Giovanni Sartor

L'intelligenza artificiale e il diritto

Giappichelli

L’intelligenza artificiale e il diritto


IUSTITIAM COLIMUS

Giovanni Sartor

L’intelligenza artificiale e il diritto

Giappichelli

VIA PO, 21 - TEL. 011-81.53.111
http://www.giappichelli.it

ISBN/EAN 978-88-921-4458-3

G. Giappichelli Editore

Questo libro è stato stampato su
carta certificata, riciclabile al 100%

Stampa: Stampatre s.r.l. - Torino

Indice

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.

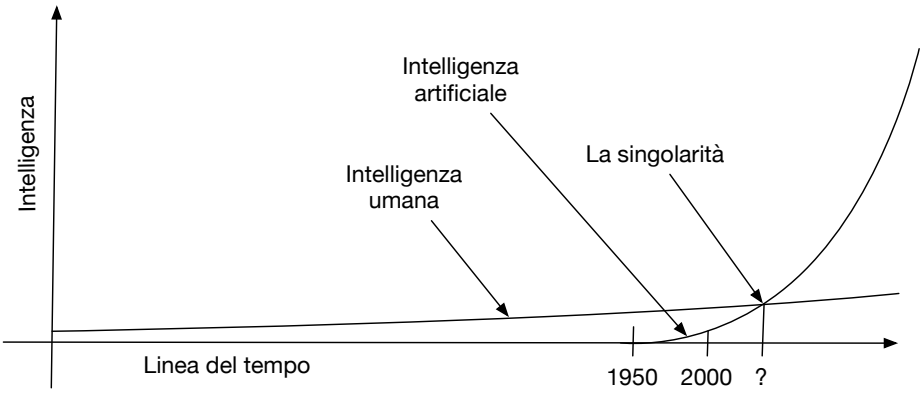

Figura 1.2: *L'evoluzione dell'intelligenza*

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]

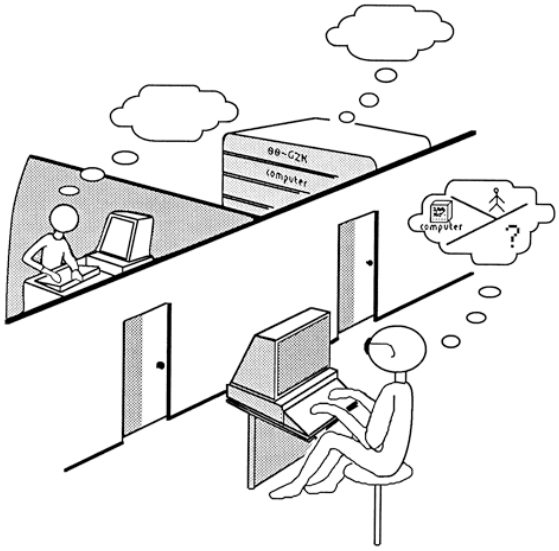

Figura 1.3: *Il test di Turing*

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-

pola strutture simboliche (per esempio, sequenze di parole o numeri) producendo altre sequenze simboliche, secondo determinati processi. Ecco come Alan Newell e Herbert Simon caratterizzano un tale sistema:

> Un sistema di simboli fisici consiste di un insieme di entità, chiamate simboli, che sono schemi fisici che possono presentarsi come componenti di un altro tipo di entità, chiamata espressione (o struttura simbolica). Pertanto, una struttura simbolica è composta di un numero di istanze (o occorrenze) di simboli correlati in qualche modo fisico (ad esempio, un simbolo può essere adiacente a un altro simbolo). In ogni istante di tempo il sistema conterrà una collezione di queste strutture simboliche. Oltre a queste strutture, il sistema contiene anche una collezione di processi che operano sulle espressioni per produrre altre espressioni: processi di creazione, modificazione, riproduzione e distruzione. Un sistema simbolico fisico è una macchina che produce nel tempo una collezione di strutture simboliche in evoluzione.[64]

Secondo gli stessi autori "un sistema simbolico fisico ha i mezzi necessari e sufficienti per l'azione intelligente generale".[65] Dato che ogni sistema di simboli fisici può essere realizzato mediante una macchina universale (come la macchina di Turing) e dato che i moderni calcolatori sono macchine universali, l'ipotesi che un sistema di simboli fisici sia capace di intelligenza implica che un calcolatore potrebbe dar vita all'intelligenza (una volta che fosse dotato di un software adeguato e di sufficiente memoria e capacità di calcolo).

> L'ipotesi del sistema di simboli fisici implica una teoria computazionale-simbolica dell'intelligenza [...] Pertanto l'ipotesi del sistema simbolico implica che l'intelligenza sarà realizzata da un calcolatore universale.[66]

Seguendo questa linea di pensiero non vi sono limiti assoluti o "filosofici" allo sviluppo dell'intelligenza automatica, si tratta solo di sviluppare tecnologie hardware, e soprattutto software, adeguate. In entrambe le direzioni ci sono stati notevoli progressi. Circa le tecnologie hardware, furono inizialmente sviluppati calcolatori specificamente dedicati all'IA, le cosiddette "macchine per il Lisp" (*Lisp machine*), ma la disponibilità di calcolatori "a scopo generale" (*general purpose*) sempre più potenti ed economici consentì di usare questi ultimi anche per le applicazioni di IA.[67] Nel campo del software le ricerche si svilupparono in due direzioni complementari: tecniche per la rappresentazione della conoscenza in strutture simboliche, e tecniche per l'elaborazione di tali conoscenze, cioè per il ragionamento automatico. Ciò corrispondeva al paradigma dell'IA "simbolica", cioè all'assunto che un sistema capace di risolvere problemi in modo intelligente (*intelligent problem solving*) debba unire due aspetti: una rappresentazione simbolica (in un linguaggio appropriato) delle conoscenze rilevanti, e la capacità di trarre conclusioni fondate su tali conoscenze. Questa idea è espressa con chiarezza da

John McCarthy e Paul Hayes (anche quest'ultimo fu tra i fondatori dell'IA), che danno la seguente definizione di intelligenza.

> Un'entità è intelligente se ha un modello adeguato del mondo (inclusi il mondo intellettuale della matematica, la comprensione dei propri scopi e altri processi mentali), se è capace di rispondere a un'ampia varietà di domande sulla base di quel modello, se può trarre informazioni ulteriori dal mondo esterno quando necessario, e può effettuare nel mondo esterno i compiti richiesti dai suoi scopi e consentiti dalle sue capacità fisiche.[68]

La nozione di intelligenza, per gli stessi autori, comprende due parti: una parte epistemologica e una parte euristica.[69]

> La parte epistemologica è una rappresentazione del mondo in una forma tale che la soluzione dei problemi derivi dai fatti espressi nella rappresentazione. La parte euristica è il meccanismo che sulla base dell'informazione risolve il problema e decide che cosa fare.[70]

Come vedremo nelle Sezione 2.1, secondo questo modello la conoscenza è espressa usando un linguaggio simbolico, nel quale i simboli sono entità linguistiche che fanno riferimento agli oggetti del dominio in esame, combinati in strutture sintattiche, e l'elaborazione della conoscenza avviene mediante ragionamenti, cioè catene di inferenze, che date certe premesse ne derivano altre, sulla base della struttura logica delle premesse. Pensiamo ad esempio alla semplice inferenza mediante la quale dalle premesse (1) "Tutti gli uomini sono animali", e (2) "Socrate è un uomo" possiamo inferire "Socrate è un animale". Da questa conclusione e dalla premessa ulteriore "tutti gli animali sono mortali" possiamo raggiungere la conclusione "Socrate è mortale." In questa inferenza compaiono simboli che fanno riferimento alle entità e ai tipi di entità su cui si ragiona (Socrate, uomo, animale, mortale) da essi si traggono inferenze giustificate dalle premesse (per es., l'inferenza sillogistica secondo la quale se tutti gli A sono B, e l'individuo x è un A, allora x è anche un B).

Un'importante estensione del paradigma appena prospettato è il tentativo di affiancare alle teorie che descrivono i diversi domini dell'azione intelligente, meta-teorie intese a indicare come le prime teorie debbano essere usate, estese, e aggiornate. In questo modo, funzioni come l'apprendimento dall'esperienza o l'effettuazione di inferenze analogiche possono essere rese compatibili, almeno in una certa misura, con una concezione tendenzialmente statica della conoscenza (la conoscenza come rappresentazione del mondo all'interno del sistema intelligente).

1.4.3 Sviluppo e crisi delle ricerche di IA

Negli anni seguenti, il programma di ricerca dell'IA simbolica diede origine a numerosi risultati. Da un lato furono sviluppati numerosi sistemi capaci di affrontare compiti

tali da richiedere intelligenza negli esseri umani (il gioco degli scacchi, la derivazione di teoremi matematici, la soluzione di problemi matematici o fisici, lo spostamento nell'ambiente fisico, ecc.). Dall'altro lato furono realizzati alcuni strumenti che facilitavano grandemente la realizzazione tali sistemi, come in particolare il linguaggio Lisp (*LISt Processing*).

Questi successi condussero a previsioni ottimistiche. Studiosi molto autorevoli si spinsero ad affermare che entro un prossimo futuro (uno o due decenni) sarebbero state disponibili macchine capaci di raggiungere l'intelligenza umana, di "svolgere ogni lavoro possa essere compiuto da un essere umano" (così Simon, nel 1965) e di "avere l'intelligenza generale di un essere umano medio" (così Minsky, nel 1970). Il lento sviluppo delle applicazioni di IA smentì però queste previsioni: passare dai semplici esempi usati nelle applicazioni pionieristiche a sistemi utili per affrontare problemi reali si rivelò molto difficile. In particolare, i sistemi di IA non riuscirono ad svolgere in modo soddisfacente attività che gli esseri umani compiono spontaneamente e apparentemente senza sforzo (come la comprensione del linguaggio, o l'identificazione di oggetti nello spazio). Si tratta delle funzioni comprese nella dotazione naturale della nostra specie, che vengono esercitate utilizzando le conoscenze attinenti al "senso comune", appartenenti a ogni persona normale (in parte innate, in parte acquisite nel corso della normale socializzazione).

Negli anni seguenti gli sforzi si concentrarono sul tentativo affrontare problemi riguardanti ambiti specialistici, la cui soluzione potesse essere derivata da un'ampia base di conoscenze. Si trattava quindi di realizzare *sistemi esperti* capaci di risolvere problemi che richiedessero una particolare competenza (*expertise*) specialistica (codificata nella base di conoscenza). Si elaborarono conseguentemente tecniche per la rappresentazione della conoscenza in forme tali da renderla elaborabile automaticamente, e procedure per utilizzare ampie basi di conoscenza.

Tra tali tecniche ebbe importanza crescente l'uso di metodi per il ragionamento automatico ispirati alla logica. La logica non solo consente di effettuare inferenze, ma ne garantisce altresì la validità: se le procedure inferenziali di un sistema esperto possono essere ricondotte alla logica, intesa in senso ampio come l'insieme dei metodi del ragionamento corretto, allora si può fare affidamento sul funzionamento del sistema stesso (assumendo che anche le premesse del sistema siano corrette). La logica computazionale (l'effettuazione automatica di inferenze logicamente corrette) conobbe, infatti, un rapido sviluppo. In particolare, all'inizio degli anni '70 fu inventato il Prolog, un linguaggio logico semplice e intuitivo, basato su una parte della logica predicativa, per il quale furono definite procedure di inferenza molto efficienti.

Tra gli anni '70 e '80 furono sviluppati numerosi sistemi esperti, e alcuni di essi conseguirono risultati significativi in diversi campi, come la diagnosi medica, l'analisi delle strutture molecolari o la progettazione di sistemi informatici.[71]

In questi anni iniziarono anche i primi studi in tema di IA e diritto e, come vedremo

nel seguito, non mancarono i tentativi di realizzare sistemi esperti di diritto. Anche nell'ambito dei sistemi esperti ai primi entusiasmi fece seguito una delusione. Si dovette costatare che i sistemi realizzati non erano in grado di sostituire la prestazione di un professionista esperto, ma al più di integrarne la competenza. L'attività di un esperto non si limita infatti all'applicazione di conoscenze preesistenti, ma richiede attività ulteriori, come l'esame del caso concreto, la ricerca delle informazioni rilevanti, la considerazione delle analogie con casi precedenti, la formulazione e la valutazione di ipotesi, l'individuazione e il superamento di incoerenze.[72] Pertanto, nello svolgimento di un compito complesso, un sistema esperto può integrarsi con le competenze umane, ma non le può, di regola, sostituire.

Inoltre, emersero alcune difficoltà inerenti allo sviluppo e la manutenzione dei sistemi esperti. In particolare, si constatava che era difficile e costoso rappresentare la conoscenza nella forma richiesta da un sistema esperto (assai più rigida e limitata rispetto al linguaggio umano) e mantenerla aggiornata, ma anche che non tutte le informazioni potevano essere espresse in questo modo e che, una volta ridotte in tale forma, le informazioni non potevano essere impiegate con la flessibilità di cui è capace l'intelligenza umana. In particolare, i modelli simbolici apparivano intrinsecamente inadatti in talune attività, che umani (e animali) compiono in modo inconscio, senza sforzo apparente, come il riconoscimento di oggetti, di volti, delle parole pronunciate nel linguaggio parlato, la traduzione automatica, e altri nei quali si tratta di identificare schemi di comunanze e analogie all'interno di dati suscettibili di presentarsi in modi diversi.

1.4.4 Dalla crisi ai primi successi

Nei primi anni '90 vi fu quindi una profonda crisi delle ricerche di IA, il cosiddetto "inverno dell'IA" (*AI winter*), un clima di generale sfiducia nei confronti delle prospettive dei sistemi intelligenti. Lo stesso termine "intelligenza artificiale" cadde in discredito, e gli studiosi che continuavano a occuparsi di sistemi intelligenti preferivano spesso qualificare la propria ricerca in altro modo (logica o linguistica computazionale, database deduttivi, sistemi probabilistici, supporto alle decisioni, ecc.). La crisi dell'IA simbolica determinò un'intensa attività di ricerca in direzioni diverse. Qui mi limito a ricordarne tre: la ripresa degli studi sui modelli computazionali dell'attività neurale, la creazione di modelli computazionali ispirati alle discipline matematiche ed economiche, l'attenzione per le dimensioni dell'azione e della comunicazione, .

