Giovanni Sartor

L'intelligenza artificiale e il diritto

Giappichelli

Prefazione

Negli anni Ottanta, la rivoluzione del personal computer ha diffuso l'informatica nella società, nell'industria, nelle professioni, nelle case. I giuristi si sono interessati ai problemi giuridici dell'informatica, dalla protezione dei dati, alla tutela del software, al diritto d'autore, ai reati informatici. Al tempo stesso hanno iniziato ad usare i computer per scrivere testi e accedere a banche dati locali e online.

Negli anni Novanta, la rivoluzione di Internet ha collegato computer e persone in una rete globale, divenuta inesauribile fonte di informazione e luogo di ogni interazione. La vita economica e sociale si è trasferita nella dimensione virtuale o, anzi, in un'infosfera fisico-virtuale piena di enormi quantità di dati digitali e abitata da macchine interconnesse, che incessantemente elaborano quei dati e comunicano, tra di loro e con le persone. I temi del diritto dell'informatica hanno assunto un'importanza crescente, spesso interessando trasversalmente diversi settori del diritto. Si pensi, per esempio, al commercio elettronico, alla protezione dei dati raccolti, alla responsabilità dei fornitori di servizi digitali, ai diritti sui dati e contenuti digitali, alla concorrenza nei mercati digitali. Al tempo stesso, le applicazioni dell'informatica hanno modificato le attività del giurista, in ambiti quali il processo telematico, la creazione di atti e documenti digitali, la documentazione giuridica online.

Oggi siamo di fronte alla terza rivoluzione, quella dell'intelligenza artificiale (IA). Attività che richiedono intelligenza, fino ad oggi svolte esclusivamente dalle persone, possono, in misura sempre maggiore, essere affidate alle macchine, che hanno acquisito capacità di ragionare, apprendere e agire. Applicazioni fino a ieri impossibili —come la comprensione vocale, la traduzione automatica, il riconoscimento di oggetti— sono alla portata di ogni smartphone. Un insieme sempre più ampio di funzioni può essere delegato a tecnologie intelligenti: decisioni automatiche, predizioni sui comportamenti di individui e gruppi, controllo su ambienti di lavoro e spazi pubblici, riconoscimento biometrico, governo di robots, guida di veicoli autonomi, ecc. Ciò solleva nuovi problemi giuridici, rispetto ai quali spesso non esistono risposte definitive. Anche la pratica del diritto è suscettibile di profonde modifiche: ai sistemi esperti, in grado di applicare norme formalizzate in modo automatico, si affiancano funzioni di apprendimento automatico, capaci di estrarre informazioni da grandi masse di dati e di costruire e applicare modelli predittivi e decisionali.

Nel primo stasimo dell'Antigone, il coro usa le seguenti parole per descrivere le capacità umane: "Molte cose sono meravigliose e terribili ma nessuna lo è più dell'uomo"[1], e aggiunge che le abilità umane, grandi "al di là di ogni speranza", possono indirizzarsi "talvolta al bene, altre volte al male".

Oggi guardiamo con lo stesso atteggiamento di ammirazione e paura alle prospettive dell'IA, di cui immaginiamo sviluppi al di là di ogni aspettativa. Così il giurista si chiede se le tecnologie dell'IA possano essere controllate e dirette dal diritto verso il bene degli individui o della società o se invece saranno rivolte a interessi particolari o addirittura finiranno per travolgere le istituzioni che oggi conosciamo. Rispetto al proprio lavoro, egli si chiede se le stesse tecnologie potranno aiutarlo ad applicare la legge con maggiore efficienza ed efficacia, contribuendo a realizzare valori di razionalità, e giustizia, o se invece finiranno per sostituire l'attività umana con la decisione automatica, o comunque per dominare su di essa, facendo del giurista stesso un servitore della macchina.

Questa attitudine di speranza e preoccupazione nei confronti dell'IA è pienamente giustificata dalle opportunità e dai rischi che le tecnologie intelligenti sembra dischiudere, e dalla grande incertezza rispetto ai loro possibili sviluppi, alle loro applicazioni, al modo in cui tali applicazioni potranno essere governate. Tuttavia, per cogliere le opportunità e rischi presenti già oggi o probabili nel vicino futuro, è necessario un approfondimento. Solo grazie ad una comprensione sufficientemente precisa delle tecnologie dell'IA e dei relativi problemi sociali e giuridici, il giurista potrà contribuire ad approntare e applicare una regolazione adeguata dell'uso dell'IA, che ne garantisca la coerenza con valori individuali e sociali. Solo su questa base, egli potrà partecipare alla definizione e all'impiego efficace delle tecnologie dell'IA nella pratica del diritto, nel rispetto dei valori giuridici.

Il presente volume intende fornire un primo, modesto e preliminare, contributo in questa direzione. Si articola in quattro capitoli.

Nel primo si introduce il concetto di IA, illustrando le diverse prospettive dalle quali si è guardato a questa disciplina e alle sue realizzazioni, individuandone limiti e prospettive.

Nel secondo si esaminano le tecnologie dell'IA. Si descrivono i due principali indirizzi per sviluppo dei sistemi intelligenti: la rappresentazione della conoscenza e l'apprendimento automatico.

Nel terzo si considerano opportunità e rischi dell'IA, e i modi nei quali l'etica e il diritto ne possono governare sviluppi e applicazioni.

Infine, nel quarto si esaminano le applicazioni giuridiche dell'IA, dai sistemi basati su regole, all'argomentazione, al ragionamento basato sui casi, all'apprendimento automatico, alla giustizia predittiva.

I riferimenti bibliografici sono limitati ai soli temi trattati, omettendo, in particolare, ogni riferimento al ricchissimo dibattito dottrinale sull'intelligenza artificiale e le decisioni algoritmiche. Per le opere in lingua inglese (che costituiscono gran parte del-

la bibliografia in materia di informatica), data la rapidità con cui si susseguono nuove edizioni e traduzioni, è sembrato preferibile indicare il testo originale.[2]

Sono grato a quanti mi hanno aiutato nel mettere a punto questo lavoro, in particolare i colleghi Raffaella Brighi, Giuseppe Contissa, Francesca Lagioia, Monica Palmirani e Antonino Rotolo, con i quali ho condiviso gli studi di Informatica Giuridica presso il CIRSFID - Alma AI e il Dipartimento di Scienze Giuridiche dell'Università di Bologna, e il Law Department dell'European University Institute di Firenze. Infine, uno speciale ringraziamento a Enrico Pattaro che mi ha incoraggiato e sostenuto nelle mie ricerche di Intelligenza Artificiale e Diritto parecchi anni fa, quando questa tematica era ancora largamente inesplorata.

Il presente volume è un risultato del progetto di ERC-Advanced CompuLaw, Grant agreement N. 833647. Ringrazio la Commissione Europea e ERCEA per il generoso sostegno alle mie ricerche.

Capitolo 1

L'intelligenza artificiale

Nel presente capitolo si introduce dapprima il concetto di IA, illustrando le diverse prospettive dalle quale si è guardato a questa disciplina e alle sue realizzazioni, in diversi contesti, fisici e virtuali. Si esamina il dibattito sulle capacità e i limiti dell'IA, e delle tecnologie di IA oggi disponibili. Si presenta infine l'evoluzione delle ricerche di IA, dalle origini fino ai nostri giorni.

1.1 Il concetto di IA

Per introdurre il concetto di IA si partirà dall'idea di intelligenza, per poi considerare come l'intelligenza possa essere "artificiale". Quindi, si esamineranno i problemi inerenti a una definizione "giuridica" di IA.

1.1.1 L'intelligenza

Come è noto, manca una definizione univoca e condivisa di intelligenza. Uno dei più autorevoli testi introduttivi in materia, l'*Oxford Companion to the Mind*, apre la trattazione della voce "intelligence" dicendo che "sono disponibili innumerevoli test per misurare l'intelligenza, ma nessuno sa con sicurezza che cosa sia l'intelligenza, e addirittura nessuno sa con sicurezza che cosa misurino i test disponibili".[3]

Si suole peraltro convenire che l'intelligenza si rivela nella capacità di svolgere diverse funzioni, come le seguenti: l'adattamento all'ambiente e in particolare a nuove situazioni), l'apprendimento dall'esperienza, la percezione, l'intuizione, il pensiero astratto, l'utilizzo efficiente di risorse limitate, la comunicazione, e così via. Tali funzioni, tanto diverse tra loro, sono unite dal fatto che consentono a chi le possieda di migliorare le proprie prestazioni, di agire in modo più efficace ed efficiente (di raggiungere meglio i propri scopi, con un minore dispendio di risorse), grazie all'acquisizione e all'elaborazione di informazioni e all'adozione di azioni conseguenti. L'intelligenza è oggetto di diverse discipline,[4] tra cui possiamo ricordare brevemente le seguenti:

- la filosofia, che fin da Platone e Aristotele ha individuato nell'intelligenza o razionalità una caratteristica fondamentale dell'uomo e ne ha fatto uno dei temi principali della propria ricerca,[5] studiando i procedimenti del pensiero (logica), i principi della conoscenza e della scienza (gnoseologia ed epistemologia), e le strutture dei concetti, nel loro collegamento con la realtà (ontologia);
- la matematica, che ha formalizzato i metodi del pensiero nei linguaggi e nelle tecniche della logica formale e della teoria della probabilità, e ha altresì affrontato i problemi della computabilità;
- l'economia, che ha elaborato tecniche per l'uso efficiente di risorse limitate, anche in contesti nei quali la determinazione e la valutazione delle conseguenze delle azioni è difficile (teoria delle decisioni) o nei quali il singolo agente deve tener conto delle scelte altrui (teoria dei giochi);
- la medicina, che ha studiato l'elaborazione delle informazioni nel cervello (neurologia) così come il funzionamento degli organi sensoriali;
- la psicologia, che ha esaminato il funzionamento della mente umana, in particolare nell'apprendimento (psicologia cognitiva), rappresentandola come un processo di elaborazione di informazioni (scienza cognitiva);
- la linguistica, che ha considerato i procedimenti che danno luogo alla formulazione e alla comprensione del linguaggio, traducendoli talvolta in programmi informatici (linguistica computazionale).

