsificati sperimentalmente.

Nuovo paradigma (guidato dai dati).

Disponendo di una quantità sufficiente di dati, dotati di un'adeguata garanzia di significatività statistica, è possibile estrarre da essi in modo automatico, mediante algoritmi di apprendimento e/o di analisi statistica, leggi e modelli di valore generale.

Data mining (e machine learning)

Data mining: insieme di metodologie, tecniche e strumenti che consentono estrarre informazioni significative da grandi quantità di dati, più o meno strutturati, mediante l'utilizzo di strumenti automatici o semi-automatici.

Le strategie di data mining si possono suddividere in:

- supervisionate - i valori di output dipendono dai valori di input e vengono utilizzati per effettuare delle predizioni (classificazione, propensione, analisi di serie storiche, regressione);
- non supervisionate - si cercano generiche relazioni fra i dati tramite tecniche di clustering e individuazione di regole di associazione dei dati.

Alcune questioni tecniche

Ci sono diverse questioni di natura tecnica che hanno un notevole impatto sul significato e la validità dei risultati ottenuti.

Quali requisiti deve soddisfare l'insieme dei dati a disposizione (dataset) per garantire qualità e affidabilità dei risultati?
Alcuni fattori: numerosità, rappresentatività, ..

Come suddividere i dati disponibili tra training set e test set?

Il problema dell'overfitting. La carta geografica di Borges, o della classificazione spinta al limite: un individuo, una classe.

Alcune questioni - Passato e futuro

Una scienza costruita sui dati storici disponibili è una scienza che guarda solo al passato, e quindi intrinsecamente conservatrice, in quanto essa è "strutturalmente" costretta a escludere le novità,

E' una scienza che costruisce modelli delle persone e dei sistemi artificiali analizzando il loro comportamento passato e li utilizza per prevedere quello futuro.

Problema: il comportamento futuro non necessitato da quello passato. La novità è una caratteristica distintiva del comportamento umano e della storia umana e, a meno che il futuro non sia completamente determinato dal passato, è anche una caratteristica distintiva della natura.

Per quanto riguarda le macchine, si può ipotizzare una qualche forma di ripetizione rigida per i sistemi più semplici (e chiusi), ma non è una caratteristica dei sistemi aperti complessi.

Alcune questioni - Il problema dei dati

I dati non esistono di per sé. I dati dipendono dall'osservatore, uomo o macchina, che li raccoglie.

Dipendono anche dai dispositivi di misurazione, poiché si possono raccogliere solo dati che possono essere effettivamente misurati/acquisiti.

Per stabilire quali sono i dati rilevanti da raccogliere (se è possibile raccoglierli), sono necessari un obiettivo e un piano per raggiungerlo. Entrambi possono essere formulati solo con riferimento ad un esplicito o modello implicito del mondo (una teoria).

Non esistono dati "neutri" e i dati "esistenti" sono sia di fatto che in linea di principio incompleti

Alcune questioni - Interpretabilità dei dati

Interpretabilità o meno dei risultati: possibilità di fornire un'interpretazione dei risultati che consenta di comprenderne le ragioni e il significato.

Human-in-the-loop: un modello/sistema che prevede l'interazione col soggetto umano (ad esempio, nei processi di decisione).

Non sempre è possibile garantire tale condizione (applicazioni finanziarie e sistemi critici dal punto di vista della sicurezza).

In alcuni casi, è essenziale (diagnosi mediche «critiche»).

Trustworthy AI (IA affidabile).

La nuova frontiera dell'intelligenza artificiale: explainable IA (trustworthy AI). L'IA di cui ti puoi fidare, perché è in grado di rendere ragione/spiegare le conclusioni/previsioni formulate.

Importanza della spiegazione in tutte le applicazioni dell'IA che hanno un impatto rilevante sulla vita delle persone e delle comunità.

Nel caso del ragionamento automatico (simbolico) la questione della spiegazione è legata alla complessità logica e/o computazionale del processo, nel caso del ragionamento sub-simbolico tale questione riguarda la natura stessa del processo che produce il risultato.

L'AI Act europeo: come coniugare libertà e responsabilità nell'IA (approvato il 14 marzo 2024 dal Parlamento Europeo).

Considerazioni finali

Nel mondo continuerà a lungo, salvo catastrofi, la crescita esponenziale della capacità delle memorie e delle reti di trasmissione e il raffinamento dei metodi per accedere e controllare gli accessi alle memorie.

Intelligenza dei dati: comprensione vs. analisi statistica (analogia con l'elaborazione del linguaggio naturale).

Non c'è coscienza (intenzionalità) senza memoria (umana), ma anche non c'è memoria (umana) senza coscienza (intenzionalità): sostanziale differenza tra memorie naturali e memorie tecnologiche (nelle memorie tecnologiche c'è un'intenzionalità derivata).