Uno sviluppo importante attiene alla ripresa dei modelli ispirati all'attività neurale. Come si è visto nelle pagine precedenti, per l'indirizzo simbolico (logico), l'essenza dell'intelligenza consiste nell'uso di regole simboliche (regole di inferenza), per manipolare espressioni simboliche (le premesse) ed estrarne nuove strutture simboliche (le conclusioni). L'IA simbolica si concentra quinti sulle tecniche per il ragionamento. Le reti neurali si ispirano invece ad un modello biologico: esse mirano a riprodurre la struttura

e la dinamica del cervello piuttosto che una sua particolare funzione (il ragionamento). Per l'indirizzo neurale l'essenza dell'intelligenza consiste nell'adattamento all'esperienza: l'apprendimento avviene fondamentalmente mediante la modifica della forza delle connessioni tra i neuroni, così da dare alla rete la capacità di reagire agli stimoli con risposte più appropriate. L'accento non è sul ragionamento, ma piuttosto su attività "intuitive", come la percezione e, più in generale, il riconoscimento di strutture (*pattern*) rilevanti all'interno di dati di input.

L'elaborazione della conoscenza da parte di una rete neurale non è simbolica ma piuttosto distribuita. Per esempio, una rete neurale che classifica immagini di animali (con il nome dell'animale di cui si tratta) non contiene al suo interno definizioni che specifichino il significato dei nomi degli animali sulla base di altri simboli linguistici. Un concetto come quello di cane non ha nella rete definizioni linguistiche che specifichino le caratteristiche dei cani (per es., un cane è un mammifero della famiglia dei Canidi onnivoro, peloso, dotato di sensi sviluppati, ecc.). Il concetto di cane è invece rappresentato dall'aspetto della rete che consente il riconoscimento del cane (i neuroni attivabili in seguito alla presentazione dall'immagine del cane e le loro connessioni). L'elaborazione della conoscenza, e in particolare l'apprendimento, non avverrà mediante ragionamenti (processi medianti i quali vengono prodotti nuovi simboli sulla base dei simboli dati), ma piuttosto mediante complesse elaborazioni numeriche, governate dal calcolo differenziale, che trasformano la rete adattandola all'esperienza.

Quindi, i modelli ispirati all'attività neurale mirano a riprodurre l'aspetto reattivo-intuitivo dell'intelligenza (comune a tutti gli animali evoluti), cioè la capacità di rispondere agli input forniti dall'esperienza e di adattare tale risposta in modo appropriato (di apprendere), senza il tramite del ragionamento e della rappresentazione esplicita della conoscenza. Le reti neurali (e in generale i modelli connessionistici) hanno offerto una nuova tecnologia (e una nuova prospettiva) non solo per affrontare compiti del senso comune (come il riconoscimento di immagini o di volti) per i quali mancano modelli teorici precisi, ma anche per compiti specialistici attinenti al riconoscimento, alla luce di esperienze precedenti, di caratteristiche non definibili con precisione (come la rischiosità di un'operazione finanziaria, la probabilità che un soggetto incorra in un sinistro, la probabilità di una patologia sulla base di una radiografia, ecografia o altro esame medico, ecc.). Gli straordinari successi ottenuti dai modelli connessionistici hanno forse messo in ombra i progressi nei modelli simbolici. Le elaborazioni simboliche si sono estese all'uso di metodiche formali tratte non solo dalla logica matematica (cui l'IA si era ispirata fin dai propri inizi) ma anche dalle logiche filosofiche (logica dell'azione, della possibilità, di obblighi e permessi) e da discipline quali il calcolo delle probabilità, la teoria della decisione, la teoria dei giochi, le teorie dell'argomentazione. È divenuto così possibile utilizzare in misura crescente strumenti di IA per il supporto alla decisione, anche in contesti caratterizzati da incertezza (come nelle previsioni atmosferiche o finanziarie) o opinabilità (come nell'ambito giuridico).

L'attenzione per gli aspetti dinamici e relazionali (l'azione, l'interazione, la comunicazione) ha consentito di superare l'idea del sistema intelligente quale mero intelletto, privo di iniziativa, che si limita a rispondere alle domande dell'utilizzatore sulla base delle conoscenze registrate al suo interno. Numerose ricerche si sono indirizzate invece verso la realizzazione di *agenti intelligenti*, capaci non solo di elaborare informazioni, ma anche di ricercare le informazioni rilevanti, di percepirle esaminando l'ambiente e di agire sulla base degli obiettivi a essi assegnati, possibilmente interagendo con altri agenti dello stesso tipo o con interlocutori umani. È stato così enfatizzato l'aspetto "robotico" dell'IA, cioè la realizzazione di *robot* quali entità capaci di azione autonoma, sia nello spazio fisico (i robot industriali, come le sonde spaziali capaci di movimento autonomo o gli aerei senza pilota) sia nello spazio virtuale di Internet (come gli agenti software utilizzabili nella ricerca di informazioni o per fare acquisti on-line).

Un ulteriore dinamica —che si sovrappone parzialmente alla dialettica tra indirizzi subsimbolici e simbolici— attiene al passaggio dalla logica alla statistica, come modalità prevalente per l'elaborazione dell'informazione, e dalla rappresentazione umana all'apprendimento automatico, come modalità prevalente per la generazione di modelli computabili (Sezione 2.2).

Nei primi anni '90 le ricerche di IA hanno incontrato Internet, che da un lato richiedeva applicazioni informatiche intelligenti, dall'altro offriva un'enorme quantità di informazioni in formato digitale, informazioni alle quali si potevano applicare tecniche di IA. Le tecniche di IA hanno trovato impiego in diversi strumenti per la rete: motori di ricerca, sistemi che forniscono raccomandazioni agli utenti, sistemi per la costruzione di siti Web, agenti software per la ricerca di informazione e l'effettuazione di transazioni commerciali, ecc.

1.4.5 L'era dell'IA

Gli sviluppi appena indicati hanno determinato la fine dell'"inverno dell'IA". È fuor di dubbio che l'IA abbia avuto un enorme successo negli anni più recenti. Da un lato l'IA si è dotata di una solida base interdisciplinare. Al nucleo originario costituito dall'informatica, la matematica e la logica, si sono aggiunte altre discipline, come la statistica, l'economia, la linguistica, le neuroscienze, la psicologia, il diritto.

D'altro lato una serie di applicazioni di successo sono state sviluppate, e sono entrate nella vita di tutti noi:

- l'estrazione di informazioni da grandi masse di dati (*data mining*), oggi utilizzata in molti ambiti: le ricerche commerciali, l'individuazione di comportamenti fraudolenti (per es., nell'uso di carte di credito) o le indagini di polizia (per es., nell' anti-terrorismo);
- la selezione di informazioni rilevanti o l'eliminazione di quelle irrilevanti con tecniche intelligenti (utilizzate per es., nei filtri *antispamming*);

- l'interpretazione degli esami medici e la consulenza medica;
- la traduzione automatica, che oggi fornisce risultati già molto utili, anche se limitati;
- i giochi, per i quali sono stati realizzati sistemi automatici capaci di raggiungere alti livelli di competenza e prodotti commerciali di notevole successo;[73]
- la gestione e la logistica (ad esempio, la pianificazione di attività imprenditoriali o di operazioni militari);
- il riconoscimento di immagini, volti, e movimenti (ad esempio, nei sistemi per la vigilanza automatica);
- i robot fisici, utilizzati nell'esplorazione di regioni inospitali della terra (come l'Antartide) o dello spazio (ad esempio, Marte), ma anche nelle attività industriali, nelle pulizie domestiche, nel gioco (i *robot pet*, nuovo tipo di animali domestici);
- i sistemi di trasporto (aerei, automobili, treni) autonomi, cioè in grado di condurre il loro carico a destinazione anche senza l'intervento di un pilota umano;
- gli agenti software, ai quali l'utilizzatore può delegare compiti da eseguire con autonomia negli spazi virtuali (dalla ricerca di informazioni, all'effettuazione di transazioni commerciali);
- la ricerca intelligente di informazioni, 'analisi di documenti, la risposta a domande in linguaggio naturale (*question-answering)*, transazioni commerciali ad alta velocità (*high speed trading*), la robotica industriale, i veicoli (sempre più) autonomi, ecc.

Sulla base di questi successi si può ipotizzare che le odierne applicazioni di successo non solo si consolideranno e perfezioneranno, ma saranno altresì accompagnate dallo sviluppo di altre applicazioni. Probabilmente, lo sviluppo dell'IA seguirà la linea intermedia nella Figura 1.4, anche se non si possono escludere scenari diversi.[74]

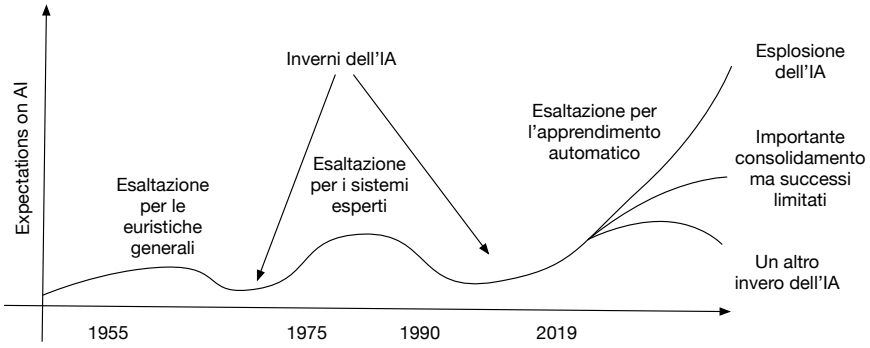

Figura 1.4: *Esaltazioni e disillusioni in tema di AI*

Capitolo 2

L'IA: tecnologie

Fino a pochi decenni fa, si assumeva che, al fine di realizzare un sistema intelligente, fosse necessario fornire al sistema stesso tutta la conoscenza necessaria per affrontare i compiti affidatigli, e che tale conoscenza dovesse essere rappresentata mediante linguaggi formali ed elaborata mediante ragionamenti automatici. Negli ultimi anni, l'attenzione di studiosi e sviluppatori si è spostata sulle tecniche per l'apprendimento automatico e sono stati realizzati numerosi sistemi che generano essi stessi la conoscenza (il modello) su cui si fondano le proprie prestazioni.

Nel presente capitolo si esamina dapprima la rappresentazione della conoscenza, per poi passare all'apprendimento automatico.

2.1 La rappresentazione della conoscenza

Come abbiamo visto nella sezione 1.3, non esistono a tutt'oggi sistemi informatici in grado di comprendere il significato del linguaggio umano, di avere accesso, in senso pieno, alla "semantica". Pertanto, le operazioni che un sistema può fare con documenti testuali dipendono solo dalle parole (considerate come mere sequenze di caratteri alfabetici e numerici) che compaiono in quei documenti e dalle strutture sintattiche impiegate per collegarle. Il sistema ragiona solo sulla base delle forme (non dei contenuti), e in questo senso il suo ragionamento è formale.

Per chiarire il concetto di sistema formale consideriamo la classica inferenza sillogistica che dalle premesse (1) "Ogni uomo è mortale", (2) "Socrate è un uomo", deriva la conclusione (3) "Socrate è mortale". Per effettuare questa inferenza non abbiamo bisogno di sapere che cosa significa uomo, che cosa significa mortale, che cosa significa Socrate. Ci basta riconoscere la struttura linguistica (la forma) delle premesse, e applicare una regola di inferenza agli elementi racchiusi in quella struttura. Anche un sistema automatico, una volta riconosciuta la forma delle premesse (ogni X è Y, z è un X, dove X e Y esprimono concetti generali, e z è il nome di un individuo), potrà essere in grado di trarre la conseguenza (z è Y): a tal fine non occorre capire i significati delle parole, basta unire il terzo termine nelle premesse (z) al secondo (Y) attraverso la copula "è".

Allo stesso modo consideriamo due semplice regole in materia di cittadinanza, che varrebbero se il cosiddetto *ius soli* fosse introdotto nel nostro paese, e due asserti di fatto relativi a due persone, Antonio e Mara.

1. SE x è nato in Italia ALLORA x è italianx
2. SE x è genitore di y E x è italianx ALLORA y è italianx.

Aggiungiamo due fatti specifici

1. *Antonio* è nato in Italia
2. *Antonio* è genitore di *Mara*.

Anche in questo caso per trarre la conclusione che Mara è italiana, non ci serve conoscere il significato delle parole "genitore", "cittadino", "italiano". Basta sostituire il nome Antonio al posto della x nella prima regola così da ottenere la regola specifica "SE Antonio è nato in Italia, ALLORA Antonio è italiano", e usare quest ultima regola in combinazione con il primo fatto (Antonio è nato in Italia), per trarre la conseguenza "Antonio è italiano". A questo punto possiamo sostituire alla x e alla y nella seconda regola rispettivamente Antonio e Mara, e dato che Antonio è genitore di Mara (il secondo fatto) e che Antonio è italiano (dal ragionamento precedente), concludere che Mara è italiana (si è usato il termine "italianx" per comprire entrambi i generi grammaticali). Anche in questo caso, il ragionamento è puramente formale e quindi automatizzabile: si basa sulle strutture sintattiche e prescinde dal significato delle parole (tranne quelle usate, secondo regole precise e univoche, per esprimere strutture sintattiche, come SE ... ALLORA ed E, nel nostro esempio), che possono essere infatti sostituite dai simboli della logica matematica o di un linguaggio di programmazione logica.[75]

Numerosi formalismi sono stati sviluppati per la rappresentazione della conoscenza (regole, concetti, schemi, logiche classiche, modali, descrittive, defeasible, schemi - *frame*, ontologie, ecc.) combinati con algoritmi per compiere diversi tipi di inferenze sulla base di tali rappresentazioni (inferenze deduttive, presuntive, induttive, probabilistiche, argomentative, basate sui casi, ecc.). Dall'interazione tra logica ed informatica è nata la programmazione logica, che consiste nell'uso "della logica simbolica per la rappresentazione esplicita di problemi e delle basi di conoscenza a questi associate, assieme all'uso di inferenze logiche controllate per la soluzione effettiva di tali problemi".[76]

Nelle pagine seguenti esamineremo brevemente il concetto di sistema basato sulla conoscenza, per poi passare ad esaminare le regole e il ragionamento basato su regole.