L'IA ha tratto ispirazione da tutte le ricerche appena menzionate, ma ha aggiunto a queste un aspetto ingegneristico: l'IA non vuole solo studiare l'intelligenza, ma si propone di costruirla, di dar vita ad artefatti intelligenti. L'obiettivo ingegneristico dell'IA non esclude che essa possa contribuire alla conoscenza dell'intelligenza umana. Come osservava Gian Battista Vico [1668-1744] *verum esse ipsum factum* (il vero è ciò che è fatto), o *verum et factum convertuntur* (il vero e il fatto si convertono l'uno nell'altro): come dallo studio dell'intelligenza umana si possono trarre utili indicazioni al fine della costruzione dell'IA, così la costruzione dell'IA (il fatto) può aiutarci a cogliere la natura dell'intelligenza (il vero) e in particolare possiamo trarne ipotesi (da verificare empiricamente) circa il funzionamento dell'intelligenza umana.[6] Poiché le facoltà conoscitive da realizzare nei sistemi di IA corrispondono, almeno in parte, alle facoltà in cui si esplica l'intelligenza naturale (umana o animale), non dobbiamo stupirci se l'IA trae ispirazione dall'intelligenza naturale, trovando in questa soluzioni appropriate alle proprie esigenze di elaborazione dell'informazione, né dobbiamo stupirci se ritroviamo nell'intelligenza naturale (nelle strutture cerebrali o nei processi mentali) alcune soluzioni ingegneristiche elaborate dall'IA.

L'affinità funzionale tra IA ed intelligenza umana non esclude peraltro che vi siano importanti differenze.

L'intelligenza umana è infatti realizzata da un hardware (le cellule cerebrali e sensoriali) profondamente diverso dall'hardware dell'IA (chip di silicio, telecamere e altri sensori). Abbiamo ancora conoscenze limitate rispetto al funzionamento del cervello umano, ma possiamo certamente affermare che la complessità dello stesso —che comprende circa 100 miliardi di neuroni, con 1000 miliardi di connessioni— va molto al di là dei sistemi artificiali oggi disponibili. Solo gli esseri umani, come vedremo, sono dotati di intelligenza generale, così da poter affrontare i diversi problemi che si presentano nel corso della loro vita; l'intelligenza dei sistemi artificiali si esplica solo in ambiti specifici, per i quali i sistemi in questione sono stati progettati.

I processi cognitivi umani sono altamente paralleli, implicando l'attivazione contemporanea di un elevato numero di neuroni, secondo modalità ancora largamente sconosciute (benché la neurologia abbia fatto enormi progressi negli ultimi anni). I sistemi artificiali sono più semplici, ma le elaborazioni elementari che essi svolgono sono molto più veloci, e possono essere applicate ad enormi masse di dati. Pertanto, i sistemi artificiali hanno prestazioni molto superiori in talune forme di elaborazione dell'informazione (come l'effettuazione di calcoli numerici o il concatenamento di un elevato numero di regole precise e predeterminate). All'opposto, in altre elaborazioni (come quelle che attengono all'interpretazione di situazioni inusuali, alla comprensione dei significati, alla formulazione di nuove ipotesi e analogie) i sistemi automatici sono assai inferiori.

Una fondamentale differenza tra il sistema nervoso umano e i "cervelli artificiali" è che il primo è immerso nel corpo. Quindi la cognizione umana interagisce in modi complessi con le funzioni biologiche svolte dagli organi e con i processi del metabolismo (si pensi ad esempio, a come una disfunzione nel corpo generi sensazioni di dolore e disagio, che a loro volta attivano processi mentali e corporei).[7] Nei secondi, anche quando si tratti di robot destinati ad operare nell'ambiente fisico, l'integrazione di aspetto corporeo e cognitivo è assente o presente in modo elementare.

1.1.2 Idee di IA

Stuart Russell e Peter Norvig, celebri studiosi di IA (e autori del più diffuso manuale in materia) distinguono i diversi modi di accostarsi all'intelligenza secondo due diverse dimensioni:[8]

- l'idea che l'intelligenza consista prevalentemente nel pensiero (rappresentazione della conoscenza e ragionamento) si contrappone all'idea che in essa l'interazione con l'ambiente (percezione e azione) svolga un ruolo preminente (o almeno altrettanto importante);
- l'obiettivo di riprodurre fedelmente le capacità intellettive dell'uomo (con tutti i loro limiti) si contrappone all'obiettivo di realizzare sistemi capaci di razionalità (cioè di elaborare informazioni o agire in modo ottimale) prescindendo dai limiti della razionalità umana.

Sistemi che pensano come esseri umani	**Sistemi che pensano razionalmente**
"Il nuovo eccitante tentativo di fare in modo che i calcolatori pensino [...] di costruire *macchine dotate di menti*, nel senso pieno e letterale"[9] "[L'automazione delle] attività che associamo con il pensiero umano, attività quali prendere decisioni, risolvere problemi, imparare"[10]	"Lo studio di facoltà mentali mediante l'uso di modelli computazionali"[11] "Lo studio delle elaborazioni che rendono possibile percepire, ragionare, e agire"[12]
Sistemi che agiscono come esseri umani	**Sistemi che agiscono razionalmente**
"L'arte di creare macchine che svolgono funzioni che richiederebbero intelligenza quando svolte da persone"[13] "Lo studio di come far fare ai calcolatori cose nelle quali, al momento, le persone sono migliori"[14]	"L'intelligenza computazionale è lo studio della progettazione di agenti intelligenti"[15] "L'IA [...] si occupa del comportamento intelligente negli artefatti"[16]

Figura 1.1: *Definizioni dell'intelligenza artificiale (IA)*

Conseguentemente, gli stessi Russell e Norvig propongono lo schema della Figura 1.1, che riporta diverse autorevoli definizioni di IA distinguendole a seconda di come si collocano rispetto alle due dimensioni appena indicate.

Al riguardo si limitiamo ad alcune brevi considerazioni. Per quanto attiene alla distinzione tra pensiero e azione, basta ricordare come il comportamento intelligente richieda il collegamento tra il momento epistemico (volto a determinare come stanno le cose, come è fatto il contesto nel quale l'agente si trova e quali dinamiche lo caratterizzano) e il momento pratico (volto a determinare il comportamento più appropriato rispetto agli interessi dell'agente, nel contesto della sua azione): gli interessi epistemici (quali cose un agente desidera conoscere) sono determinati anche dagli obiettivi pratici dell'agente (da che cosa esso intenda realizzare o conservare) , e i modi del perseguimento degli obiettivi pratici (e il giudizio preliminare sulla possibilità di raggiungere tali obiettivi) dipendono dalle nostre conoscenze epistemiche. Per esempio, un viaggiatore è interessato a conoscere qual è la soluzione più rapida ed economica per arrivare ad una destinazione nel momento in cui desidera andarci, e la scelta su come viaggiare (per treno, automobile o aereo) dipende dalle conoscenze che il viaggiatore ha ottenuto su tempi, costi e impatto climatico (se è interessato agli impatti delle proprie azioni sul pianeta).[17]

L'attenzione dell'IA per l'aspetto pratico è cresciuta negli anni più recenti quando —in parallelo con sviluppi tecnologici di cui parleremo nelle pagine seguenti, come in particolare la creazione di robot fisici e bot virtuali (software)— si sono sviluppate indagini volte a cogliere il comportamento razionale nella relazione tra l'agente e il suo ambiente. Tali ricerche hanno enfatizzato aspetti dell'intelligenza non riducibili al ragionamento in senso stretto, come la percezione e la capacità di esplorare attivamente l'ambiente. Una formulazione estrema di questa tesi può ritrovarsi nelle seguenti parole di Rodney Books, pioniere della ricerca nella robotica comportamentale, oltre che inventore e imprenditore di successo (tra i prodotti di iRobot, l'impresa da lui fondata con altri ricercatori, c'è Roomba, l'aspirapolvere robotico di cui sono state vendute milioni di copie):

> Il comportamento di risoluzione di problemi, il linguaggio, la conoscenza di esperti e la sua applicazione, e la ragione, sono tutti semplici una volta che siano disponibili l'essenza dell'esistere e del reagire. Questa essenza è l'abilità di spostarsi in un ambiente dinamico, percependo ciò che sta attorno a un livello sufficiente per realizzare il necessario mantenimento di vita e riproduzione. Questa parte dell'intelligenza è quella in cui l'evoluzione ha concentrato il suo tempo—essa è molto più difficile. Credo che la mobilità, una visione acuta e l'abilità per eseguire compiti collegati alla sopravvivenza in un ambiente dinamico forniscano una base necessaria per lo sviluppo di vera intelligenza.[18]

Per quanto attiene alla distinzione tra l'obiettivo di riprodurre pienamente il pensiero umano (comprese le sue irrazionalità) e quello di sviluppare invece procedure cognitive razionali, molto dipende dall'obiettivo di un'applicazione di IA: simulare l'uomo o affrontare nel modo migliore certi problemi. Bisogna però ricordare che la conoscenza dei procedimenti cognitivi e deliberativi umani è ancora assai limitata: vi sono molte cose che l'uomo riesce a fare in modo appropriato spontaneamente, senza sapere in che modo riesca a raggiungere tale risultato.

La natura ci ha dotato di capacità adatte ad affrontare in modo adeguato il mondo in cui ci troviamo[19] e siamo in grado di utilizzare tali facoltà pur senza conoscere le modalità del loro funzionamento, e quindi, a maggior ragione, senza conoscere le ragioni a sostegno di tali modalità. Ciò vale non solo per capacità specifiche (come quella di riconoscere le facce delle persone che incontriamo) ma anche per le nostre generali capacità linguistiche, logico-matematiche, e in generale per le competenze richieste nella soluzione di problemi.