2.1.1 I sistemi basati sulla conoscenza

La struttura di un sistema basato sulla conoscenza. è rappresentata nella Figura 2.1. Si noti che gli esseri umani compaiono come utenti del sistema e come creatori della base di conoscenza del sistema stesso. La creazione della conoscenza è affidata ad

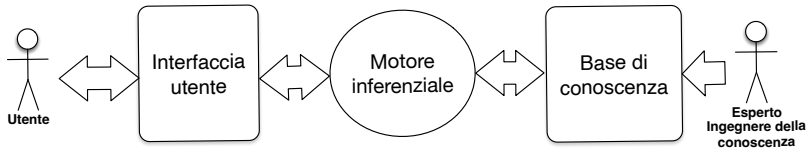

Figura 2.1: *Il sistema basato sulla conoscenza*

esperti nella materia su cui verte l'uso del sistema, possibilmente aiutati da esperti nella rappresentazione formale della conoscenza (i cosiddetti ingegneri della conoscenza).

Nel modello del sistema basato sulla conoscenza, il sistema informatico dispone di due componenti fondamentali (oltre all'interfaccia verso l'utente: una *rappresentazione della conoscenza* rilevante (la base di conoscenza), e metodi per il *ragionamento automatico* applicabili a tale rappresentazione (il motore inferenziale).

2.1.2 Il ragionamento mediante regole

Ricordiamo che il ragionamento consiste in generale nel passaggio da certe ragioni (premesse, obiettivi, ecc.), a conclusioni giustificate da tali ragioni. Tale passaggio avviene in generale secondo modelli generali o schemi di ragionamento, che sono forniti dalla logica (intesa in senso ampio).

Per esempio, le due premesse (1) *A* e (2) "SE *A* ALLORA *B*" giustificano la conclusione (3) *B*, quali che siano i particolari contenuti espressi da *A* e *B*.

Al riguardo, è importante distinguere due tipi di ragionamento, sui quali torneremo nel seguito: il ragionamento conclusivo e il ragionamento defeasible (che in italiano può essere tradotto con "presuntivo", "disfattibile" o anche "defettibile"). Nel ragionamento conclusivo (detto anche deduttivo) la verità delle premesse garantisce la verità delle conclusioni: è impossibile che le premesse siano vere e la conclusione falsa. Nel ragionamento defeasible la verità delle premesse fonda solo la presunzione delle conclusioni: se le premesse sono vere, anche la conclusione si assume essere vera a meno che non risulti applicabile un'eccezione. Come esempio di ragionamento conclusivo si consideri il classico sillogismo.

1 Tutti gli uomini sono mortali,
2 Socrate è un uomo,
  PERTANTO
3 Socrate è mortale.

È impossibile che siano valide le premesse (1) e (2) e sia invalida la conclusione (3). Consideriamo invece l'inferenza seguente:

1 I labrador sono (normalmente) cani pacifici,
2 Fido è un labrador,
PERTANTO (presumibilmente)
3 Fido è un cane pacifico.

In questo caso siamo giustificati dalle premesse (1) e (2) a concludere che Fido è pacifico, ma tale conclusione varrà solo fino a quando non abbiamo evidenza in contrario (per esempio, abbandoneremmo tale conclusione se Fido iniziasse a ringhiare, o ci venisse detto che ha aggredito un passante).

Nel diritto le inferenze sono spesso defeasible. Le regole generali, infatti, come espresse dai legislatori, dalla giurisprudenza o dalla dottrina, intendono cogliere le situazioni "normali", e sono completate da eccezioni rivolte a casi speciali. Così diciamo che chi causa colpevolmente un danno ha l'obbligo di risarcirlo, ma tale regola generale è limitata da eccezioni che escludono il risarcimento in presenza di incapacità, legittima difesa, stato di necessità, ecc. Nuove eccezioni possono poi essere introdotte per trattare casi non coperti dalle eccezioni espressamente previste, rispetto ai quali la regola generale si riveli inadeguata (per esempio, eccezioni al principio secondo il quale spetta al danneggiato fornire prova della negligenza della colpa della controparte).

Non è possibile illustrare in questa sede i metodi del ragionamento logico (o, meglio, dei diversi ragionamenti autorizzati da diversi modelli logici), né il modo in cui quei metodi possono essere applicati da sistemi informatici. Ai nostri fini è sufficiente illustrare il tipo di ragionamento più frequente in ambito giuridico, cioè l'applicazione di regole, intese come enunciati *condizionali*, che collegano un *antecedente* a un *conseguente* consentendo di inferire il secondo dal primo (più esattamente una specificazione del secondo da una specificazione del primo. L'antecedente può consistere di una sola proposizione atomica, o di una combinazione (una congiunzione o una disgiunzione) di tali proposizioni, dette *condizioni*, mentre il conseguente consiste usualmente di una singola proposizione (positiva o negata), detta anche *conseguenza*.

Il concetto di regola come struttura condizionale corrisponde all'idea diffusa che le regole giuridiche colleghino una fattispecie astratta e una conseguenza giuridica, anch'essa astratta: la fattispecie astratta è l'antecedente della regola e la conseguenza giuridica astratta ne è il conseguente. L'antecedente può consistere di un solo elemento o, invece, di una congiunzione di elementi. Quali esempi di regola così intesa si considerino, gli enunciati seguenti: "se una persona causa colpevolmente un danno e il danno è ingiusto, allora lo deve risarcire", o "se una persona è nata da un cittadino italiano, allora quella persona è cittadino italiano".

Il ragionamento che sarà qui esaminato consiste semplicemente nella derivazione di un'istanza specifica (concreta) del conseguente di una regola, data una corrispondente istanza dell'antecedente della stessa. Nella terminologia giuridica diremo che si tratta di derivare un effetto giuridico concreto (una specificazione dell'effetto giuridico astrattamente previsto dalla regola), data una fattispecie concreta (una specificazione della

fattispecie astratta). Ecco un semplice esempio. Assumiamo che Lucia, stizzitasi perché Renzo continuava a chattare con un'amica, abbia gettato il telefonino di lui dalla finestra, causando la rottura dello schermo, e applichiamo a questa fattispecie concreta la regola generale della responsabilità civile:

1 SE una persona x ha intenzionalmente causato a un'altra persona y un danno z E z è ingiusto ALLORA x deve risarcire il danno z a y
2 *Lucia* ha intenzionalmente causato a *Renzo* il danno $RotturaTelefonino$
3 $RotturaTelefonino$ è ingiusto
PERTANTO (presumibilmente)
4 *Lucia* deve risarcire il danno $RotturaTelefonino$ a *Renzo*

Si possono ricondurre al ragionamento mediante regole anche inferenze che nel linguaggio naturale sono espresse in altro modo. Per esempio, anche il sillogismo deduttivo sulla mortalità di Socrate può essere rappresentato come applicazione di una regola (la premessa (1) dell'esempio):

1 SE x è umano, ALLORA x è mortale
2 $Socrate$ è umano
PERTANTO
3 $Socrate$ è mortale

Lo stesso può farsi per inferenze defeasible come la seguente:

1 Chi acquista la prima casa gode della riduzione al 2% dell'imposta di registro
2 *Lucia* acquista la sua prima casa
PERTANTO (presumibilmente)
3 *Lucia* gode della riduzione al 2% dell'imposta di registro

Si tratta di un'inferenza defeasible, perché vale solo in assenza di eccezioni (per esempio, la riduzione non si applica se si tratta di un'abitazione di lusso). Anche questa inferenza può essere riformulata come applicazione di una regola.

1 SE una persona x acquista la sua prima casa, ALLORA x gode della riduzione al 2% dell'imposta di registro
2 *Lucia* acquista la sua prima casa
PERTANTO (presumibilmente)
3 *Lucia* gode della riduzione dell'imposta di registro

Nell'esempio seguente si riportano alcune possibili regole in tema di borse di studio:

r1: x ha diritto alla borsa di studio SE
  (a) x soddisfa i requisiti soggettivi E
  (b) x soddisfa i requisiti di merito

r2: x soddisfa i requisiti soggettivi SE
  (a) x è cittadino comunitario E
  (b) l'età di x è inferiore a 30 anni

r3: x soddisfa i requisiti di merito SE
  (a) x ha sostenuto almeno 4 esami E
  (b) la media dei voti di x è superiore al 28

r4: x è cittadino comunitario SE
  (a) x è cittadino dello stato y E
  (b) y appartiene all'Unione Europea.

Dati i fatti seguenti:

f1: l'età di $Sofia$ è 25 anni
f2: $Sofia$ ha sostenuto 5 esami
f3: la media di voti di $Sofia$ è 29
f4: $Sofia$ è cittadino dello stato *Romania*
f5: *Romania* appartiene all'Unione Europea

L'applicazione delle regole sopra indicate condurrà alla conclusione che

c$*$: $Sofia$ ha diritto alla borsa di studio.

Tale conclusione potrà essere raggiunta nel modo seguente:

1. Applicando la regola r4 ai fatti f4 e f5 si trae una prima conclusione:
   (c1) $Sofia$ è cittadina comunitaria.
2. Applicando la regola r3 ai fatti 2 e 3, può trarre una seconda conclusione:
   (c2) $Sofia$ soddisfa i requisiti di merito.
3. Applicando la regola r2 alla conclusione c1 e al fatto f1 si trae una terza conclusine: (c3)$Sofia$ soddisfa i requisiti soggettivi.
4. Applicando la regola r1 alle conclusioni c2 e c3 si trae la conclusione finale:
   (c$*$) $Sofia$ ha diritto alla borsa di studio.

Il ragionamento appena illustrato è chiamato *concatenamento in avanti* (*forward chaining*). Dato un insieme di fatti noti (incluse le conclusioni già derivate in precedenza),

il concatenamento in avanti esamina se le condizioni di qualche regola risultino soddisfatte. In caso positivo aggiunge la conclusione della regola ai fatti noti; passa poi a considerare se, grazie alle nuove conclusioni, sia possibile trarre, allo stesso modo, le conclusioni di altre regole (aggiungendo anch'esse ai fatti noti), e così via. Il processo termina quando si sia raggiunta la conclusione desiderata o si sia stabilita l'impossibilità di raggiungerla.

Il concatenamento all'indietro procede nel senso opposto. Nel nostro caso il processo per rispondere al quesito se Sofia abbia diritto alla borsa di studio sarebbe il seguente:

1. $Sofia$ ha diritto alla borsa di studio (conclusione finale c$*$) se $Sofia$ soddisfa le condizioni della regola r1, cioè ha i requisiti soggettivi e di merito.
2. $Sofia$ ha i requisiti soggettivi (conclusione c1) se soddisfa le condizioni della regola r2, cioè è cittadina europea e ha un'età inferiore ai 30 anni.
3. $Sofia$ è cittadina europea (conclusione c2) se soddisfa le condizioni della regola r4, cioè $Sofia$ è cittadina di uno stato che appartiene all'Unione Europea,
4. Invero, $Sofia$ soddisfa le condizioni della regola r4 ed è quindi è cittadina europea (c2), poiché è cittadina della Romania (fatto f4), stato che appartiene all'unione europea (fatto f5).
5. Invero, $Sofia$ ha i requisiti soggettivi (conclusione 2), poiché oltre ad essere cittadina europea (conclusione c2) ha meno di 30 anni (fatto f1).
6. $Sofia$ soddisfa i requisiti di merito (conclusione 3) se soddisfa le condizioni della regola r3, cioè ha sostenuto più di 4 esami e la sua media è superiore al 28.
7. Invero, $Sofia$ soddisfa i requisiti di merito poiché $Sofia$ ha sostenuto più di 4 esami (fatto r2) e ha un'età inferiore ai 30 anni (fatto f1).
8. Invero, $Sofia$ ha diritto alla borsa di studio (conclusione c$*$) poiché soddisfa i requisiti soggettivi (conclusione c2) e di merito (conclusione c3).

Come illustra l'esempio, data una conclusione da dimostrare, il ragionamento all'indietro ricerca una regola il cui conseguente coincida con la conclusione cercata. Procede quindi a esaminare se le condizioni di quella regola possano essere a loro volta dimostrate, in quanto conclusioni di altre regole o fatti noti. Il processo termina quando si risale a un insieme di fatti noti dal quale si deriva (mediante una concatenazione di regole) la conclusione cercata, o si sia riscontrata l'impossibilità di rinvenire quei fatti.

Per quanto riguarda il ragionamento defeasible, esistono numerosi modelli logici per trattare le eccezioni e i conflitti tra regole.[77] Qui mi limito a ricordare il modello più semplice, usato nel Prolog, il linguaggio di programmazione logica più diffuso. In questo linguaggio troviamo la "negazione per fallimento" (*negation by failure*) (usualmente

espressa con il termine NOT, ma che qui per chiarezza esprimiamo con la locuzione "NON RISULTA CHE"): la proposizione "NON RISULTA CHE *A*" si considera soddisfatta a condizione che siano falliti tutti i tentativi di dimostrare la proposizione *A*, sulla base delle informazioni disponibili (quelle incluse nella base di conoscenza). Seguendo questa idea la regola di cui sopra sul beneficio fiscale per acquisto di una prima casa potrebbe essere formulata nel modo seguente:

1. x ha diritto alla riduzione dell'imposta di registro SE
   (a) x acquista l'immobile y] E
   (b) y è la prima casa di x E
   (c) NON RISULTA CHE c'è un'eccezione alla concessione dei benefici prima casa per y
2. c'è un'eccezione alla concessione dei benefici prima casa per y SE
   (a) y è un'abitazione di lusso.

Date queste premesse, se non risulta che l'abitazione acquistata sia un'abitazione di lusso (questa informazione non è derivabile dalla base di conoscenza), allora si potrà concludere che la persona gode del beneficio. Assumiamo di aggiungere alla base di conoscenza i fatti seguenti:

1. *Lucia* acquista l'immobile "Palazzotto di Don Rodrigo"
2. "Palazzotto di Don Rodrigo" è un'abitazione di lusso

Date queste informazioni non è possibile concludere che Lucia ha diritto al beneficio fiscale, poiché, usando la regola (2) si conclude che c'è un'eccezione alla concessione dei benefici prima casa per il Palazzotto di Don Rodrigo, e quindi, non risulta soddisfatta l'ultima condizione per la concessione del beneficio fiscale (quella che richiede che non risultino eccezioni).

Nella logica e nei linguaggi di programmazione logica normalmente si usa una sintassi meno intuitiva di quella appena illustrata, che impone che i predicati siano rappresentati da un'unica parola, seguita da variabili e costanti tra parentesi. Per esempio, nel linguaggio Prolog, l'ultima regola si dovrebbe rappresentare nella forma seguente (dove la virgola "," tra proposizioni esprime la congiunzione "e", e il simbolo ":-" sta per "se"):

1. ha_diritto_alla_riduzione_dell_imposta_di_registro(X):-
   (a) acquista_l_immobile(X,Y),
   (b) è_la_prima_casa_di (Y,X),
   (c) c_è_una_eccezione_alla_concessione_dei_benefici_prima_casa_per(Y).

Oggi esistono sistemi nei quali il giurista (e in generale, l'utente) può utilizzare un linguaggio più intuitivo e comprensibile, affidando eventualmente al sistema, se necessario, il compito di tradurre quel linguaggio in un formalismo logico standard.[78]

È importante ricordare che la correttezza di un'inferenza —il fatto che essa sia conforme ai principi della logica— non garantisce la correttezza della sua conclusione.[79] Per esempio, entrambi i seguenti ragionamenti, sono corretti, ma non possono essere accolte entrambe le conclusioni che ne discendono.

Prima inferenza.

1. x può accedere alla casella di posta z SE
   (a) x è il datore di lavoro di y E
   (b) z è la casella aziendale che x ha assegnato a y
2. *Lucia* è il datore di lavoro di *Renzo*
3. `renzo@filandaManzoni.it` è la casella aziendale che *Lucia* ha assegnato a *Renzo*
   PERTANTO
4. *Lucia* può accedere alla casella di posta `renzo@filandaManzoni.it`

Seconda inferenza.

1. x NON può accedere alla casella di posta z SE
   (a) x è il datore di lavoro di y E
   (b) z è la casella aziendale che x ha assegnato a y
2. *Lucia* è il datore di lavoro di *Renzo*
3. `renzo@filandaManzoni.it` è la casella aziendale che *Lucia* ha assegnato a *Renzo*
   PERTANTO
4. *Lucia* NON può accedere alla casella di posta `renzo@filandaManzoni.it`

Le conclusioni di tali ragionamenti sono infatti incompatibili (o Lucia può accedere alla casella di posta, o non può farlo) e quindi non possono essere entrambe accolte. Tuttavia, l'inferenza di ciascuna conclusione è giustificata dalle corrispondenti premesse. Il problema è che non possono essere accolte le premesse di entrambe le decisioni: o è valida la regola secondo la quale il datore di lavoro può accedere alla posta del dipendente, o è valida la regola secondo cui lo stesso non vi può accedere. Non si tratta di un problema di inferenza logica, ma piuttosto di interpretazione del diritto vigente.

La correttezza della conclusione tratta da un certo insieme di premesse mediante un'inferenza, è, in generale, garantita solo dalla compresenza delle seguenti condizioni:

- la correttezza delle premesse;
- la correttezza dello schema di inferenza;
- l'assenza di inferenze prevalenti contro l'impiego delle premesse o contro l'esecuzione dell'inferenza.

Le prime due condizioni sono sufficienti a garantire conclusioni tratte secondo il ragionamento deduttivo (che presuppone, che premesse e inferenze non ammettano eccezioni), mentre la terza condizione si applica qualora si ricorra al ragionamento defeasible (vedi Sezione 4.1.5).

2.1.3 Successi e limiti del modello logico

Gli importanti risultati teorici nel campo della rappresentazione della conoscenza e del ragionamento automatico furono seguiti da risultati applicativi significativi, ma non così dirompenti come ci si aspettava, non tali da cambiare le regole del gioco. Numerosi sistemi esperti furono realizzati in diversi ambiti —medicina, diritto, ingegneria, chimica, ecc.— che contenevano ampie basi di conoscenza, combinate con motori inferenziali. Quei sistemi spesso non ebbero successo o ebbero un successo limitato.

Il limite più importante nei sistemi esperti era infatti rappresentato dal "collo di bottiglia" della rappresentazione della conoscenza. Tali sistemi infatti possono trarre inferenze solo dalla propria base di conoscenza. Quindi, se la base di conoscenza non include tutte le informazioni rilevanti per risolvere il problema proposto, la risposta sarà inadeguata. Ci si accorse che, in molti settori e per molte applicazioni, l'obiettivo di fornire in anticipo tutta la conoscenza rilevante —per la soluzione di qualsiasi quesito potesse essere rivolto al sistema— era assai difficile e in molti casi impossibile.

In particolare, sarebbe stato necessario che il sistema acquisisse anche le conoscenze tacite, che rimangono inespresse nell'interazione umana e spesso non sono facilmente o compiutamente esprimibili nel linguaggio[80] e quelle del senso comune, normalmente non esplicitate poiché si presume che tutte le conoscano. Inoltre, possono esservi diversi modi di esprimere le stesse conoscenze, cosicché le rappresentazioni formali della conoscenza possono riflettere le idiosincrasie del loro autore. Infine, la conoscenza non è statica, ma dinamica, e si richiede un suo continuo aggiornamento.

In generale, solo in ambiti ristretti i sistemi basati sulla rappresentazione esplicita della conoscenza hanno condotto ad applicazioni di successo. Nell'ambito del diritto alcuni sistemi basati sulla conoscenza sono stati impiegati con successo nella pratica legale ed amministrativa, in particolare nell'applicazione della normativa fiscale e della sicurezza sociale (vedi Sezione 4.1.3). Tuttavia, gli studi e le applicazioni nel campo dei sistemi basati sulla conoscenza giuridica non hanno prodotto fondamentali trasformazioni del sistema giuridico e dell'applicazione del diritto.

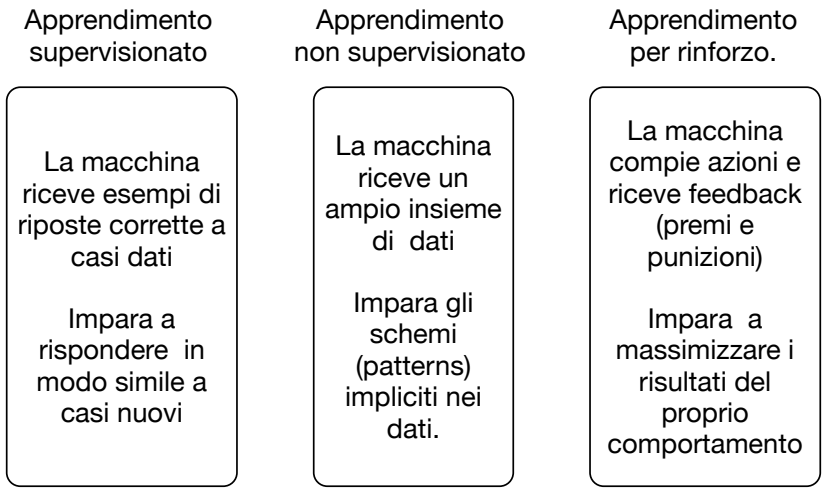

Figura 2.2: *Tipi di apprendimento automatico*

2.2 L'apprendimento automatico

L'intelligenza artificiale ha fatto enormi passi avanti da quando si è concentrata sulle tecnologie per l'apprendimento automatico, applicandole a grandi masse di dati. Nei sistemi basati sull'apprendimento automatico la conoscenza relativa all'ambito di applicazione del sistema non è più fornita dall'uomo; è invece costruita dal sistema, sulla base degli esempi forniti dai dati cui il sistema stesso ha accesso. Questo indirizzo ha condotto ad un grande numero di applicazioni di successo in molti settori, dalla traduzione automatica, all'ottimizzazione industriale, al marketing, alla visione robotica, al controllo dei movimenti, e altri ancora. Grazie ai metodi di apprendimento automatico, le macchine possono estrarre le informazioni rilevanti dai dati di input e così imparare a svolgere in modo adeguato le funzioni a esse affidate. In questo modo, come già osservava Alan Turing negli anni '50, una macchina capace di apprendere può operare in modi che non sono stati anticipati dai suoi creatori e addestratori, e anche senza che questi conoscano i dettagli del funzionamento interno della macchina.[81]

2.2.1 Indirizzi nell'apprendimento automatico

Si distinguono tre principali indirizzi nell'apprendimento automatico: l'apprendimento supervisionato (*supervised*), l'apprendimento per rinforzo (*reinforcement* e l'apprendimento non-supervisionato (*unsupervised*) (vedi Figura 2.2).

L'apprendimento supervisionato. Nell'apprendimento supervisionato la macchina apprende mediante supervisione, cioè attraverso una fase di istruzione o addestramento in cui le viene fornito un ampio insieme di esempi, ciascuno dei quali combina la descrizione di un caso alla risposta corretta allo stesso. Su questa base la macchina stessa costruisce un modello generale, applicabile anche a casi nuovi, parzialmente diversi da quelli nell'insieme di addestramento.

Ecco alcuni esempi: in un sistema destinato a riconoscere oggetti (per esempio, a classificare animali) che compaiono in immagini online, ogni immagine nell'insieme di addestramento è etichettata con il tipo di oggetto che contiene (per es., cane, gatto, coniglio, ecc.); in sistemi per la traduzione di testi, ogni (frammento) di documento nel linguaggio sorgente è appaiato alla sua traduzione; nei sistemi per la selezione del personale, la descriziondie ciascun candidato valutato in passato (età, esperienza, studi, ecc.) è collegata all'esito della selezione (o alla valutazione della prestazione lavorativa del candidato nel caso in cui questi sia stato assunto); nei sistemi che propongono suggerimenti (*recommender system*) sugli acquisti, le caratteristiche di ciascun consumatore sono associate agli oggetti acquistati dallo stesso; nei sistemi che esaminano richieste di mutuo, i dati raccolti su ciascuna richiesta già valutata (in particolare le informazioni sul richiedente) sono associati alla decisione adottata (o anche, rispetto alle richieste accettate, a informazioni sul successivo adempimento o inadempimento da parte del richiedente).

Come illustrano questi esempi, l'addestramento di un sistema non sempre richiede che un istruttore umano si assuma il compito di fornire esempi di risposte corrette al sistema. In molti casi, l'insieme di addestramento può essere raccolto "nella natura" (*in the wild*), cioè dal web aperto, o comunque da dati storici, concernenti attività svolte in passato. Per esempio, immagini o facce disponibili sulle reti sociali (nelle pagine aperte al pubblico), già etichettate dagli utenti dei servizi, possono essere prelevate ("raschiate" o *scraped*, in inglese) e usate per l'addestramento di classificatori automatici. Similmente, nel caso prototipico di giustizia predittiva –la predizione di decisioni giudiziarie future– gli esempi consistono in precedenti (casi decisi in passato) registrati in archivi di giurisprudenza, e ciascun esempio associa la descrizione dei fatti di un precedente alla decisione adottata nello stesso (vedi Sezione 4.3).

L'apprendimento per rinforzo. L'apprendimento per rinforzo è simile all'apprendimento supervisionato, nel senso che entrambi comportano addestramento mediante esempi. Tuttavia, nell'addestramento per rinforzo, il sistema non ha bisogno di un istruttore, poiché apprende dai risultati delle proprie azioni, cioè dalle ricompense o penalità (punti guadagnati o persi) che sono collegati ai risultati a quelle azioni. Per esempio, nel caso di un sistema destinato a partecipare a un gioco (per esempio, gli scacchi), le ricompense possono essere collegate alle vittorie e le penalità alle sconfitte; in un sistema che apprende a giocare in borsa, le ricompense possono essere collegate ai guadagni ottenuti

e le penalità alle perdite; in un sistema che impara a inviare messaggi pubblicitari mirati, le ricompense possono essere collegate ai click degli utenti.

In ogni caso, il sistema basato sul rinforzo osserva i risultati delle proprie azioni e si autosomministra le ricompense e penalità appropriate. Essendo diretto a massimizzare il proprio punteggio complessivo, il sistema apprende così a compiere le azioni che più probabilmente conducono a risultati collegati a ricompense (vittorie, guadagni, click) e a evitare le azioni che più probabilmente conducono a risultati collegati a penalità (sconfitte, perdite, nessun click). I sistemi basati sul rinforzo operano integrando sfruttamento (*exploitation*) ed esplorazione: in ogni situazione (stato dell'ambiente in cui operano) tendono a riprodurre le azioni che in passato si sono rivelate maggiormente efficaci in tale situazione, ma adottano di tanto in tanto azioni diverse, scelte a caso, per sperimentarne l'efficacia. Per esempio, un sistema che apprende a giocare a scacchi, oltre a riprodurre le mosse che si siano rivelate più efficaci (rispetto alle diverse configurazioni dei pezzi sulla scacchiera), sperimenterà di tanto in tanto mosse diverse, che potrebbero rivelarsi preferibili a quelle adottate in passato. In questo modo il sistema continua ad aggiornare la lista delle mosse più efficaci alla luce dell'esperienza, e quindi a migliorare le proprie prestazioni.

Nel campo della giustizia predittiva non sembra operino già oggi sistemi basati sul modello dell'apprendimento per rinforzo. A tale fine bisognerebbe che fosse possibile determinare le conseguenze delle decisioni del sistema e valutarle come positive e negative. Per esempio, il rinforzo potrebbe essere positivo (ricompensa) o negativo (punizione) a seconda che il giudice umano accolga il suggerimento del sistema, o lo respinga, oppure a seconda che la proposta del sistema venga confermata o rigettata in appello.

L'apprendimento non supervisionato. Nell'apprendimento non-supervisionato, infine, il sistema intelligente impara senza ricevere istruzioni, né da fonti esterne (apprendimento supervisionato), né dai risultati delle proprie attività (apprendimento per rinforzo). Le tecniche dell'apprendimento non-supervisionato sono usate in particolare per il raggruppamento (cosiddetta clusterizzazione – *clustering*), cioè per riunire insiemi di oggetti che presentano somiglianze o connessioni rilevanti (i documenti che riguardano gli stessi oggetti o problemi, le persone con simili caratteristiche, le parole che hanno simili significati o funzioni).

Per esempio, nell'ambito di un'indagine, può essere utile riunire i documenti simili in gruppi distinti, per individuare quelli che attengono al caso in esame. Oppure in un'indagine su precedenti giudiziari finalizzata all'adozione di nuove politiche criminali, si potrebbe chiedere al sistema di raccogliere i casi simili, e quindi esaminare, per esempio, le connessioni tra certi tipi di reati e certe caratteristiche delle loro occorrenze (per esempio, l'uso di droghe o di armi).