A questo riguardo è opportuno ricordare il concetto di *razionalità limitata*, elaborato dallo studioso di IA (e premio Nobel per l'economia) Herbert Simon. Scelte che appaiono irrazionali con riferimento a un concetto ideale di razionalità (non assicurando un risultato ottimale, cioè il migliore risultato possibile) possono invece apparire appropria-

te (razionali nella misura in cui ci è possibile esserlo) quando si considerino i limiti delle nostre capacità conoscitive e la complessità dell'ambiente.[20]

La nostra stessa ragione ci vieta di sprecare le nostre energie nell'impossibile ricerca della scelta ottimale, e ci richiede invece di seguire procedure cognitive fallibili, ma rapide ed economiche (richiedenti un impegno limitato delle nostre risorse mentali) che conducano a risultati sufficientemente buoni (anche se non ottimi) nella maggior parte dei casi:

> Non possiamo, entro limiti computazionali praticabili, generare tutte le alternative ammissibili e comparare i loro rispettivi vantaggi. Né possiamo riconoscere l'alternativa migliore, anche se siamo abbastanza fortunati da generarla subito, finché non le abbiamo viste tutte. Realizziamo scelte sufficientemente buone ricercando alternative in un modo tale da consentirci, di regola, di trovarne una di accettabile dopo una ricerca limitata.[21]

Queste procedure fallibili tese a economizzare le energie richieste dall'impiego della ragione sono chiamate *euristiche*.[22]

Ciò che può apparire un difetto della razionalità umana (una forma di irrazionalità), può invece rivelarsi una procedura cognitiva appropriata per una razionalità limitata: emulare (copiare) l'intelligenza umana, anche in aspetti apparentemente irrazionali (o solo limitatamente razionali) può talvolta condurre a soluzioni efficaci. I sistemi informatici hanno enormi capacità di calcolo e memoria. Tuttavia, le euristiche diventano necessarie anche per i sistemi informatici, quando i dati accessibili siano limitati, o quando il problema da affrontare presenti un'elevata complessità computazionale.

Alcuni studiosi hanno affermato la necessità di distinguere nettamente, anche sotto il profilo concettuale e terminologico, le capacità cognitive artificiali ed umane. Tale necessità discenderebbe dal fatto che la terminologia psicologica, o "cognitiva" —intelligenza, conoscenza, ma anche percezione, credenza, intenzione, o volontà, razionalità— ha la funzione di descrivere la mente e le competenze degli esseri umani. L'estensione di questi termini a enti artificiali condurrebbe ad attribuire a tali enti attitudini e capacità che, almeno nei limiti delle tecnologie odierne, essi non possono avere. In particolare, si è escluso che sistemi artificiali possano dirsi intelligenti. Ad essi può essere riconosciuta solo una capacità di "agire smart," senza intelligenza,[23] o una capacità di comunicare, senza comprendere i significati.[24]

Altri studiosi, invece, preferiscono usare la terminologia cognitiva anche per descrivere comportamenti e attitudini dei sistemi artificiali. Pertanto i termini cognitivi sono intesi in un significato astratto (semplificato, e quindi più generale), così che essi siano applicabili sia agli esseri umani sia alle macchine. Seguendo questa seconda prospettiva, anche ad agenti artificiali si possono attribuire le capacità cui fanno riferimento quei termini: essi possono cogliere aspetti della realtà (percezione) e dotarsi di rappresentazioni di tali aspetti (credenze), possono avere obiettivi da perseguire (desideri, intenzioni, sco-

pi), possono elaborare informazioni per trarne ulteriori contenuti (mediante inferenze epistemiche e pratiche) e agire di conseguenza.[25] L'uso di concetti psico-sociale astratti non esclude ovviamente che si possano specificare e chiarire le notevolissime differenze tra le competenze umane e quelle dei sistemi artificiali.[26]

1.1.3 Un concetto giuridico di IA

Nelle pagine precedenti si sono sviluppate alcune considerazioni sul concetto di IA, come inteso dai ricercatori che si occupano di questa materia. Quale esempio paradigmatico possiamo considerare la definizione di IA proposta da John McCarthy, uno dei pionieri di questa disciplina:

> [L'IA] è la scienza e l'ingegneria del fare macchine intelligenti, specialmente programmi intelligenti per computer. È connessa al compito simile di usare i computer per comprendere l'intelligenza umana, ma l'IA non ha la necessità di limitarsi a metodi che sono biologicamente osservabili.[27]

Ci possiamo però interrogare se questo concetto sia adeguato alla prospettiva del giurista, il quale deve dotarsi di concetti sufficientemente precisi, che consentano ai destinatari delle norme e a chi ne deve assicurare l'attuazione, di distinguere gli oggetti o fenomeni cui si applicano quei concetti da quelli cui gli stessi concetti non si applicano. Se non abbiamo un concetto condiviso di intelligenza, o comunque non è possibile stabilire in modo preciso che cosa sia intelligente e che cosa non lo sia, come possiamo distinguere i sistemi informatici "intelligenti" da quelli privi di intelligenza, al fine di applicare solo ai primi le norme sull'IA? Il problema giuridico è divenuto reale rispetto alla Proposta di Regolamento sull'IA (Legge sull'IA) recentemente presentata dalla Commissione Europea. Il Regolamento, al fine di delimitare il proprio ambito di applicazione, all'articolo 3, definisce un sistema di IA (sistema di IA) in questo modo:

> un software sviluppato con una o più delle tecniche e degli approcci elencati nell'allegato I, che può, per una determinata serie di obiettivi definiti dall'uomo, generare output quali contenuti, previsioni, raccomandazioni o decisioni che influenzano gli ambienti con cui interagiscono.

L'allegato I dello stesso Regolamento distingue tre tecnologie che caratterizzano l'IA (il cui impiego sembra anzi sufficiente per attribuire la natura di "sistema di IA" ai software basati su di esse):

> a) approcci di apprendimento automatico, compresi l'apprendimento supervisionato, l'apprendimento non supervisionato e l'apprendimento per rinforzo, con utilizzo di un'ampia gamma di metodi, tra cui l'apprendimento profondo (*deep learning*);

b) approcci basati sulla logica e approcci basati sulla conoscenza, compresi la rappresentazione della conoscenza, la programmazione induttiva (logica), le basi di conoscenze, i motori inferenziali e deduttivi, il ragionamento (simbolico) e i sistemi esperti;

c) approcci statistici, stima bayesiana, metodi di ricerca e ottimizzazione.

Una diversa definizione era stata fornita dallo High Level Expert Group on AI (AI HLEG, costituito dalla Commissione Europea) nel rapporto predisposto ai fine dell'elaborazione di una strategia europea sull'IA, che ha preceduto la Proposta di Regolamento:

> I sistemi di intelligenza artificiale (IA) sono sistemi software (e possibilmente hardware) sviluppati da esseri umani che, dato uno scopo complesso, operano nella dimensione fisica o digitale percependo il loro ambiente mediante l'acquisizione di dati, interpretando le strutture di dati strutturati e non strutturati raccolte, ragionando sulla conoscenza o elaborando l'informazione, derivata da questi dati e decidendo le migliori azioni da compiere per raggiungere i goal dati. I sistemi di IA possono usare regole simboliche o apprendere un modello numerico, e possono anche adattare il loro comportamento analizzando come l'ambiente sia influenzato dalle loro azioni precedenti.[28]

Questa definizione elenca molte funzioni importanti nei sistemi di IA. Bisogna peraltro ricordare che la maggior parte dei sistemi di IA compie solo una frazione delle attività elencate nella definizione, essendo dedicate esclusivamente a singole funzioni come le seguenti: riconoscimento di pattern (classificazione di oggetti all'interno di immagini, identificazione di persone in base a caratteristiche biometriche, analisi di attitudini e sentimenti, ecc.), traduzione (da un linguaggio all'altro), filtro di informazioni indesiderate (spam, violenza, pornografia, ecc.), selezione di informazioni (pubblicità o notizie mirate). Alcuni sistemi invece combinano diverse capacità, come i veicoli autonomi, che debbono essere in grado di identificare gli oggetti che incontrano, ma anche di pianificare il percorso da effettuare, e autogovernarsi nel viaggio verso la meta.

Lo High-Level Expert Group descrive l'IA come segue:

> Come disciplina scientifica, l'IA include numerosi approcci e tecniche, come l'apprendimento automatico (di cui l'apprendimento profondo e l'apprendimento per rinforzo sono esempi specifici), il ragionamento automatico (che include la pianificazione, la schedulazione, la rappresentazione della conoscenza e il ragionamento, la ricerca e l'ottimizzazione) e la robotica (che include il controllo, la percezione, i sensori e gli attuatori, così come l'integrazione di ogni altra tecnica in sistemi ciber-fisici).[29]

Questa pur ampia caratterizzazione delle ricerche di IA omette alcuni settori importanti, come la comprensione e generazione del linguaggio naturale (il linguaggio parlato dagli esseri umani), una funzione fondamentale nei sistemi che operano su dati testuali, o che interagiscono con le persone. È molto difficile cogliere l'IA mediante una definizione che sia al tempo stesso precisa ed esauriente, poiché l'IA non è un'unica disciplina scientifica e tecnologica, ma piuttosto una gamma disparata di metodi e tecniche applicate a un amplissimo e diversificato insieme di obiettivi scientifici, tecnologici, e industriali. Quindi, l'interpretazione giuridica del concetto non potrà che essere teleologica, così da abbracciare il più possibile tutti e soli i sistemi che presentano i rischi e le opportunità delle tipiche applicazioni intelligenti.

Questa prospettiva sembra peraltro essere stata adottata anche nel Regolamento appena citato. Infatti, le norme più significative del Regolamento si applicano solo ai sistemi che il Regolamento stesso classifica come sistemi ad alto rischio. Ciò che conta, ai fini dell'applicazione del regolamento, non è la qualifica di "sistema di IA", ma piuttosto il fatto che il sistema in questione rientri in una delle categorie di sistemi ad alto rischio (vedi Sezione 3.7.2).