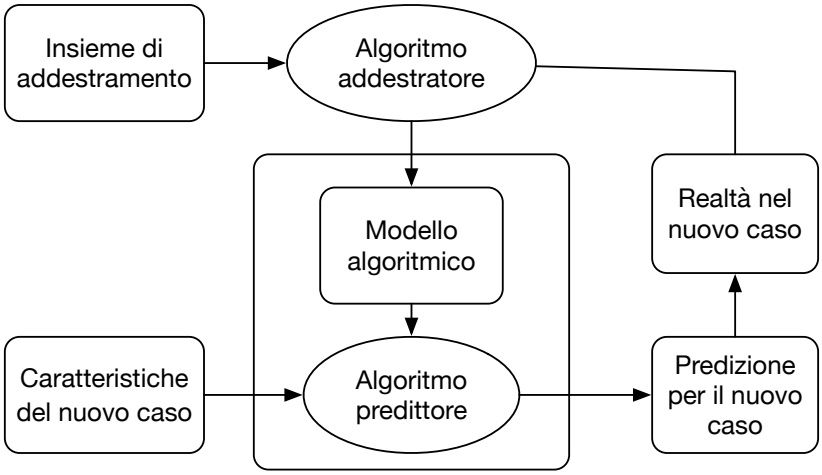

Figura 2.3: *L'apprendimento supervisionato*

2.2.2 L'apprendimento supervisionato: addestramento e costruzione di un modello

In tutti i sistemi per l'apprendimento automatico possiamo distinguere due aspetti: l'algoritmo discente (*learning*, detto anche algoritmo addestratore, *training*), che apprende (costruisce il modello) usando gli esempi forniti al sistema, e l'algoritmo o modello appreso (il risultato dell'apprendimento).

La Figura 2.3 raffigura il processo dell'apprendimento supervisionato. L'algoritmo discente usa l'insieme di addestramento per costruire un modello del compito che il sistema dovrà svolgere. Tale modello può essere visto, in generale, come una funzione matematica, cioè come un meccanismo che collega possibili input a output corrispondenti.

Per esempio, il modello potrebbe collegare possibili immagini di animali alle parole che denotano la specie animale corrispondente (immagini di gatto alla parola "gatto", di cane alla parola "cane", ecc.) o potrebbe collegare descrizioni dei "fatti" (naturali o giuridici) presenti in possibili casi giudiziari all'indicazione di decisioni corrispondenti.

Il modello non si limita a riprodurre gli esempi presenti nell'insieme di addestramento (a fornire la stessa soluzione per i casi che coincidano esattamente con uno di tali esempi), ma ne offre una generalizzazione: esso può essere applicato anche a nuovi casi, che differiscono in qualche aspetto da ciascuno degli esempi su cui si è basato l'addestramento. Nelle pagine seguenti ci limitiamo ad illustrare l'apprendimento supervisionato (Figura 2.3).

Il modello predisposto dall'algoritmo discente è poi usato da un diverso algoritmo, l'algoritmo predittore, per fornire risposte sperabilmente corrette a casi nuovi. Se gli esempi più simili al caso nuovo (rispetto alle caratteristiche maggiormente suscettibili di influenzare l'esito) hanno avuto una certa risposta, l'algoritmo predittore potrà proporre la stessa risposta nel caso nuovo. Per esempio, se nell'insieme di addestramento le immagini di animali più simili (negli aspetti correlati alla classificazione) alla nuova immagine proposta al sistema sono etichettate come immagini di gatto, anche la nuova immagine sarà classificata nello stesso modo; se nell'insieme di addestramento i richiedenti di prestiti le cui caratteristiche si avvicinano a quelle del nuovo candidato sono classificati come inaffidabili, il sistema classificherà nello stesso modo il nuovo candidato; se in passato i lavoratori aventi caratteristiche più simili al nuovo candidato sono stati assunti, il sistema predirà l'assunzione del nuovo candidato. Infine, in ambito giudiziario (ai fini della giustizia predittiva, vedi Sezione 4.3) se i casi maggiormente simili al nuovo caso, rispetto agli aspetti suscettibili di influenzare la decisione nell'una o nell'altra direzione (secondo il modello costruito dal sistema), hanno condotto a una certa decisione, la stessa sarà proposta anche nel nuovo caso.

2.2.3 Predizioni e correlazioni

Le risposte di un sistema basato sull'apprendimento automatico sono solitamente chiamate predizioni (*prediction*), ma l'uso di questo termine nel contesto dell'apprendimento automatico differisce dal suo significato usuale.

Innanzitutto, tali "predizioni" non sempre riguardano il futuro. In alcuni casi si tratta di anticipazioni del futuro (per esempio, la predizione della probabilità che il richiedente un prestito sarà in grado di rimborsarlo), ma in altri casi la "predizione" riguarda il presente o il passato: si pensi ad un sistema che "predice" la classificazione del contenuto di un'immagine o la paternità di una sottoscrizione.

Inoltre, la predizione in alcuni casi riguarda un evento suscettibile di verificarsi indipendentemente dalla predizione stessa, in altri casi è invece un suggerimento, che potrà essere accolto o meno da chi può realizzare l'evento stesso. Per esempio, la "predizione" che il giudice deciderà in un certo modo una controversia potrà essere considerata dall'avvocato come la previsione di un esito possibile, dal giudice come un suggerimento da prendere in considerazione ai fini della decisione. La stessa distinzione tra previsione e suggerimento si applica alla predizione circa la concessione di un prestito o altro beneficio: l'indicazione che costituisce una previsione per il cliente o cittadino, diventa un suggerimento per il funzionario competente.

È importante sottolineare che un sistema che effettua predizioni automatiche opera sulla base di correlazioni, cioè di relazioni probabilistiche tra dati di input e esiti possibili. Una correlazione consiste nel fatto che alla presenza di certi dati di input corrisponde una maggiore probabilità di un certo esito (correlazioni positive), o una minore probabi-

lità dello stesso (correlazione negativa). Tali correlazioni sono incorporate nel modello costruito dall'algoritmo di addestramento, che a dati di input correlati positivamente ad un esito favorevole (per esempio, l'elevato patrimonio o livello educativo del richiedente un prestito) tende a far corrispondere una predizione favorevole, e l'opposto nel caso di correlazione negativa. Solitamente gli effetti di tutte le correlazioni rilevanti di cui il sistema può tener conto vengono aggregate in un punteggio (*score*), che esprime la probabilità che nel caso in esame la classificazione sia positiva anziché negativa.

Nel valutare l'impiego di un sistema automatico per effettuare predizioni, bisogna distinguere se i dati contenuti nell'insieme di addestramento siano costituti da scelte passate di esperti umani, o invece da eventi indipendenti da tali scelte. Prendiamo in considerazione, per esempio, due sistemi utilizzati per la valutazione delle richieste di prestito. Il primo sistema ha appreso a valutare tali richieste sulla base di un insieme di addestramento che associa le informazioni su richieste presentate in passato alle decisioni corrispondenti da parte del funzionari competenti (accettazione o rigetto). Il secondo sistema invece usa un insieme di addestramento che associa le richieste accolte all'esito del prestito (restituzione o mancata restituzione). Nel primo caso, il sistema apprende a predire le decisioni che i funzionari della banca avrebbero dato in circostanze analoghe; nel secondo caso, il sistema apprende a predire la realizzazione del risultato desiderato (la restituzione dei prestiti concessi). Nel primo caso, il sistema riproduce le virtù (accuratezza, imparzialità, equità) e i vizi (imprecisioni, pregiudizi, iniquità) dei funzionari; nel secondo caso esso anticipa in modo più obiettivo gli esiti desiderati o temuti.

Considerazioni analoghe potrebbero farsi per un sistema destinato ad operare nell'ambito della giustizia, ad esempio, per determinare se concedere o meno la libertà vigilata. Il sistema potrebbe essere addestrato: (*a*) sulla base di un insieme di addestramento che associa richieste di libertà vigilata alle corrispondenti decisioni dei giudici o (*b*) sulla base di un insieme di addestramento che associa le stesse richieste di libertà vigilata al comportamento successivo dell'imputato (indicando se questi si sia comportato correttamente o, invece, si sia sottratto alla pena o abbia reiterato il reato).

2.2.4 L'apprendimento supervisionato: Esempi

Quale semplice esempio di apprendimento supervisionato, la Figura 2.4 mostra un piccolo insieme di addestramento (alcune decisioni in tema di libertà vigilata) e l'albero di decisione che può essere costruito (appreso) automaticamente sulla base di quell'insieme di addestramento (si tratta di un albero a orientamento invertito, nel quale dalla radice in alto si dipartono i rami verso il basso). L'albero di decisione riprende e generalizza l'informazione implicita nell'insieme di addestramento nella forma di una combinazione di test, da effettuarsi sequenzialmente, procedendo sui rami dei rami dell'albero, dall'alto verso il basso. Il primo test chiede se l'imputato sia stato implicato in un reato correlato alla droga. Se la risposa è positiva abbiamo raggiunto il fondo dell'albero (una sua fo-

Insieme di addestramento					
Predittori					Esito
Caso	Lesione	Droga	Arma	Precedenti	Decisione
1	Nessuna	No	No	Sì	Sì
2	Grave	Sì	No	Gravi	No
3	Nessuna	No	Sì	No	Sì
4	Grave	Sì	No	Sì	No
5	Leggera	Sì	Sì	Sì	No
6	Nessuna	Sì	Si	Gravi	No
7	Nessuna	No	Sì	Sì	No

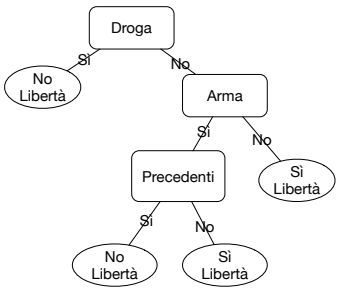

Figura 2.4: *Insieme di addestramento e albero di decisione*

glia) con la conclusione che la libertà vigilata è negata. Se la risposta è negativa (non si tratta di droga), si passa al secondo test, che chiede se l'imputato abbia usato un'arma, e così via.

Si osservi che l'albero di decisione non include informazioni riguardanti il tipo di lesione causata dal reato. Ciò è dovuto al fatto che tutte le decisioni contenute nell'insieme di addestramento possono essere spiegate senza far riferimento a questa informazione: applicando l'albero di decisione otteniamo, per ciascun caso nell'esempio, la decisione adottata dai giudici in quel caso. L'algoritmo per la costruzione dell'albero di decisione ha costruito il modello esplicativo più semplice tra quelli che consentono di spiegare tutti i casi, omettendo le informazioni inutili a tal fine.

Nella Figura 2.4 possiamo chiaramente distinguere l'input e l'output dell'addestramento. La tabella è l'insieme di addestramento, e l'albero di decisione è il modello. L'algoritmo discente è il software che costruisce l'albero di decisione a partire dalla tabella. L'albero di decisione codifica la logica (così come ricostruita dal software addestratore) delle decisioni umane rappresentate nell'insieme di addestramento. L'algoritmo predittore è il software che predice la decisione di nuovi casi usando l'albero di decisione. Grazie all'albero di decisione è possibile non solo indicare un risultato (rilascio o detenzione), ma anche fornire una spiegazione, cioè indicare le ragioni che sostengono quel risultato. A tal fine basta ripercorrere il cammino che ha condotto al risultato. Assumiamo, per es., che nel caso in esame vi siano gravi lesioni, ma non vi siano droga, né armi, né precedenti penali. La risposta sarà negativa, e la spiegazione sarà costituta dal fatto che nel caso in esame non vi sono né droga né precedenti.

Nell'esempio appena riportato, l'albero di decisione riproduce la logica dei decisori umani le cui decisioni sono comprese nell'insieme di addestramento, e riflette quindi vizi e virtù di quei decisori. Per esempio, nell'albero di decisione, il fatto che l'imputazione riguardi un reato collegato agli stupefacenti è sufficiente perché sia negata la libertà vigilata. Ci potremmo chiedere se questo criterio di decisione sia imposto dal diritto, o

corrisponda a una scelta discrezionale. In questa seconda ipotesi ci possiamo chiedere se esso sia equo o rifletta dei pregiudizi dei decisori rispetto a chiunque sia coinvolto nel traffico illegale di stupefacenti (perché chi è coinvolto in un reato che comporta l'uso di armi, in assenza di precedenti penali, è lasciato in libertà solo se non vi è collegamento con gli stupefacenti?).

Inoltre, si osservi che l'albero di decisione fornisce risposte anche rispetto a casi che non coincidono esattamente con alcun esempio compreso nell'insieme di addestramento. Per esempio, l'insieme di addestramento non comprende alcun caso di un reato connesso all'uso di droga nel quale non siano state usate armi e né ci siano precedenti penali. Tuttavia, il sistema fornisce una risposta anche in questo caso: la libertà vigilata dovrebbe essere negata, poiché ciò accade in tutti i casi compresi nell'insieme di addestramento in cui ci sia uso di droga. Ci si può chiedere se questa generalizzazione sia preferibile secondo criteri sociali, etici e anche giuridici. In un caso in cui ci sia uso di droghe ma non vi siano armi né lesioni, potrebbe forse essere preferibile concedere la libertà vigilata. Inoltre, si consideri come la predizione del sistema si basi solo su certe caratteristiche dei casi, quelle usate per descrivere i precedenti (la presenza di lesioni, droga, armi, e precedenti penali). Altri aspetti potenzialmente rilevanti per determinare l'opportunità di concedere la libertà vigilata —come il carattere dell'imputato, le sue connessioni con ambienti criminali, o il suo comportamento nel processo— non sono considerate

L'esempio illustra come i suggerimenti forniti da un sistema automatico si basino sempre su generalizzazioni estratte da casi simili. I critici delle decisioni automatizzate affermano che tali decisioni non possono quindi rispondere al requisito che ogni decisione su un caso penale sia individuale, si fondi cioè sulle caratteristiche specifiche del caso in esame. Anche i sistemi più avanzati si limiterebbero a mettere nello stesso secchio i casi simili (cosiddetto *bucketing*) senza considerare sul serio le circostanze di ogni caso: la decisione su una particolare persona si baserebbe solo sull'appartenenza di quella persona a una certa classe. Tale classe ricomprenderebbe, infatti, tutte le persone che, nell'insieme di addestramento, sono assimilate all'individuo in questione, condividendo con lo stesso le caratteristiche emerse come significative nei precedenti già considerati (per esempio, il fatto di essere stati coinvolti in reati attinenti al possesso o traffico di stupefacenti). All'individuo coinvolto nel nuovo caso verrebbe associata una predizione (il diniego della libertà vigilata) sulla base del fatto che in passato a persone simili a lui (anch'esse coinvolte nel traffico di stupefacenti) tale libertà è stata negata. Analoghe considerazioni si possono applicare a decisione automatizzate, sfavorevoli all'interessato, in altri ambiti, come la mancata assunzione a un posto di lavoro, o il diniego di un mutuo.