1.2 L'IA nel contesto

Nella presente sezione si esamina il rapporto tra IA e altri temi, ad essa strettamente collegati: gli algoritmi, i *big data* (grandi masse di dati, o megadati), la robotica e l'intelligenza ambientale

1.2.1 Algoritmi

Il termine "algoritmo" è spesso usato per far riferimento, in modo preminente, se non esclusivo, alle applicazioni di IA, e lo ritroviamo in locuzioni come "processi decisionali algoritmici" (*algorithmic decision-making*), "governance algoritmica ", "costituzionalismo algoritmico", e così via. Infatti, è tanta oggi l'attenzione per le tematiche dell'IA, che questa viene spesso identificata con la dimensione algoritmica nel suo complesso, pur costituendone solo un aspetto.

È quindi importante ricordare che gli algoritmi, quali procedure suscettibili di applicazione automatica, hanno un campo di utilizzo che si estende al di là dei sistemi di IA, ricoprendo ogni sistema informatico. Essi possono essere molto semplici, come quello che specifica come ordinare liste di parole o come trovare il massimo comune divisore tra due numeri (il cosiddetto algoritmo di Euclide). Essi possono essere invece molto complessi, come gli algoritmi per la cifratura o la compressione di file digitali, il riconoscimento vocale, o la previsione nella finanza. Ovviamente, non tutti gli algoritmi riguardano l'IA, ma ogni sistema di IA, come ogni sistema informatico, comprende algoritmi, alcuni dei quali svolgono compiti che attengono direttamente a funzioni di IA.

Gli algoritmi dell'IA svolgono diverse funzioni epistemiche e pratiche (attinenti al ragionamento, alla percezione, alla classificazione, alla pianificazione, alla decisione, ecc.). Alcuni algoritmi si limitano ad applicare conoscenze preesistenti, altri realizzano forme di apprendimento, contribuendo a creare o modificare il modello su cui si basa il funzionamento del sistema di cui fanno parte. Per esempio, un sistema di IA per il commercio elettronico potrebbe limitarsi ad applicare regole predeterminate (per es. applicare sconti ai consumatori che soddisfanno certe condizioni) ma potrebbe anche imparare e usare correlazioni tra caratteristiche e attività degli utenti e loro preferenze (per raccomandare acquisti) e sviluppare e selezionare strategie efficaci per l'attività commerciale (per negoziare online, o ottimizzare la gestione finanziaria).

Benché di regola un sistema per IA comprenda molti algoritmi, dalla cui interazione risultata il funzionamento del sistema stesso, lo possiamo anche vedere come un singolo algoritmo complesso, che comprende gli algoritmi che svolgono funzioni specifiche, così come gli algoritmi che orchestrano le funzioni del sistema attivando gli algoritmi di più basso livello. Per esempio, un bot che risponda a quesiti in linguaggio naturale comprenderà una combinazione orchestrata di algoritmi per il riconoscimento vocale (dalle onde sonore emesse alle parole pronunciate), l'individuazione delle strutture sintattiche, il recupero della conoscenza rilevante, la generazione di risposte, ecc.

Come si vedrà nel seguito, nei sistemi capaci di apprendimento, la componente più importante non è il modello algoritmico costruito (in parte) dal sistema per eseguire i compiti ad esso affidati. Il nucleo del sistema è piuttosto l'algoritmo per l'apprendimento, che genera o sviluppa —sulla base dei dati cui il sistema ha accesso— il modello algoritmico, affinché quest'ultimo possa meglio svolgere quei compiti. Per esempio, in un sistema classificatore che riconosce immagini attraverso una rete neurale, l'elemento cruciale non è la rete neurale, ma piuttosto l'algoritmo per l'apprendimento (l'algoritmo "discente" - *learning*) che modifica la struttura della rete neurale (il modello algoritmico) cambiando i pesi delle sue connessioni, in modo che essa migliori le proprie prestazioni nel classificare gli oggetti di interesse (e.g., animali, suoni, volti, attitudini, sentimenti, ecc.).

La tesi secondo cui un sistema di AI consiste di algoritmi (e di dati) sembra essere messa in dubbio dal fatto che il comportamento dei sistemi di AI, e in particolare quelli che usano metodi per l'apprendimento automatico, non sembrano operare secondo istruzioni predeterminate; al contrario adattandosi a nuovi contesti e informazioni, sviluppano nuovi comportamenti, non previsti dal creatore del sistema. Questa prospettiva potrebbe essere suggerita da una recente sentenza del Consiglio di Stato (Sezione Terza, sentenza 25 novembre 2021 n. 7891), in cui si afferma quanto segue:

> la nozione comune e generale di algoritmo riporta alla mente una sequenza finita di istruzioni, ben definite e non ambigue, così da poter essere eseguite meccanicamente e tali da produrre un determinato risultato; nondimeno se la nozione è applicata a sistemi tecnologici, è ineludibilmente collega-

> ta al concetto di automazione ossia a sistemi di azione e controllo idonei a ridurre l'intervento umano, di cui il grado e la frequenza dipendono dalla complessità e dall'accuratezza dell'algoritmo che la macchina è chiamata a processare. Cosa diversa è l'intelligenza artificiale, in cui l'algoritmo contempla meccanismi di machine learning e crea un sistema che non si limita solo ad applicare le regole software e i parametri preimpostati (come fa invece l'algoritmo tradizionale) ma, al contrario, elabora costantemente nuovi criteri di inferenza tra dati e assume decisioni efficienti sulla base di tali elaborazioni, secondo un processo di apprendimento automatico.

La definizione appena riportata, se intesa a tracciare una distinzione tra l'ambito degli algoritmi in senso proprio, e quello dell'IA, solleva due problemi.

In primo luogo, non necessariamente un sistema di IA usa metodi di apprendimento automatico; il concetto di AI, come comunemente inteso, include sistemi che compiono inferenze inferenza sulla base di rappresentazioni della conoscenza fornite dall'uomo (Sezione 2.1). Ciò resta vero, anche se oggi sono soprattuto i sistemi per l'apprendimento automatico a sollevare interesse, aspettative e preoccupazioni.

In un secondo luogo, come indica la stessa sentenza citata, anche i sistemi di AI si basano su algoritmi, seppure "non tradizionali" e in particolare algoritmi per l'inferenza e l'apprendimento. Nel caso dei sistemi basati sull'apprendimento automatico (vedi Sezione 2.2), tanto il programma informatico mediante il quale il sistema apprende (l'algoritmo che apprende), quanto il modello mediante il quali il sistema risponde agli input (l'algoritmo appreso, per es. la rete neurale addestrata) possono essere visti come algoritmi, in un senso ampio. L'algoritmo appreso, peraltro, può essere oscuro per noi, nel senso che non riusciamo a capire la funzione, rispetto al risultato finale, dei singoli passi attraverso cui l'algoritmo stesso si sviluppa (Vedi Sezione 2.2.5). In questo senso, possiamo forse dire che i modelli oscuri sono algoritmi per la macchina; non sono fatti per noi (la stessa qualifica può applicarsi peraltro ad ogni codice oggetto di un programma informatico, cioè al risultato della compilazione del codice sorgente scritto dal programmatore nel linguaggio binario del computer, al fine della sua esecuzione).

Si può parlare di algoritmo, in senso ampio, anche con riferimento ai software che sperimentano variazioni casuali, per approntare nuove soluzioni da verificare con l'esperienza. È questo il caso dei sistemi che si basano sull'apprendimento con rinforzo. Questi sistemi, oltre a riprodurre le combinazioni di azioni che hanno già avuto maggior successo, scelgono a caso nuove azioni, per sperimentarne l'efficacia. Similmente, gli algoritmi genetici generano nuove soluzioni mediante la ricombinazione, con variazioni (mutazioni) casuali degli soluzioni preesistenti, anch'essi privilegiando le soluzioni che hanno avuto maggiore successo.

In conclusione, per cogliere la relazione tra algoritmi e IA, sembra preferibile allargare il concetto di algoritmo e distinguere, al suo interno gli algoritmi di IA— in base alle tecnologie che li caratterizzano e alle funzioni che svolgono— anziché restringere

tale concetto ai soli "algoritmi tradizionali" escludendo dal suo ambito i programmi per l'IA.

1.2.2 Big Data

Il temine Big Data (grandi masse di dati) viene applicato a enormi raccolte di dati che è difficile trattare usando le tecnologie informatiche solitamente impiegate per la gestione di dati digitali (le basi di dati o i sistemi documentali). Tali masse di dati sono caratterizzate delle cosiddette tre V: enorme Volume, alta Velocità (nel cambiamento) e grande Varietà. Altre caratteristiche talvolta associate ai big data sono la bassa Veracità (alta probabilità che alcune informazioni siano inaccurate) e l'alto Valore (l'utilità, correlata all'ampiezza della massa, ricavabile dai dati attraverso tecniche di analisi).

I dati che compongono i Big Data possono essere creati dagli umani, ma più spesso sono raccolti automaticamente, da dispositivi che raccolgono informazioni dal mondo fisico (e.g., telecamere nelle strade, sensori di dati ambientali, dispositivi per esami medici, sensori applicati a prodotti nell'industria o nel commercio, ecc.) o che mediano attività economiche e sociali collegando gli individui facendoli partecipare a organizzazioni socio-tecniche (transazioni del commercio elettronico e del governo elettronico, tracciamento delle attività su Internet, ecc.).

Da una prospettiva sociale e giuridica ciò che è maggiormente rilevante rispetto alle grandi masse di dati, cioè che rende "grande" una massa di dati, è una caratteristica funzionale: la possibilità di usare quei dati per finalità di "analitica" (*analytics*), cioè per scoprire correlazioni e fare predizioni. A tal fine, come vedremo nel seguito, sempre più spesso si utilizzano tecnologie di IA basate sull'apprendimento automatico, che consentono di estrarre modelli predittivi da grandi insiemi di dati. I Big Data possono riguardare il mondo fisico e digitale non-umano (dati astronomici, ambientali, biologici, industriali, tecnologici), così come gli umani e le loro relazioni (dati sulle reti sociali, la salute, la finanza, i trasporti, ecc.).

1.2.3 Robotica

L'IA costituisce il nucleo della robotica, la disciplina che si occupa di costituire agenti fisici che compiano compiti che richiedono la manipolazione del mondo fisico. Secondo la definizione dell'High Level Expert Group, la robotica può essere definita come IA in azione nel mondo fisico (anche chiamata IA "incorporata", *embodied*).