A questa critica si può rispondere osservando che le proposte del sistema diventano più individualizzate quanti più sono i predittori che esso prende in considerazione e quanto più ampio il suo insieme di addestramento. Inoltre, anche i giudizi umani si

Caso	Lesione	Droga	Arma	Precedenti	Decisione
8	Nessuna	No	No	Sì	No

Figura 2.5: *Un nuovo caso*

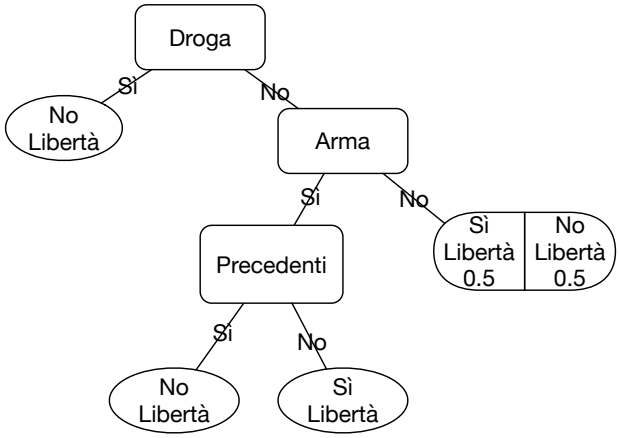

Figura 2.6: *Albero di decisione con incertezza*

basano su generalizzazioni basate su esempi passati, in cui il giudicante è stato coinvolto o di cui ha avuto notizia, ragion per cui l'assoluta individualizzazione sfuggirebbe anche al decisore umano.[82]

Un'altra critica può far riferimento alla possibilità che i dati stessi siano incoerenti, non in quanto errati, ma perché riflettono un contrasto reale. Come potrebbe un modello che fornisce un'unica soluzione a ciascun quesito propostogli, dare un'immagine adeguata di una realtà caratterizzata da conflitti di giudizi, come tipicamente accade in ambito giuridico? Immaginiamo, per esempio, che nell'insieme di addestramento sia inserito un nuovo caso, con valori identici al primo caso della lista, ma con decisione opposta (Figura 2.5). Per rappresentare questo conflitto nell'albero di decisione, potremmo specificare che il ramo Droga=no, Arma = no dà luogo a decisioni alternative, ciascuna associata a una probabilità che corrisponde alla proporzione delle decisioni corrispondenti (Figura 2.6).

Usando questo albero di decisione, il sistema potrà dare, a chi chieda una predizione per il caso in cui una persona abbia commesso un reato in cui non vi siano stupefacenti né armi, la previsione che vi è il 50% di probabilità per il rilascio e 50% per la detenzione.

È possibile altresì estrarre regole da un insieme di addestramento —dette regole di associazione (*association rule*)— e valutarne l'affidabilità. Nel caso del nostro albero di

decisione, ogni cammino che parte dalla radice dell'albero e giunge fino alla decisione corrisponde ad una regola siffatta, per esempio:

1. SE Droga=sì, ALLORA Decisione= no

I due criteri più significativi per valutare una regola estratta da esempi sono quelli del supporto e della confidenza:

- il supporto è il rapporto tra il totale dei casi e quelli in cui siano soddisfatti sia l'antecedente sia il conseguente della regola;
- la confidenza è il rapporto tra i casi in cui è soddisfatto l'antecedente e quelli cui, oltre all'antecedente, è soddisfatto anche il conseguente

Il supporto ci dice, quindi, in quale proporzione di casi, tra tutti quelli compresi nella base di dati, la regola risulta applicabile e soddisfatta (vi verificano tanto l'antecedente quanto il conseguente); la confidenza ci dice invece in quale proporzione di casi, tra quelli in cui la regola risulta applicabile (ha luogo l'antecedente), la regola risulta soddisfatta (ha luogo il conseguente).

Rispetto al dataset riportato nella Figura 2.4, esteso con il nuovo caso della Figura 2.5, la regola SE Droga=sì, ALLORA Decisione= no ha supporto 0,5 (metà dei casi riguardano droga e in essi viene negata la libertà), e confidenza 1 (in tutti i casi in cui ci sia droga, la libertà viene negata).

Invece tanto la regola SE Droga=no, Armi=no ALLORA Decisione=Sì, quanto la regola SE Droga=no, Armi =no ALLORA Decisione=No hanno un supporto pari a 1/8=0,15. Infatti, solo un caso su 8 ha i valori della prima regola (Droga=no, Armi=no, Decisione=sì) e lo stesso vale per valori della seconda (Droga=no, Armi=no, Decisione=No). Inoltre, entrambe le regole hanno una confidenza di 0,5, poiché tra i casi che soddisfanno l'antecedente comune ad esse (Droga=no, Armi=no), una metà (uno su due) ha il valore del conseguente della prima (Decisione=sì), e una metà ha il valore del conseguente della seconda (Decisione=no).

Per un ulteriore esempio di apprendimento supervisionato, si consideri l'insieme di addestramento riportato nella Figura 2.7. Anche in questo caso l'algoritmo di apprendimento ha fornito delle generalizzazioni discutibili, quali le regole secondo cui il prestito deve sempre essere negato ai giovani (indipendentemente dal loro reddito) e concesso agli anziani. Si noti che nel caso di un giovane ad alto reddito —caso non contenuto nell'insieme di addestramento— le regole estratte generano conclusioni contraddittorie. Usualmente, per ottenere decisioni sufficientemente accurate, un sistema deve prendere in considerazione un numero adeguato di fattori e il suo insieme di addestramento deve comprendere un insieme molto ampio di esempi.

Insieme di addestramento			
Predittori			Esito
Nome	**Età**	**Reddito**	**Decisione**
Marco	Giovane	Basso	No
Luisa	Giovane	Basso	No
Antonio	Media	Alto	Sì
Anna	Media	Basso	No
Giuseppe	Avanzata	Medio	Sì
Carla	Avanzata	Alto	Sì

Regole
SE Età=Giovane ALLORA Decisione = No
SE Reddito=Alto ALLORA Decisione =Sì
SE Età=Media E Reddito = Basso ALLORA Decisione = No
SE Età=Avanzata ALLORA Decisione= Sì

Figura 2.7: *Apprendimento supervisionato: richieste di prestito*

2.2.5 Le tecnologie per l'apprendimento automatico: sistemi trasparenti e opachi

L'apprendimento automatico usa diversi metodi: gli alberi di decisione, la regressione statistica, le macchine a vettori di supporto (*support vector machine* (vedi Paragrafo 4.2.1), gli algoritmi evolutivi, le reti neurali, ecc. Tali metodi differiscono non solo nelle prestazioni predittive ma anche nella capacità di fornire spiegazioni, e spesso vi è una tensione tra i due obiettivi: i sistemi che forniscono le prestazioni più accurate sono più opachi, cioè meno capaci di giustificare le proprie decisioni.

Alcuni dei modelli appresi in modo automatico sono caratterizzati dal fatto che il loro funzionamento non si basa su una sequenza di passi ciascuno dei quali collega premesse dotate di significato a conclusioni sostenute da tali premesse (come nei sistemi basati su regole o gli alberi di decisione). Invece, l'attivazione di quei modelli comporta calcoli complessi intesi a riprodurre le correlazioni statistiche tra caratteristiche di input e risultati da predire. Oggi il modello più influente è rappresentato probabilmente dalle reti neurali che presenteremo nel paragrafo seguente.

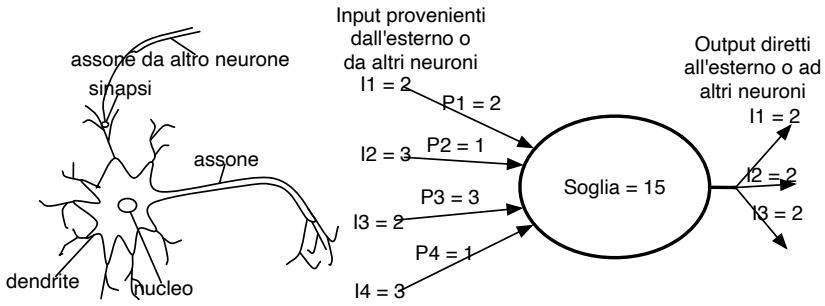

Figura 2.8: *Neurone naturale e neurone artificiale*

Le reti neurali. Le reti neurali sono sistemi informatici che consistono di nodi (i cosiddetti neuroni) collegati da link ai quali sono assegnati pesi numerici. Come si è osservato nella Sezione 1.4.4, la ricerca sulle reti neurali è ispirata all'idea che l'intelligenza possa ottenersi riproducendo l'hardware cerebrale (i neuroni) piuttosto che riproducendo i processi del pensiero cosciente (il ragionamento). Conseguentemente, tale ricerca assume che il comportamento intelligente risulti innanzitutto dall'adattamento flessibile all'ambiente (di cui sono capaci tutti gli animali dotati di un sistema nervoso), piuttosto che dal ragionamento.

La struttura informatica utilizzata per riprodurre questo tipo di apprendimento è la rete neurale, una struttura composta da unità chiamate *neuroni*, e da collegamenti tra le stesse unità. I neuroni artificiali si ispirano alla struttura dei neuroni presenti nel nostro cervello, come risulta dal confronto tra neurone naturale e artificiale riportato nella Figura 2.8. Nelle reti neurali realizzate mediante software (anziché mediante dispositivi elettrici, interruttori e fili), neuroni e collegamenti non sono oggetti materiali ma strutture informatiche: i neuroni sono oggetti software che reagiscono ai dati di input a essi forniti inviando possibilmente degli output ai neuroni ad essi collegati. Ai collegamenti sono assegnati pesi, cioè coefficienti secondo i quali i segnali passanti attraverso i collegamenti stessi sono amplificati o ridotti. [83]

Il funzionamento di ogni neurone è stabilito da funzioni logico-matematiche. Il neurone, quando riceve determinati segnali (valori numerici), verifica se quei segnali abbiano raggiunto il livello (la soglia) richiesta per la propria attivazione. Se il livello non è stato raggiunto, il neurone rimane inerte; se invece il livello è stato raggiunto, il neurone si attiva, inviando a sua volta segnali ai neuroni con esso connessi. Per esempio, nell'immagine riportata a destra nella Figura 2.8 al neurone sono inviati stimoli di valore 2, 3, 2 e 3. I pesi applicati a tali stimoli sono rispettivamente 2, 1, 3, e 1. Moltiplicando gli stimoli inviati al neurone per i relativi pesi, si ottengono gli input forniti al neurone: $2 * 2 = 4$; $3 * 1 = 3$; $2 * 3 = 6$; $3 * 1 = 3$. Il valore ottenuto sommando gli input

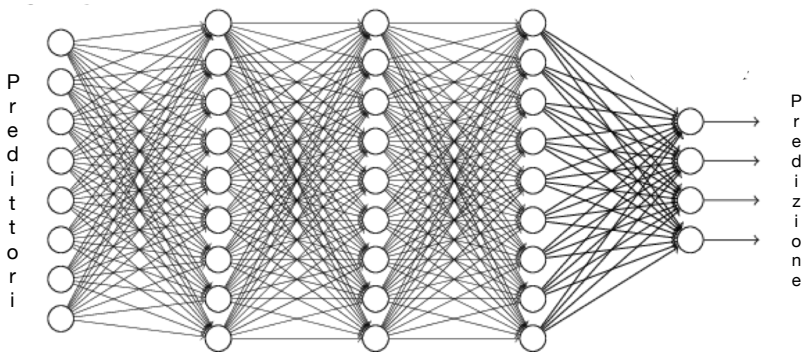

Figura 2.9: *Rete Neurale*

$(4+3+6+3)$, cioè 16, è al disopra della soglia (15) del neurone, che conseguentemente si attiverà, inviando messaggi di valore 2 ai neuroni a esso collegati.

Combinando i neuroni otteniamo una rete neurale. Come risulta dalla Figura 2.9, alcuni neuroni della rete ricevono input dall'esterno (ad esempio da una base di dati, da una telecamera che riceve immagini dall'ambiente, da una tavoletta sulla quale si tracciano disegni o caratteri); alcuni neuroni, i cosiddetti neuroni nascosti, sono collegati solo con altri neuroni; alcuni neuroni, infine, inviano il loro output all'esterno della rete (nella Figura 2.9, per semplicità si è omessa l'indicazione dei pesi, tranne che nel primo livello, per non appesantire l'immagine).

La tecnica più comune per addestrare una rete neurale consiste nel proporre alla rete una serie di esempi corretti, cioè un serie di coppie < input, output>, dove l'output indica il risultato corretto per l'input corrispondente. Per esempio, se vogliamo addestrare una rete nel riconoscimento dei caratteri alfabetici ogni esempio consisterà di un segno grafico unito al carattere corrispondente.

L'elaborazione degli esempi avviene nel modo seguente. Il sistema determina la propria risposta rispetto all'input indicato nell'esempio. Se la risposta differisce dall'output nell'esempio, la rete si riorganizza, cambiando la propria configurazione (i collegamenti o i pesi associati a essi) in modo da poter dare la risposta corretta (la stessa indicata nell'esempio) di fronte alla riproposizione dello stesso input. Dopo un conveniente addestramento, la rete acquista l'abilità di dare risposte corrette non solo nei casi contenuti nell'insieme degli esempi, ma anche in casi analoghi.