> Un robot è una macchina fisica che deve affrontare la dinamica, le incertezze e le complessità del mondo fisico. La percezione, il ragionamento, l'azione, l'apprendimento, come le capacità di interazione con altri sistemi sono solitamente integrati nell'architettura di controllo del sistema robotico. In aggiunta all'IA, altre discipline giocano un ruolo nella progettazione e nel

> funzionamento del robot, come l'ingegneria meccanica e la teoria del controllo. Esempi di robot includono manipolatori robotici, veicoli autonomi (per esempio, automobili, droni, taxi volanti), robot umanoidi, aspirapolvere robotici, ecc.[30]

Peraltro, il termine "robot", o semplicemente "bot", accompagnato dall'aggettivo digitale o software è spesso usato per far riferimento ad agenti digitali che interagiscono in modo attivo con il mondo digitale, per esempio, effettuando transazioni commerciali (come vendere e acquistare titoli sul mercato delle azioni e delle obbligazioni). Ai nostri fini basta sottolineare come l'IA costituisca l'aspetto preminente sia dei robot fisici, sia dei bot digitali, pur accompagnandosi con altre discipline. Gli aspetti fondamentali dei robot fisici sono ben colti dalla seguente definizione:

> una macchina, situata nel mondo, che sente, pensa e agisce. Pertanto, un robot deve avere sensori, capacità di elaborazione che emula alcuni aspetti della cognizione, e attuatori. I sensori sono necessari per ottenere informazione dall'ambiente. I comportamenti reattivi (come il riflesso da stiramento negli umani) non richiedono alcuna abilità cognitiva profonda, ma l'intelligenza a bordo è necessaria se il robot deve svolgere compiti significativi autonomamente, e l'attuazione è necessaria per consentire al robot di esercitare forze sul suo ambiente. In generale, queste forze risulteranno nel movimento dell'intero robot o di uno dei suoi elementi.[31]

Un robot può utilizzare diversi sensori: video-camere o laser per sondare l'ambiente, dispositivi come il GPS per determinare la propria ubicazione, giroscopi o acceleratori per misurare il proprio movimento. Gli effettori o attuatori possono essere avere varie forme e funzioni: gambe, ruote, articolazioni, pinze, ecc.

I robot possono essere classificati in tre categorie principali: robot manipolatori, robot mobili, e manipolatori mobili.[32]

I robot manipolatori sono ancorati fisicamente al proprio posto di lavoro, e si presentano tipicamente nella forma di bracci meccanici mobili. La maggior parte di essi (milioni di unità) vengono impiegati nelle catene di montaggio. Alcuni manipolatori sono usati in ambito sanitario, ad esempio, per aiutare i chirurghi nell'effettuazione di operazioni che richiedono assoluta precisione.

I robot mobili si spostano nell'ambiente, con vari strumenti di locomozione (gambe, ruote, eliche, etc.). Molti robot mobili sono usati in ambienti ristretti, dove svolgono funzioni limitate, ad esempio, la pulizia dei pavimenti, il taglio dell'erba, ecc. Altri robot mobili sono invece dotati della capacità di affrontare missioni di ampio raggio, anche in spazi condivisi con gli esseri umani. Negli ultimi anni abbiamo assistito alla progressiva robotizzazione delle automobili, che si sono dotate di sensori per riconoscere ostacoli, e della capacità di effettuare autonomamente operazioni di guida, come il parcheggio e il

mantenimento della direzione sulla strada. Stanno ormai entrando in funzione automobili senza pilota, capaci di condurre autonomamente il proprio carico (di persone e cose) a destinazione senza interventi umani (come nel caso della Google-car, priva di volante). I *rover*, veicoli di superficie usati nelle esplorazioni extraterrestri (sulla Luna o su Marte), possono muoversi per lungo tempo (anche per più anni) con autonomia, affrontando territori sconosciuti. Sono già oggi numerosi i veicoli aerei senza pilota (*Unmanned Air Vehicles* - UAV) usati nella sorveglianza, nei lavori agricoli, o in operazioni militari. Veicoli robotici sottomarini (*Autonomous Underwater Vehicles* - AUV) sono impiegati per esplorare le profondità marine.

I robot manipolatori mobili uniscono manipolazione e movimento. Si pensi ad esempio ai dispositivi robotici usati per disinnescare bombe. A questa categoria appartengono i robot antropomorfi o umanoidi, dotati di un corpo dotato di arti e testa, che mima la struttura fisica degli umani.

Appartengono alla robotica, ampiamente intesa, anche le protesi con capacità cognitiva, destinate a sostituire parti del corpo umano, come gli arti, o l'apparato per l'udito o la visione. Infine, esistono robot multicorpo (*multibody*), che consistono di gruppi o sciami di dispositivi separati che si auto-coordinano.

1.2.4 Intelligenza ambientale

Tra i profili emergenti dell'intelligenza artificiale va ricordata l'*intelligenza ambientale* (*ambient intelligence*). Si tratta dell'inserimento nell'ambiente fisico di dispositivi automatici dotati della capacità di elaborare informazioni e, anzi, di esibire comportamenti intelligenti. Tali dispositivi possono assorbire informazioni sia dall'ambiente fisico sia dalla rete informatica, e di operare in entrambi gli ambiti. Essi sono destinati a inserirsi nell'ambiente in modo ubiquo e invisibile, governando macchine di vario genere, e facendo sì che l'ambiente stesso si adatti automaticamente alle esigenze dell'uomo. Possono comunicare tra loro e con altri dispositivi digitali, ma anche percepire i mutamenti dell'ambiente e reagire agli stessi.

Si immagini una casa nella quale la porta si apra automaticamente ogni qualvolta la telecamera riconosca uno degli abitanti, la cucina si attivi per riscaldare la cena al momento opportuno, il frigorifero proceda automaticamente a ordinare i prodotti mancanti, la combinazione ottimale di umidità e temperatura sia mantenuta costante (tenendo conto, altresì, del costo del riscaldamento), l'armadietto sanitario si occupi di indicarci le medicine da prendere secondo il piano stabilito dal medico, l'impianto stereo proponga brani musicali, tenendo conto dei nostri gusti e addirittura del nostro stato d'animo, ecc. Si immagini altresì che sia possibile dialogare con la casa stessa e con i vari dispositivi che ne fanno parte (per esempio, chiedendo al forno di attivarsi per cucinare l'arrosto e allo stereo di proporci un brano di Brahms o dei Maneskin). Ecco come questo scenario

è presentato da Philips, la nota casa produttrice di elettronica di consumo, che tra i primi usò il termine "intelligenza ambientale":

> Questa è la nostra visione dell'"intelligenza ambientale': persone che vivono facilmente (comodamente) in un ambiente digitale nel quale i dispositivi elettronici sono sensibili ai bisogni delle persone, personalizzati secondo le loro esigenze, anticipatori rispetto ai loro comportamenti e reattivi alla loro presenza.[33]

L'Unione Europea ha fatto propria la prospettiva dell'intelligenza ambientale, dedicando a essa un ampio spazio nell'ambito dei propri progetti di ricerca. Si tratta di una prospettiva che, accanto agli aspetti positivi, manifesta diversi profili problematici, rispetto ai quali si rendono necessarie garanzie giuridiche, profili che vanno dalla tutela dei dati personali, alle responsabilità per i danni causati dalle apparecchiature intelligenti, alla protezione dell'interessato rispetto alle possibilità di sfruttamento e manipolazione che possono essere realizzate controllando le macchine intelligenti, e tramite esse il comportamento dei loro utilizzatori (ad esempio, rispetto a scelte di consumo o di acquisto), e così via.

Lo sviluppo delle scienze fisiche e delle tecnologie per l'elaborazione della materia ci aveva consegnato un mondo materiale "disincantato",[34] nel quale ci rapportavamo agli oggetti assumendo che il loro comportamento sia esclusivamente e pienamente accessibile secondo le leggi fisiche, obbedendo alle quali gli oggetti stessi svolgono la funzione loro assegnata. Oggi lo sviluppo dell'intelligenza artificiale ambientale sembra ricreare un mondo "incantato" nel quale ci accostiamo agli oggetti in modo analogo a quello con cui interagiamo con le persone, riproducendo quindi schemi del pensiero animistico, proprie del mondo del mito e della fiaba. In un vicino futuro —per taluni aspetti già presente oggi, per esempio nell'uso di assistenti personali intelligenti, quali Alexa e Google home— potremo capire il funzionamento degli oggetti più comuni (dalla cucina, al frigorifero, all'automobile) solo assumendo che l'oggetto in questione persegua certi obiettivi (attinenti alle nostre esigenze, così come l'oggetto stesso riesce a coglierle) scegliendo i mezzi che ritiene più adatti al loro conseguimento. Interagiremo con gli oggetti intelligenti adottando uno stile comunicativo, cioè interrogandoli sulle iniziative che stanno adottando, e indicando a essi i risultati da realizzare o i modi per raggiungerli (così come faremmo con un collaboratore domestico).[35] Immaginiamo per esempio di rientrare in casa e di chiedere alla cucina che cosa possa prepararci per cena (dopo aver interrogato il frigorifero sulle sue disponibilità), che questa si informi sulle nostre preferenze, e conseguentemente suggerisca particolari menu, ci indichi i tempi di cottura (o quelli necessari per approvvigionarsi di materie prime non disponibili in casa), e così via. Il mondo incantato dell'intelligenza ambientale può però diventare un mondo stregato, nel quale gli oggetti (o chi li governa) ci manipolano, ci sfruttano, operano a

nostro danno. Di qui l'importanza che, anche nel campo dell'intelligenza ambientale, alla tecnologia si affianchino l'etica e il diritto.