Quale sia l'input, quale l'output, quale l'insieme di addestramento di una rete neurale, dipende dalla funzione che si vuole affidare alla rete. Nel caso di una rete destinata alla classificazione di immagini, l'input sarà costituto dai punti di colore dai quali è costituita l'immagine, l'output dalla classificazione dell'immagine (per esempio, come

immagine di uomo, una donna, o qualche animale), l'insieme di addestramento da un insieme di immagini etichettate (come uomo, donna, cane, gatto, ecc.). Nel caso di rete destinata a classificare dei testi, l'input potrà essere costituito da una rappresentazione numerica del testo, l'output dalla classificazione del testo in base al suo contento, all'attitudine che esprime, o alla sua funzione argomentativa; l'insieme di addestramento sarà costituito da un insieme di testi, etichettati con le caratteristiche rilevanti del loro contenuto.

Quando il caso da valutare è sottoposto alla rete, le caratteristiche del caso sono tradotte in valori numerici. Per esempio, ogni frase di una sentenza può essere trasformata in un vettore di numeri binari —una sequenza di cifre 1 o 0, una per ciascuna parola del vocabolario— che indica quali parole siano presenti o assenti nella frase: uno 0 nella posizione dedicata ad una certa parola indica che l parola è assente nella frase, mentre un 1 indica che essa vi compare (per un esempio di questa rappresentazione, la cosiddetta "borsa di parole" (*bag of words*, vedi Sezione 4.2.1). Allo stesso modo, possono essere codificate informazioni sulla presenza o assenza delle caratteristiche rilevanti di un caso (per esempio, se vi siano droga, armi, ecc.). Questi valori numerici vengono acquisiti dai neuroni di input e trasmessi attraverso i link ai neuroni connessi (dopo essere stati aumentati o diminuiti a seconda dei pesi delle connessioni). Ogni neurone applica delle funzioni matematiche ai valori ricevuti, e determina di conseguenza se e quali valori saranno trasmessi ad altri neuroni. Il risultato finale —l'output fornito dalla rete— è determinato dall'insieme delle interazioni tra i neuroni.

Qui non possiamo approfondire l'esame delle tecnologie delle reti neurali. Ci basta osservare che le reti neurali, effettuano un'elaborazione "subsimbolica". Una rete non consiste di simboli —quali entità linguistiche che, come i vocaboli del linguaggio umano, esprimono concetti e fanno riferimento a certi tipi di oggetti— ma piuttosto di vettori di numeri. Le elaborazioni di quei numeri determinano tanto l'apprendimento della rete, quanto la sua risposta a nuovi casi.

Per esempio, se si vuol creare una rete neurale che preveda l'esito di casi giudiziari (Sezione 4.3), bisogna innanzitutto identificare quali caratteristiche dei casi in esame possano essere rilevanti per l'esito di quei casi e stabilire come rappresentare numericamente quelle caratteristiche. Bisogna quindi costruire una rete intesa a collegare (attraverso più livelli di nodi intermedi) quelle caratteristiche, rappresentate nei nodi di input agli esiti possibili, rappresentati dai nodi di output. La rete è addestrata su casi reali e ipotetici (l'insieme di addestramento) finché —rispetto ai casi contenuti in tale insieme— non fornisca risultati corretti, cioè quando, date le caratteristiche di un esempio, la rete risponde con l'esito associato a quell'esempio. La rete addestrata può essere applicato a nuovi casi, di cui ancora non si conosca l'esito. La rete stessa (il software che l'attiva) risponderà alla descrizione del nuovo caso suggerendo l'esito dei casi passati più simili al nuovo caso, rispetto alle caratteristiche correlate all'esito stesso (quelle che concorrono a determinarne la probabilità).

Scatole nere e spiegazione. Si noti che, a differenza di un albero di decisione, una rete neurale non si presta a fornire direttamente spiegazioni comprensibili. È possibile determinare in quale modo il sistema abbia raggiunto una certa decisione, esaminando come l'input fornito abbia determinato l'attivazione di certi neuroni, e come l'output di questi abbia determinato l'attivazione di altri neuroni (in base ai pesi assegnati alle connessioni neuronali in seguito all'addestramento della rete). Tuttavia, questa informazione non rende esplicite le ragioni per le quali è stata data una certa risposta, non è una giustificazione comprensibile dalla mente umana. La difficoltà di comprendere il funzionamento di una rete cresce all'aumento del numero dei nodi, dei livelli in cui essi sono organizzati e della complessità delle loro connessioni. Si possono così realizzare reti per il cosiddetto apprendimento profondo (*deep learning*), in grado di apprendere anche da dati non strutturati. Le reti neurali, come altri modelli del cui funzionamento non si riesca a dare spiegazioni, sono dette "opache" o anche "scatole nere" (*black box*).

Numerose ricerche si propongono di realizzare tecnologie che consentano spiegazioni comprensibili all'uomo anche per l'attività di tali sistemi, ma i risultati di tali ricerche sono ancora molto limitati. Di qui la necessità di bilanciare efficienza (accuratezza nelle predizioni) e spiegabilità nella scelta del sistema predittivo da utilizzare in un determinato ambito. In molti casi (per esempio, nei sistemi robotici o nella diagnosi medica o manutentiva) è tendenzialmente l'efficienza a prevalere: preferiamo un sistema che commetta meno errori rispetto a un sistema più fallibile, ma capace di giustificare le proprie determinazioni adducendo ragioni a sostegno di tali determinazioni. L'esigenza di fornire spiegazioni è, invece, fondamentale quando il sistema sia usato per l'esercizio di funzioni di rilevanza pubblica, specialmente quando siano in gioco interessi in contrasto e le esigenze del controllo siano preminenti, come accade rispetto all'azione amministrativa e ancor più nell'ambito della giurisdizione.

Esistono numerose ricerche che sviluppano diversi indirizzi per fornire tecnologie per ottenere spiegazioni comprensibili all'uomo del funzionamento di tali sistemi. Alcuni orientamenti di ricerca guardano all'interno del sistema il cui comportamento si vuole spiegare (per es., guardano ai risultati ottenuti dai diversi livelli della rete e cercando di capire come essi si collegano a caratteristiche dell'input). Altri adottano una prospettica estrinseca rispetto allo sviluppo delle spiegazioni. Si accetta l'idea che il nucleo del sistema sia una scatola nera, che dà risposte fornire spiegazioni. Ad esso si affianca un secondo sistema, che ha la funzione di razionalizzare per le risposte del sistema, adducendo possibili ragioni e argomenti a sostegno di tali risposte (senza che ragioni e argomenti influiscano sul funzionamento del primo sistema) .

La capacità di fornire spiegazioni ai risultati forniti da sistemi opachi è per ora assai limitata e in molti ambiti le maggiori prestazioni sono offerte dai sistemi opachi. Anche quando un sistema si configuri come una scatola nera, il cui interno sia imperscrutabile, alcune analisi critiche del suo comportamento sono possibili. In particolare, è possibile determinare quali predittori abbiano determinato il risultato fornito dal sistema mediante

analisi di sensibilità, cioè controllando sistematicamente quali cambiamenti dei valori dei predittori comportino quali cambiamenti negli output del sistema.

Ad esempio, per ottenere una spiegazione del risultato fornito da un sistema destinato a valutare l'affidabilità creditizia di chi richiede un mutuo, potremmo controllare come la valutazione del sistema cambi modificando la posizione lavorativa, il reddito, il patrimonio o anche la residenza o il luogo di nascita dell'interessato. Per esempio, assumiamo che un sistema per la valutazione di richieste di mutuo abbia dato una risposta negativa a un richiedente che abbia dichiarato un reddito di 2.000 euro mensili. Sostituiamo quindi il valore del reddito indicato dal richiedente (2.000) con valore più elevato, per esempio, 4. 000, lasciando gli altri dati immutati (*ceteris paribus*), e sottoponiamo nuovamente il caso, così modificato, al sistema. Se la risposta del sistema diventa positiva possiamo concludere che una spiegazione del diniego del prestito consiste nel basso reddito del richiedente (si parla, in questi casi, di spiegazione "contrastiva").

In taluni casi, la natura del dato rilevante (il cui cambiamento fa cambiare il risultato) può condurci a conclusioni significative anche per una valutazione giuridica. Se per esempio la residenza, invece che il reddito, risultasse determinante (cambiando il quartiere di residenza, si passa dal diniego del prestito alla sua concessione), ci potremmo chiedere se il sistema operi in modo arbitrario, o addirittura discrimini iniquamente gli individui a secondo della loro etnia o condizione sociale, che possono essere statisticamente correlate alla residenza o al luogo di nascita.

2.2.6 L'integrazione di rappresentazione della conoscenza e apprendimento automatico

L'apprendimento automatico si avvale di tecniche "guidate dai dati", che nella maggior parte dei casi non conducono ad una rappresentazione della conoscenza comprensibile all'uomo. Un esempio paradigmatico è rappresentato dalle reti neurali nelle quali, come abbiamo visto, la conoscenza è rappresentata dai nodi della rete (con le relative funzioni di attivazione) e dalle connessioni (con i relativi pesi) tra i nodi stessi.

Tuttavia, ciò non significa che i metodi per la rappresentazione della conoscenza e per il ragionamento abbiano perso rilevanza. Infatti, in molti settori, i modelli logici possono essere complementari all'apprendimento automatico. Questi modelli possono contribuire a spiegare il funzionamento dei sistemi per l'apprendimento automatico, controllare e governare il loro funzionamento secondo standard normativi (regole etiche, giuridiche, o ispirate da ragioni di opportunità), validarne i risultati, e sviluppare le implicazioni logiche degli stessi risulti secondo conoscenza concettuale e teorie scientifiche. Nella comunità dell'intelligenza artificiale la necessità di combinare in molti casi modellazione logica e apprendimento automatico è riconosciuta da molti, benché vi siano diverse opinioni su come raggiungere questo obiettivo e sugli aspetti da coprire con i la modellazione logica o con metodi di apprendimento, trasparenti o opachi.[84]

Capitolo 3

L'IA: opportunità, rischi e norme

Questo capitolo affronta gli aspetti sociali dell'uso dell'intelligenza artificiali: le opportunità che essa offre, i rischi che comporta, e le norme etiche e giuridiche che ne possono disciplinare l'utilizzo. In particolare, si considerano le decisioni automatiche, con riguardo ai temi dell'equità e della non-discriminazione.[85] Si esaminano i rischi derivanti dall'uso dell'IA per la profilazione, l'influenzamento e la manipolazione degli individui. Infine si considera il rapporto tra AI e le norme sociali, etiche e giuridiche.

3.1 Dati e predizioni

Come si è visto nella Sezione 2.2.2, predire significa passare da aspetti conosciuti di un caso —una persona, un oggetto, un evento, o una situazione–, i valori dei cosiddetti predittori (anche chiamati variabili indipendenti o caratteristiche, *features*), a un aspetto non conosciuto dello stesso caso, l'obiettivo da predire (chiamato anche variabile dipendente o etichetta). Negli ultimi anni si è assistito ad un impiego sempre più ampio di sistemi informatici a fini predittivi. L'uso di tecniche per l'apprendimento automatico è divenuto preminente, venendosi così a creare una sinergia tra raccolta di dati (finalizzata alla creazione automatica di modelli predittivi) e applicazioni basate sull'apprendimento automatico.

3.1.1 Predizioni e apprendimento automatico

L'uso di sistemi informatici per anticipare eventi e comportamenti futuri può svolgersi in forme diverse, grazie a diverse tecnologie.

Sistemi informatici per elaborazioni statistiche sono disponibili da tempo. Tali sistemi sono largamente utilizzati per la valutazione predittiva di casi individuali, in settori quali l'assicurazione e il credito. Per esempio, metodi statistici possono essere usati per determinare la probabilità che un individuo possa decedere in un certo arco di tempo, o possa non essere in grado di restituire il credito richiesto.

Anche un sistema basato su regole e concetti può essere usato a fini predittivi. Se al sistema si fornisce la descrizione di un nuovo caso, il sistema potrà applicare la sua base di conoscenza a tale descrizione, e trarne conclusioni interpretabili come predizioni.

Per esempio, si immagini un cittadino interroghi un sistema basato sulla conoscenza per sapere se egli abbia diritto ad una certa prestazione previdenziale. La risposta del sistema potrà essere intesa dallo stesso cittadino come la predizione del contenuto di una decisione futura da parte dei decisori competenti. La predizione sarà affidabile nella misura in cui (a) le regole e concetti nella base di conoscenza del sistema corrisponda alla comprensione e interpretazione del diritto da parte di quei decisori e (b) la descrizione del caso corrisponda alla qualificazione dello stesso che potrebbe essere compiuta dagli stessi decisori.

Tuttavia, nel corso degli ultimi decenni, si è assistito ad una importante evoluzione: le funzioni predittive sono state affidate in misura sempre maggiore a sistemi basati sull'addestramento automatico. Le predizioni di questi sistemi, come osservato nella Sezione 2.2.2, si fondano su modelli, costruiti automaticamente, che collegano i valori dei predittori all'obiettivo da predire. Per esempio, nel campo medico, il modello appreso potrà collegare sintomi (test diagnostici) a possibili patologie. Nel campo della pubblicità online il modello potrà collegare aspetti comportamentali (gli amici nelle reti sociali, i "like" espressi, i siti visitati, e così via) a certe preferenze e propensioni rispetto a possibili acquisti. Nel campo del credito, il modello potrà collegare caratteristiche del richiedente un prestito alla probabilità che il prestito sarà restituito nei termini o all'opportunità di concedere il prestito stesso.

Grazie all'integrazione delle risorse informatiche oggi disponibili —tecniche avanzate di IA, enormi masse di dati, e grandi potenze di calcolo— è possibile fondare le previsioni e valutazioni automatiche su grandi insiemi di esempi, ciascuno dei quali può comprendere informazioni dettagliate. Di conseguenza si sono potuti realizzare sistemi in grado di compiere predizione con livelli di accuratezza elevati, o comunque adeguati al contesto di utilizzo.