1.3 I limiti dell'IA

Lo sviluppo delle tecnologie di IA è stato accompagnato da un intenso dibattito sui limiti di tali tecnologie, e sulle prospettive dei loro sviluppi futuri. Nelle pagine seguenti si esaminerà questo dibattito da tre diverse prospettive. Dapprima si introdurrà la distinzione tra IA con competenze onnicomprensive (intelligenza generale artificiale) o invece specifiche (intelligenza speciale artificiale). Poi si passerà alla parallela distinzione tra IA forte (che riproduce l'intelligenza umana) e debole (che si limita a simularla). Infine, si esaminerà il tema della comprensione dei significati (la semantica) da parte di sistemi artificiali.

1.3.1 Intelligenza specifica e intelligenza generale

In linea di principio, le ricerche di IA possono condurre a due risultati distinti, seppure connessi: l'intelligenza specifica artificiale (*artificial special intelligence*), e l'intelligenza generale artificiale (*artificial general intelligence*).

All'intelligenza specifica artificiale (detta anche "ristretta", *narrow*) appartengono tutte le applicazioni di IA oggi disponibili: si tratta di sistemi capaci di ottenere risultati utili in attività che richiedono intelligenza, con prestazioni, in alcuni casi, di livello umano o anche sovrumano. Per esempio, nel riconoscimento di immagini o di volti, l'IA ha già raggiunto prestazioni paragonabili a quelle di un umano esperto; nel gioco degli scacchi, è invece capace di prestazioni sovrumane, superiori a quelle dei migliori giocatori. L'IA specifica può essere impiegata con profitto anche in compiti nei quali i sistemi informatici sono ancora inferiori agli umani, ma in cui il loro impiego risulta conveniente per ragioni di costo e rapidità. Ad esempio, anche se le traduzioni artificiali hanno una qualità inferiore rispetto a quelle prodotte da esperti umani, esse trovano applicazione in diversi contesti di utilizzo. In molti casi, il risultato migliore può ottenersi unendo intelligenza artificiale e umana. Per esempio, nell'identificazione di contenuti online vietati o comunque pregiudizievoli, l'analisi effettuata da sistemi di IA, intesa a rilevare i contenuti potenzialmente da rimuovere, può essere efficacemente combinata con la valutazione umana dei casi identificati dalla macchina.[36]

Mentre le applicazioni di intelligenza specifica artificiale sono limitate agli obiettivi ristretti per i quali sono state sviluppate, un'intelligenza generale artificiale dovrebbe possedere la maggior parte delle abilità cognitive umane, al livello umano, o anche a un livello sovrumano. Illustri studiosi e tecnologi hanno espresso opinioni assai diverse sia sulla probabilità che l'intelligenza generale artificiale sarà realizzata, sia sulle prospettive che essa aprirebbe.

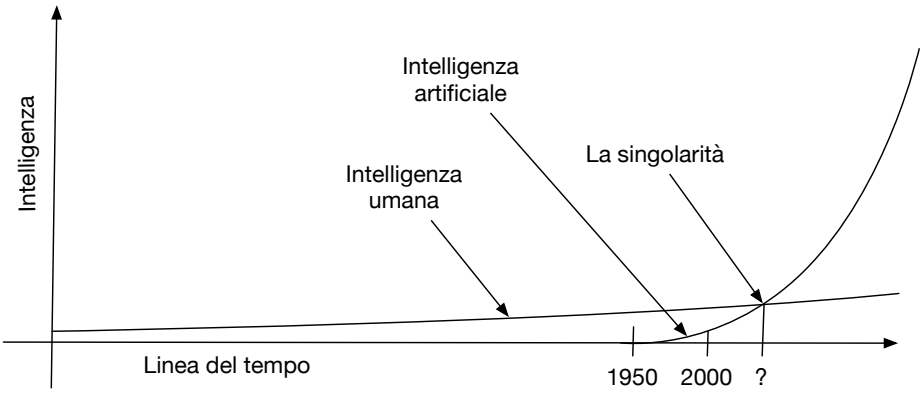

Figura 1.2: *L'evoluzione dell'intelligenza*

Innanzitutto, vi è chi esclude che l'intelligenza generale artificiale sia una prospettiva realistica, alla luce delle tecnologie oggi disponibili e dei loro possibili sviluppi. Quindi non c'è ragione né di preoccuparsi né di entusiasmarsi per essa.[37]

Altri, invece, ritengono che la realizzazione futura di sistemi dotati di intelligenza generale artificiale sia una probabilità concreta. Benché gli scienziati siano in disaccordo su quando l'intelligenza generale verrà ad esistere, sembra che la maggior parte di essi ritenga che questo risultato potrà realizzarsi entro il secolo presente.[38]

A questa prospettiva alcuni guardano con preoccupazione. Un sistema dotato di intelligenza generale artificiale potrebbe migliorare sé stesso e presto superare l'intelligenza umana.[39] A questo punto, grazie alla sua intelligenza sovrumana potrebbe acquisire capacità non più controllabili . Rispetto a tale IA ci troveremmo in una condizione di inferiorità simile a quella degli animali rispetto a noi.[40] Alcuni importanti scienziati e tecnologi (come Steven Hawking, Elon Musk, e Bill Gates) hanno infatti richiamato la necessità di anticipare questo rischio esistenziale per l'umanità (un rischio che riguarda la stessa esistenza della nostra specie). A loro parere, sarebbe necessario individuare fin d'ora le misure per prevenire la nascita dell'intelligenza generale artificiale, o per dirigerla verso risultati benefici all'umanità, assicurando che tale intelligenza si allinei ai valori umani, e più in generale sviluppi attitudini benevole.

Altri, invece guardano favorevolmente allo sviluppo di un'IA che raggiunga e poi superi i limiti dell'intelligenza umana. La realizzazione dell'intelligenza generale artificiale potrebbe rappresentare il momento magico, la "singolarità", a partire dalla quale si scatena uno sviluppo accelerato della scienza e della tecnologia, che potrebbe condurre non solo a risolvere i problemi odierni dell'umanità, ma anche a superare i limiti biologici dell'esistenza umana (la malattia, l'invecchiamento, ecc.) e a distribuire l'intelligenza (umana e artificiale) nel cosmo (Figura 1.2).[41]

Ai nostri fini, non occorre prendere posizioni rispetto alle tesi appena enunciate. L'intelligenza generale artificiale potrà realizzarsi, in ogni caso, solo tra qualche decennio (secondo le previsioni più ottimistiche). Pertanto, il dibattito su di essa oggi riguarda l'anticipazione e la valutazione di possibili scenari futuri, ma non tocca ancora la politica e il diritto. Solo sulla base di esperienze più ampie con sistemi avanzati di IA di applicazione progressivamente più generale, sarà possibile comprendere l'ampiezza e la prossimità dei rischi per l'umanità e individuare i modi migliori per affrontarli.

1.3.2 IA forte e IA debole

Alla distinzione appena tracciata –tra intelligenza specializzata artificiale e intelligenza generale artificiale– si sovrappone un'altra distinzione, quella tra IA *forte* (*strong artificial intelligence*) e IA *debole* (*weak artificial intelligence*).

Secondo la caratterizzazione che ne dà John Searle, illustre studioso del linguaggio e della mente, l'IA forte muove dall'assunto che anche i calcolatori siano capaci di stati cognitivi e di pensiero (nel modo in cui ne è dotato un essere umano) e conseguentemente si propone di costruire menti artificiali.[42] Per l'IA forte "il calcolatore appropriatamente programmato è realmente una mente, si può cioè dire letteralmente che i calcolatori dotati dei programmi giusti capiscono e hanno stati cognitivi".[43] L'IA debole invece si propone di realizzare sistemi artificiali capaci di svolgere compiti complessi, sistemi che possono mimare (simulare) aspetti dei processi cognitivi umani, ma che non possono riprodurre quegli stessi processi. I sistemi di IA oggi disponibili –e più in generale quelli basati sui computer– non sarebbero in grado di pensare, non possederebbero una mente.

Il dibattito circa la possibilità di sviluppare, mediante elaboratori elettronici, forme di IA forte, cioè vere menti artificiali, può essere fatto risalire al fondamentale contributo di Alan Turing, che già nel 1936 si interrogava non solo sulla possibilità di sviluppare macchine intelligenti, ma anche su come verificare quando e in quale misura questo risultato potesse considerarsi raggiunto. A tale fine egli proponeva un test ispirato a un gioco di società, il "gioco dell'imitazione", nel quale una persona interroga due interlocutori di sesso diverso, al fine di determinare chi di questi sia l'uomo e chi la donna (senza avere contatto diretto con gli stessi). Nel gioco di Turing lo scopo dell'interrogante è invece quello di distinguere l'interlocutore umano e l'interlocutore elettronico, il calcolatore (Figura 1.3).[44] Si avrà la prova che l'IA è stata realizzata quando un sistema informatico riuscirà a ingannare l'interrogante, facendogli credere di essere una persona (quando l'interrogante, nel gioco dell'imitazione, attribuirà l'identità umana con la stessa probabilità all'interlocutore umano e a quello elettronico). Ecco come il test è presentato da Turing stesso:

> Sostituirò la domanda ['Possono le macchine pensare?'] con un'altra, che è strettamente connessa con la prima e può essere espressa con parole relativamente non ambigue. La nuova forma del problema può essere descritta

> nei termini di un gioco che possiamo chiamare 'il gioco dell'imitazione'. È un gioco con tre persone, un uomo (*A*), una donna (*B*), e un interrogante (*C*) che possono essere dell'uno o dell'altro sesso. L'interrogante sta in una stanza separata dalle altre due. Lo scopo del gioco per l'interrogante è determinare quale degli altri due sia l'uomo e quale la donna. Egli conosce i due mediante le etichette *X* e *Y*, e alla fine del gioco egli dice '*X* è *A* e *Y* è *B*' oppure '*X* è *B* e *Y* è *A*' [...] Per far sì che i toni di voce non aiutino l'interrogante, le risposte dovrebbero essere scritte, o meglio, dattiloscritte. La soluzione ideale è avere una telescrivente che comunica tra le due stanze. [...] Lo scopo del gioco per il terzo giocatore (*B*, cioè la donna) è aiutare l'interrogante. A tal fine la migliore strategia per lei consiste nel dare risposte veritiere. Ella può aggiungere cose del tipo 'Io sono la donna, non dar retta a lui!' alle proprie risposte, ma ciò non serve a nulla in quanto anche l'uomo può fare simili commenti. Ci poniamo ora la domanda 'che accadrebbe se una macchina prendesse il posto di *A* nel gioco?' L'interrogante deciderebbe erroneamente con la stessa frequenza quando il gioco si svolge in questo modo rispetto a quando il gioco riguarda un uomo e una donna? Questa domanda sostituisce la domanda originaria 'Possono le macchine pensare?'.[45]

Nessun sistema ha ancora superato il test di Turing, e anzi nessun sistema si è avvicinato a questo risultato.[46] Se ne può trarre una conclusione rassicurante: l'IA è ancora lontana dal raggiungere l'intelligenza umana, nel campo della comunicazione linguistica non ristretta a temi e formulazioni specifiche.

1.3.3 L'IA e la comprensione dei significati

Il test di Turing solleva un importante problema teorico, che ci possiamo porre in astratto, indipendentemente dalla possibilità concreta di realizzare oggi, o nel prossimo futuro, un sistema che superi il test. Ci possiamo cioè chiedere se un sistema che, in ipotesi, riuscisse a superare il test sarebbe una vera IA, o invece sarebbe solo un mero "idiota sapiente" (che finge di essere intelligente senza esserlo, simula una mente senza possederla). Infatti, il test di Turing è puramente comportamentale: per superarlo è sufficiente che la macchina si comporti come un essere umano, non è necessario che esso abbia veramente una mente, dei pensieri.

L'argomento della stanza cinese Vi è stato pertanto chi, come John Searle, ha affermato l'impossibilità teorica di realizzare sistemi informatici capaci di attività mentale (di pensiero in senso proprio), quali che siano le prestazioni offerte dagli stessi (anche se tali prestazioni comportino il superamento del test di Turing).[47]

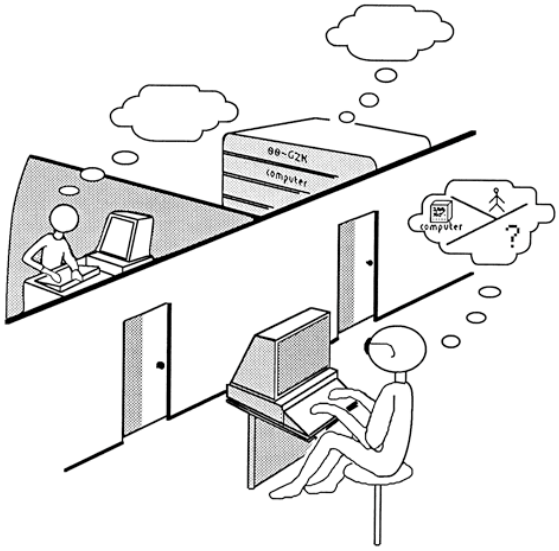

Figura 1.3: *Il test di Turing*

Per criticare le pretese dell'IA forte Searle ha sviluppato un celebre esperimento mentale, il cosiddetto "argomento della stanza cinese". Egli ci invita a immaginare che una persona capace di parlare solo la lingua inglese (non il cinese) sia chiusa in una stanza dotata di una fenditura verso l'esterno. La stanza contiene dei fogli di carta e un enorme volume. Il volume è un manuale di istruzioni che specifica come, una volta ricevuto un input consistente in una sequenza di caratteri cinesi, si debba produrre un output consistente in un'altra sequenza degli stessi caratteri. Le regole collegano a ogni input l'output appropriato (la risposta che giudicheremmo appropriata in una conversazione tra persone che conoscono il cinese), ma esse sono formali, nel senso che fanno riferimento solo alla struttura sintattica della comunicazione prescindendo dal significato dei messaggi: per ogni sequenza di simboli di input tali regole indicano la sequenza di caratteri di output corrispondente e possono quindi essere applicate senza conoscere il significato delle parole formate con quei caratteri.

Ecco come funziona la stanza cinese (come opera la persona al suo interno). Dalla fenditura viene immesso un foglio di carta che riporta caratteri cinesi (incomprensibili a chi non conosca questa lingua). Seguendo esattamente le istruzioni del manuale, la persona nella stanza scrive su un foglio bianco la risposta (i caratteri cinesi) che le regole del manuale collegano ai caratteri indicati nei fogli di input, e spinge il foglio attraverso la fenditura.

Le risposte che escono dalla camera cinese, in ipotesi, sono indistinguibili da quelle che potrebbero essere fornite da una persona capace di parlare il cinese. Di conseguenza, (tralasciando il problema dei tempi di risposta) la camera cinese riuscirebbe a superare il test di Turing (l'interrogante non sarebbe in grado di stabilire se sta dialogando con la stanza o con un cinese). Searle sostiene però che la persona all'interno della stanza cinese ha solo manipolato simboli a lei incomprensibili: anche se quella persona si comporta come un parlante cinese, le è precluso l'accesso al significato dei simboli che ricopia. Ora, uscendo dalla metafora, la persona nella stanza cinese è il calcolatore, guidato da un software (il manuale di istruzioni). Pertanto, Searle conclude che anche un calcolatore capace di conversare come un essere umano non è capace di pensieri, non ha una mente, si limita alla cieca manipolazione di simboli.

Numerosi studiosi di IA hanno raccolto la sfida di Searle, e hanno contestato il suo argomento. Alcuni hanno obiettato che, anche se l'uomo all'interno della stanza cinese non capisce il cinese, l'intero sistema (la stanza, la persona, e il manuale di regole) è in grado di capire il cinese, possedendo la capacità di rispondere a input in quella lingua producendo output appropriati nella stessa. L'errore di Searle consisterebbe nell'astrarre da tale sistema una sola componente (l'elemento che effettua le trasformazioni simboliche, corrispondente alla persona nella stanza). Sarebbe come chiedersi se la funzione mentale umana consistente nell'effettuazione di operazioni di ragionamento sia sufficiente a comprendere una lingua, una volta separata dalla memoria, dalle conoscenze, dai sensi, ecc. Un'ulteriore critica attiene al fatto che tanto la mente umana quanto il calcolatore elaborano informazioni con velocità ed efficienza enormemente superiori rispetto all'operatore della stanza cinese. L'impressione che la stanza non capisca il cinese si basa su questa circostanza, non applicabile all'elaborazione informatica. Pertanto, non sarebbe giustificato estendere ad ogni sistema informatico le conclusioni concernenti la stanza cinese.

Altri hanno osservato che la conclusione che la persona nella stanza (o la stanza nel suo insieme) non comprenda il cinese è determinata dal fatto che la comprensione di un linguaggio richiede la capacità di connettere le parole ai loro referenti reali, il che presuppone l'esperienza degli oggetti di cui parla il linguaggio (o almeno di alcuni di essi). Questo limite di un calcolatore isolato non si applicherebbe però ai sistemi automatici che uniscano capacità percettive (e possibilmente motorie) a quelle attinenti all'elaborazione e alla registrazione delle informazioni. Di conseguenza, i limiti della stanza cinese non sono limiti dell'IA: essi possono essere superati estendendo il sistema con dispositivi capaci di movimento e dotati di appropriati sensori.

Altri infine hanno osservato che l'intelligenza è un fenomeno emergente da comportamenti meccanici (non intelligenti) anche nel caso del cervello umano: anche l'intelligenza umana nasce da processi non intelligenti, le operazioni "meccaniche" (i processi chimici e fisici) che hanno luogo all'interno dei singoli neuroni del cervello umano e nei contatti (sinapsi) tra gli stessi. Allo stesso modo le operazioni mecca-

niche che avvengono all'interno del calcolatore programmato potrebbero dare origine all'intelligenza.[48]

La fondazione extralinguistica del linguaggio Bisogna distinguere due questioni circa l'esperimento mentale della "stanza cinese".

Una prima questione riguarda i sistemi informatici oggi disponibili. Ci possiamo chiedere se questi sistemi siano in grado di comprendere il linguaggio nel modo in cui lo comprendono gli umani, di avere consapevolezza dei significati delle parole e della connessione tra parole e mondo. La risposta al riguardo sembra essere negativa. Per ora dai sistemi informatici non hanno accesso, se non in misura molto limitata alla dimensione della semantica. Essi si limitano ad un "pensiero cieco" che consiste nell'elaborazione di numeri, o altri simboli, senza avere consapevolezza dei relativi significati.[49] Si tratta di un'elaborazione simile al nostro ragionamento quando eseguiamo rapidamente dei calcoli matematici, applicando regole prestabilite, senza riflettere sul significato delle regole e dei processi attivati dalla loro esecuzione.

Per esempio, un sistema per la traduzione automatica (come Google Translate) non conosce il significato dei testi nel linguaggio sorgente, né nel linguaggio obiettivo, né ha alcuna cognizione dei referenti dei termini dei due linguaggi nel mondo fisico o sociale. Il sistema applica in modo cieco le correlazioni statistiche —apprese da esempi di traduzioni passate— tra combinazioni di parole nell'uno e nell'altro linguaggio (possibilmente con l'aiuto di ontologie e altre risorse linguistiche che specificano le correlazioni logiche tra le parole). Si è pertanto osservato che il successo nella traduzione automatica non mostra che le macchine oggi comprendano il linguaggio umano, ma piuttosto che è possibile effettuare traduzioni aggirando o eludendo "l'atto della comprensione del linguaggio": il sistema si comporta come se comprendesse il linguaggio, mentre invece opera senza averne consapevolezza.[50] Analoghe considerazioni si applicano ai sistemi che generano testi che sembrano scritti da umani, come GPT-3 (*Generative Pre-trained Transformer*-3).[51] Queste valutazioni riguardano anche i sistemi di IA usati in ambito giuridico, come quelli dedicati alla cosiddetta giustizia predittiva, cioè alla previsione dell'esito di un caso sulla base di una descrizione del caso stesso o degli documenti presentati dalle parti (vedi Sezione 4.3). Tali sistemi non operano sulla base di una comprensione dei fatti del caso e delle norme da applicazione, ma su inferenze o correlazioni di tipo sintattico).

Questo aspetto (e limite fondamentale) dell'IA riguarda la fondazione (*grounding*) del significato. Nella comunicazione umana il linguaggio non si limita a combinare parole, esso fa riferimento al mondo fisico e sociale. Per capire pienamente che cosa significhi un enunciato non basta collegare tra loro le parole che lo compongono, ed esaminare i rapporti tra quelle parole e altre parole (per esempio, usando un dizionario). Bisogna invece collegare le parole alle cose cui si riferiscono, e gli enunciati alle situazioni che descrivono, costituiscono o prescrivono, nel contesto in cui quelle parole

vengono usate. Anche le comunicazioni più semplici, come gli ordini "Chiudi la finestra!", "Per favore, patente e libretto di circolazione!" possono essere comprese solo da chi sia in grado di capire a quali cose e azioni il parlante faccia riferimento.[52]

La comprensione piena del linguaggio presuppone infatti l'esperienza del mondo. Per renderci conto di questo fatto, immaginiamo che su un lontano pianeta si sia in grado di cogliere le trasmissioni radio che si svolgono sulla terra e che questa sia l'unica informazione accessibile sul nostro pianeta. Gli abitanti di quel pianeta, se dotati di elevante competenze statistiche, potrebbero determinare quali parole vengano usate più frequentemente, quali tendono ad essere compresenti, a quali condizioni, ecc. Ma gli extraterrestri non sarebbero in grado di comprendere il significato delle nostre comunicazioni. Anche se essi avessero accesso ad un dizionario terrestre —che indichi come certe parole siano riconducibili a combinazioni di altre parole del linguaggio umano— non avrebbero consapevolezza degli oggetti e situazioni cui le parole si riferiscono e, quindi, non avrebbero "rappresentazioni mentali" che colgano i significati corrispondenti.

Tuttavia, i limiti dell'IA di oggi non debbono indurci ad escludere in modo definitivo realizzazioni future che includano la progressiva, seppur parziale (nei limiti delle tecnologie via via disponibili), comprensione del linguaggio e del rapporto tra parole e mondo. Sistemi artificiali possono, infatti, in linea di principio, dotarsi di una qualche capacità di "fondare" i significati: già oggi essi possono collegare parole e immagini, e qualora abbiano una dimensione robotica, anche stabilire rudimentali connessioni tra parole e interazioni con cose e situazioni.[53]

1.4 Breve storia dell'IA

Da secoli l'uomo è affascinato e al tempo stesso impaurito dalla possibilità di realizzare entità artificiali intelligenti. Solo a partire dagli anni '50, si è passati dalle rappresentazioni fantastiche (nella letteratura, le arti figurative e il cinema) alla realtà.

1.4.1 L'IA prima dell'IA

Già nei miti dell'antica Grecia si possono rinvenire automi intelligenti: Pigmalione scolpì Galatea, una statua vivente (grazie all'intervento divino); il dio Efesto poteva creare esseri di bronzo animati, come Talos, il leggendario guardiano di Creta. Passando dal mito all'ingegneria meccanica, possiamo menzionare nell'antichità gli automi costruiti da Erone di Alessandria (vissuto nel primo secolo, e inventore, tra l'altro, del motore a vapore), usati per animare le divinità nei templi. In epoche più vicine, possiamo ricordare il mito del Golem di Praga, creato per difendere il ghetto ebraico da attacchi antisemiti, che sfuggì al controllo del suo creatore.

Il termine *robot* trae origine dall'opera teatrale *R.U.R. (Rossum's Universal Robots)* (sigla che sta per "Rossumovi univerzální roboti", cioè "i robot universali di Rossum"),

pubblicata nel 1920 dallo scrittore cecoslovacco Karel Capek [1890-1938]. I robot di Capek sono androidi costruiti per servire gli uomini, ma si ribelleranno ai loro padroni e ciò causerà la fine dell'umanità.[54]

Il tema del rapporto tra IA e intelligenza umana troverà sviluppo in numerose opere di fantascienza. Mi limito a ricordare l'opera di due autori, Arthur C. Clarke e Isaac Asimov.

Clarke[55] immaginò il calcolatore HAL (*Heuristically programmed ALgorithmic computer*), reso famoso dal film "2001 Odissea nello Spazio", diretto da Stanley Kubrick. HAL —capace non solo di ragionare, ma anche di comprendere il linguaggio umano (non solo tramite il suono, ma anche "leggendo le labbra"), di avere emozioni e di cogliere le emozioni altrui— acquista una psicologia umana, anzi troppo umana: prima per impedire che si vengano a conoscere i suoi errori e poi per proteggere sé stesso e la missione che gli è stata affidata, si rivolge contro gli astronauti al cui viaggio avrebbe dovuto sovrintendere.

Nel linguaggio di oggi diremmo che l'esempio di HAL attiene al disallineamento di valori (*value misalignment*), cioè al fatto che il comportamento di HAL, pur motivato dall'obiettivo ad esso assegnato, nel perseguimento di quell'obiettivo si discosta dai valori umani. Come far sì che il comportamento dei sistemi intelligenti si conformi e rimanga conforme ai valori umani è un aspetto chiave dell'etica dell'IA, tanto più importante quanto quei sistemi sono autonomi. Un aspetto particolarmente importante attiene al fatto che il perseguimento di uno scopo, senza tener conto degli effetti collaterali, può comportare conseguenze aberranti. Si è osservato[56] che un'IA sovrumana al fine di raggiungere un obiettivo apparentemente buono come quello di massimizzare la felicità degli umani potrebbe realizzare "istanziazioni perverse", come il forzato "impianto di elettrodi nei centri del piacere dei nostri cervelli." Per evitare che i sistemi intelligenti, nel perseguimento degli obiettivi ad essi affidati, adottino scelte pregiudizievoli agli interessi e valori umani, si è suggerito che essi dovrebbero perseguire un meta-scopo preminente su ogni obiettivo specifico loro assegnato: aiutare le persone a raggiungere gli scopi che esse desiderano perseguire.[57]

Asimov analizza il problema del rapporto tra gli uomini e sistemi di IA (robot) in numerosi volumi e racconti, nei quali egli supera l'usuale schema dell'artefatto che si ribella al suo creatore. Nei racconti di Asimov i robot sono di regola esseri benevoli, il cui funzionamento si ispira alle tre leggi della robotica:

1. Un robot non può nuocere a un essere umano o consentire, mediante la propria omissione, che un essere umano subisca danno.
2. Un robot deve obbedire agli ordini impartitigli da esseri umani, eccetto quando questi ordini confliggano con la prima legge.
3. Un robot deve proteggere la propria esistenza, fintantoché tale protezione non confligga con la prima o la seconda legge.[58]

In seguito, Asimov aggiungerà una legge ulteriore, la Legge Zero: "un robot non può danneggiare l'umanità o consentire, mediante la propria mancanza di azione, che essa venga danneggiata".[59] Questa legge, essendo superiore alle altre tre, consente ai robot di opporsi a singoli individui per il bene dell'umanità e, come evidenzia lo stesso Asimov, si rivela assai problematica. Robot benevoli la scoprono e l'adottano, al fine impedire agli esseri umani di autodistruggersi, ma nell'applicarla i robot debbono affrontare problemi di difficile soluzione: come determinare che cosa rappresenti il bene dell'umanità, e come stabilire che cosa possa favorirlo o pregiudicarlo a lungo termine? Inoltre, tale legge può favorire un eccessivo paternalismo da parte dei robot o addirittura essere usata da questi quale giustificazione opportunistica (razionalizzazione) di azioni criminali.

Nell'opera di Asimov, la benevolenza dei robot non esclude un aspetto problematico: la disponibilità di servitori robotici, con capacità superiori per molti aspetti a quelle umane, può indurre chi se ne serve a diventare dipendente dai propri schiavi meccanici, ad adagiarsi nella comodità, rinunciando all'iniziativa, rifiutando ogni rischio. Il tema della dipendenza dell'uomo dai robot ripropone così il tema della dipendenza del padrone dai propri schiavi, illustrato da Hegel nella sua *Fenomenologia dello spirito*.[60] Delegheremo tanta parte della nostra vita ai nostri aiutanti elettronici da perdere la capacità di pensare e agire autonomamente? Interporremo in tale misura i nostri schiavi elettronici tra noi e la soddisfazione dei nostri desideri (come direbbe Hegel) da divenire completamente passivi, ci trasformeremo in capricciose e inutili "macchine desideranti", avendo trasferito ai nostri schiavi elettronici tutte le attività produttive e comunicative necessarie per soddisfare le nostre voglie, così come le competenze e le conoscenze richieste a tal fine? E un robot morale e razionale perfetto sarà capace di servire persone moralmente imperfette, accetterà di farsi strumento di avidità e meschinità?[61]

1.4.2 Gli entusiasmi dei pionieri e il paradigma dell'IA simbolica

La ricerca scientifica e tecnologica sull'IA iniziò tra gli anni '40 e gli anni '50. Già nel 1943 Walter Pitts and Warren Sturgis McCulloch (due collaboratori di Norbert Wiener, l'inventore della cibernetica) mostrarono come reti di neuroni artificiali potessero elaborare informazioni, dando avvio alla ricerca sulle reti neurali (vedi Sezione 2.2.5).[62]

La nascita dell'IA viene tuttavia solitamente ricondotta a una celebre conferenza tenutasi a Dartmouth (New Hampshire, USA), che riunì per un mese alcuni tra i principali pionieri della materia. Lo scopo esplicito della riunione era lo studio dell'intelligenza automatica, partendo dall'ipotesi che "ogni aspetto dell'apprendimento e ogni altra caratteristica dell'intelligenza possa in principio essere descritto con tale precisione che si possa costruire una macchina capace di simularlo".[63]

La tesi fondamentale che ispirava gli studiosi riuniti a Dartmouth era infatti espressa dalla famosa *ipotesi del sistema simbolico fisico* (*physical system hypothesis*), cioè dall'ipotesi che l'intelligenza possa risultare dal funzionamento di un sistema che mani-