Per esempio, la pubblicità mirata si può basare su esempi (raccolti in un dataset) ciascuno dei quali collega le caratteristiche e il comportamento di un consumatore (sesso, età, condizione sociale, acquisti pregressi, pagine visitate nel web, ecc.) alle sue risposte ai messaggi pubblicitari. La valutazione di una domanda di assunzione può fondarsi su esempi ciascuno dei quali unisce le caratteristiche di una persona (educazione, impieghi, test attitudinali) all'esito della procedura di assunzione che ha riguardato la stessa persona (e possibilmente alla valutazione delle sue successive prestazioni lavorative). La predizione della propensione al recidivismo può basarsi su esempi ciascuno dei quali combina le caratteristiche di un condannato (educazione, impieghi, condizione familiare, test psicologici, ecc.) con l'indicazione se lo stesso abbia reiterato il reato o no. La diagnosi di patologie o l'indicazione di terapie personalizzate può basarsi su esempi ciascuno dei quali collega le caratteristiche di un paziente ai risultati dei test cui lo stesso è

stato sottoposto, alle corrispondenti condizioni mediche o all'esito di possibili terapie.

Si noti che tutti i sistemi basati sull'apprendimento automatico sono fallibili: le loro inferenze hanno un grado di incertezza poiché non si può escludere che un nuovo caso, pur assimilabile a casi precedenti rispetto alle caratteristiche considerate dal sistema, non si conformi alla predizione. Per esempio, anche quando il sistema prevede che il richiedente un mutuo sarà un debitore affidabile, ciò potrebbe non verificarsi (per esempio, in seguito a un improvviso dissesto finanziario del richiedente stesso). Bisogna valutare caso per caso, se il livello di accuratezza fornito dal sistema sia adeguato rispetto allo scopo per il quale è utilizzato e agli interessi in gioco.

In altri ambiti, come quello medico, si potrà richiedere un'accuratezza elevata. Per esempio, nella diagnostica si dovrà richiedere che siano minimizzati i falsi positivi, poiché la mancata individuazione di una patologia può avere effetti gravissimi per il paziente.

In altri casi prestazioni funzionalmente adeguate si possono anzi ottenere con un basso livello di accuratezza. Per esempio, in un sistema che predice la propensione agli acquisti al fine dell'invio di pubblicità mirata, la predizione che il destinatario del messaggio cliccherà sul link pubblicitario e poi acquisterà il prodotto, può giustificare l'invio della pubblicità mirata anche se la probabilità è molto inferiore al 50%.

3.1.2 Sinergia tra dati e IA

A causa del bisogno di apprendere esaminando grandi masse di dati (gli esempi da cui estrarre correlazioni e generalizzazioni), l'IA è "affamata" di dati, e questa fame ha stimolato raccolte di dati sempre più estese. A sua volta, la disponibilità di grandi raccolte di dati ha stimolato nuove o più ampie applicazioni dell'apprendimento automatico. Si è così generata una rincorsa tra dati e IA, che ha favorito la crescita di entrambi.

La formazione di grandi masse di dati in formato elettronico ha preceduto, peraltro, lo sviluppo delle applicazioni di IA basate sui dati. Infatti, la raccolta di dati elettronici è il naturale sottoprodotto dell'uso di ogni tipo di sistema informatico. Ogni qualvolta un'attività venga mediata dal computer si attiva un duplice flusso: il sistema informatico fornisce dati utili allo svolgimento di quell'attività (per esempio un acquisto online), e contemporaneamente registra dati su di essa (e sul coinvolgimento degli interessati).[86] Grandi quantità di dati sono raccolte ogni secondo dai computer che partecipano a transazioni commerciali (nel commercio elettronico, ma anche negli scambi nel mondo fisico in cui un sistema informatico registri informazioni, emetta ricevute, ecc.), dai sensori che controllano e forniscono dati a sistemi computerizzati (automobili di recente fabbricazione o dispositivi per la "casa intelligente"), dai flussi di lavoro di organizzazioni pubbliche e private (banche, trasporti, gestione delle imposte, scuole, ecc.), e dai sistemi che gestiscono servizi online (accesso ad internet, reti sociali, motori di ricerca, ecc.).

Negli ultimi decenni, questi flussi di dati sono stati integrati in un'infrastruttura globale per l'elaborazione delle informazioni, incentrata su Internet, ma non limitata ad essa. Per esempio, usiamo Internet per accedere ai servizi bancari, mentre le banche utilizzano reti dedicate, come la rete SWIFT, per effettuare transazioni finanziarie.

Le telecomunicazioni oggi coinvolgono circa 30 miliardi di dispositivi elettronici interconnessi —computer, telefoni intelligenti, macchine industriali, telecamere, ecc.— che generano enormi masse di dati. Questa infrastruttura di sistemi informatici e telecomunicazioni è divenuta l'intermediario universale per comunicare, acquisire informazioni, accedere a servizi pubblici o privati. Essa consente ai cittadini di fare acquisti, accedere a servizi finanziari, pagare le imposte, ottenere servizi pubblici, ecc., e al tempo stesso registra informazioni sulle attività effettuate attraverso di essa. Algoritmi spesso basati sull'IA-mediano l'accesso a contenuti e servizi, selezionano le informazioni da fornire alle persone, stabiliscono quali opportunità (per es., possibili acquisti o offerte di lavoro) presentare loro.

Nell'esaminare il fenomeno della "datificazione" l'attenzione del giurista tende a soffermarsi sui dati personali, cui è dedicata la disciplina della protezione dei dati. Tuttavia, la fame di dati dell'IA si estende ad ogni tipo di informazioni: dati metereologici (da usare per le previsioni del tempo), ambientali (da usare per la valutare lo stato dell'ambiente, prevenire rischi), relativi a processi e prodotti industriali (da usare per individuare prodotti difettosi, anticipare i guasti, epianificare la produzione), e molti altri.

3.2 La predizione automatica

L'uso predittivo di grandi masse di dati, usando tecnologie per l'apprendimento automatico comporta grandi opportunità ma anche rischi da non sottovalutare.

3.2.1 IA e big data: nuove opportunità

Le tecnologie dell'IA, applicate ai big data possono dare grandi benefici: miglior accesso all'informazione; accresciuta generazione e più ampia distribuzione della conoscenza; riduzione dei costi, maggiore produttività e creazione di valore; nuove opportunità di lavoro creativo e ben pagato; servizi pubblici e privati personalizzati, che tengano conto di preferenze e esigenze di ciascuno; gestione sostenibile dell'energia, dei servizi di pubblica utilità e della logistica; nuovi servizi per l'informazione e la consulenza; maggiore trasparenza nelle decisioni pubbliche e private; rimedi contro pregiudizi e discriminazioni.

L'IA può consentire grandi progressi in molti ambiti. Essa può consentire agli scienziati di scoprire, correlazioni, formulare ipotesi, e sviluppare modelli supportati da prove empiriche; ai medici di fare diagnosi più accurate e proporre terapie più efficaci; alle imprese di anticipare le dinamiche del mercato e i propri bisogni, conoscere gli impatti

delle proprie scelte, così da adottare decisioni più efficienti e socialmente responsabili; ai consumatori di fare scelte più informate e ottenere servizi personalizzati; alle autorità pubbliche di anticipare i rischi, ottimizzare la gestione dei beni pubblici (come l'ambiente) e meglio coordinare le attività dei cittadini (il traffico, i consumi di energia, ecc.). Ulteriori progressi, di cui oggi possiamo avere solo una vaga idea, possono realizzarsi in futuro, come afferma Ray Kurzweil, un influente inventore, futurologo, e direttore della ricerca in ingegneria (director of engineering) presso Google:

> Attraverso le tecnologie dell'informazione possiamo affrontare le grandi sfide dell'umanità, come mantenere un ambiente salutare, fornire le risorse per una popolazione crescente (incluse energia, cibo e acqua), vincere le malattie, estendere grandemente la longevità umana, e eliminare la povertà. Solo estendendo noi stessi con tecnologia intelligente possiamo trattare il livello di complessità di cui c'è bisogno.[87]

In alcuni casi, l'IA può sostituire completamente alcune attività umane (per es., nelle automobili senza pilota, nei robot usati per la pulizia dei pavimenti, in alcuni compiti concernenti la pianificazione e la programmazione in logistica). In molti altri casi, invece, l'IA è complementare rispetto alle capacità umane, aumentando le nostre capacità umane di conoscere e operare. Nell'era dell'IA, diventa possibile un nuovo tipo di collaborazione tra umani e macchine, che supera il modello classico nel quale le macchine compiono solo compiti ripetitivi e standardizzati. La possibilità di questa integrazione era già stata anticipata all'inizio degli anni '60 da JK Licklider, uno scienziato che ha svolto un ruolo importante nello sviluppo di Internet. Egli affermava che nel futuro la cooperazione tra esseri umani e computer avrebbe compreso attività creative, cioè

> adottare decisioni e controllare situazioni complesse senza dipendenza inflessibile da programmi predeterminati.[88]

La sfida del futuro, nel mondo del lavoro, è combinare umani e macchine in modi nuovi, che non si limitino ad accrescere l'efficienza, ma contribuiscano a preservare e accrescere la creatività, la competenza, e la soddisfazione dei lavoratori

3.2.2 IA e big data: nuovi rischi

Lo sviluppo dell'IA comporta non solo grandi opportunità ma anche gravi rischi per gli individui, i gruppi e l'intera società. I sistemi di IA e i dati da essi utilizzati possono offrire nuove occasioni per attività illegali: essi possono essere oggetto di attacchi o possono essere strumenti per commettere atti criminali (per esempio, veicoli autonomi possono essere utilizzati per omicidi o atti terroristici, algoritmi intelligenti per frodi o reati finanziari). Anche al di fuori di attività criminali, le possibile ricadute negative dell'IA non debbono essere sottovalutate.

Rischi per il lavoro. Si è osservato come la sostituzione di sistemi intelligenti al lavoro umano può svalutare il lavoro di chi può essere rimpiazzato dalle macchine. Molti rischiano di perdere la "corsa contro le macchine",[89] e quindi di perdere il proprio lavoro o comunque di vedere svalutate le proprie competenze. Di conseguenza, i lavoratori interessati possono essere colpiti da povertà ed emarginazione, e perdere la fiducia in sé stessi e nella società. Si pensi all'impatto delle auto autonome su tassisti e camionisti, o all'impatto dei chatbot (robot software per la conversazione) intelligenti sui lavoratori nei call center. Ciò comporta, l'esigenza di garantire una vita dignitosa anche a chi abbia perso il proprio lavoro in seguito all'automazione, ma anche quella di offrire nuove soddisfacenti opportunità di lavoro, in attività socialmente utili.

Inoltre, un ambiente di lavoro controllato e governato da sistemi informatici —come i magazzini di spedizioni maggiormente automatizzati, nei quali il sistema indica a ciascuno lavoratore quale pacco prelevare e dove depositarlo– può sottoporre il lavoratore a continue sorveglianza, direzione e pressione. In un contesto di questo tipo sono disconosciute fondamentali esigenze della persona: sviluppare competenze in modo da poter operare con efficacia, agire con autonomia secondo la propria iniziativa, essere connessi con altre persone in modo significativo.[90]

Infine, l'IA, consentendo alle grandi imprese digitali di ottenere enormi profitti con forza lavoro limitata, concentra la ricchezza in chi investe in tali imprese e in chi sa meglio progettare e utilizzare le tecnologie da esse realizzate. Ciò favorisce modelli economici in cui "il vincitore prende tutto" (*winner takes all*) sia nei rapporti tra imprese (dove prevalgono posizioni monopolistiche determinate dall'accesso privilegiato o esclusivo a dati e tecnologie) sia nei rapporti tra lavoratori (dove prevale chi è in grado di svolgere funzioni di alto livello, non automatizzabili).[91] Questi impatti negativi sul lavoro non sono tuttavia conseguenze inevitabili dell'IA. La formazione dei lavoratori, modelli d'interazione uomo macchina che enfatizzano e decentrano creatività e iniziativa, misure di redistribuzione e di accesso ai dati e alle tecnologie, possono consentire a tutti di beneficiare dei frutti dell'IA.[92]

Abusi dell'IA nelle attività economiche. Anche al di fuori delle attività chiaramente illegali, attori motivati dal profitto possano usare l'IA per perseguire interessi economici in modi dannosi per gli individui e la società. Le imprese commerciali e i governi, combinando IA e dati, possono sottoporre cittadini, utenti, consumatori e lavoratori a una sorveglianza pervasiva, limitare le informazioni e le opportunità cui gli stessi hanno accesso, e manipolarne le scelte in direzioni contrastanti con i loro interessi.

Gli abusi sono incentivati dal fatto che molte imprese di Internet —come ad esempio le maggiori piattaforme che ospitano contenuti generati dagli utenti— operano in mercati a due o più lati: i loro servizi principali (per esempio, ricerca, gestione di reti sociali e accesso a contenuti) vengono offerti a utenti che non operano per scopi commerciali, ma

i ricavi provengono dagli inserzionisti, o da chi sia comunque interessato a influenzare gli utenti, come nel caso della pubblicità e della propaganda politica personalizzata.[93]

Pertanto, le piattaforme non si limitano a raccogliere le informazioni utili a meglio indirizzare pubblicità personalizzate, ma usano ogni mezzo disponibile per trattenere gli utenti, in modo che essi siano esposti a messaggi pubblicitari o ad altri tentativi di persuasione. Ciò conduce non solo ad una massiva raccolta di dati sugli individui, a danno della privacy, ma anche ad un'influenza pervasiva sul comportamento degli stessi individui, a danno non solo dell'autonomia dei singoli ma anche di interessi collettivi.

Per esempio, la manipolazione del comportamento dei consumatori può condurre all'esclusione di individui e gruppi da scambi e opportunità, e di conseguenza a un cattivo funzionamento dei mercati. Inoltre, algoritmi guidati dalla ricerca del profitto possono convergere su strategie anti-concorrenziali, a danno dei concorrenti ma anche dei consumatori.[94] Ulteriori problemi derivano dalla possibilità che sistemi di IA interferiscano con la formazione dell'opinione pubblica, o adottino a determinazioni inique o discriminatorie (entrambi questi temi saranno esaminato nelle pagine seguenti).

IA e impatti negativi sull'opinione pubblica. Usi pregiudizievoli dell'AI possono emergere anche nel settore pubblico e l'IA può avere effetti negativi sulla dialettica politica. I governi possono servirsi dell'IA non solo per scopi politici e amministrativi legittimi (per es., efficienza, risparmi sui costi, servizi migliorati), ma anche per anticipare e controllare il comportamento dei cittadini in modi che limitano le libertà individuali e interferiscono con il processo democratico. La manipolazione dell'opinione politica, basata sulla profilazione e l'invio di messaggi mirati può condurre alla polarizzazione e all'estremizzazione delle opinioni, e quindi al deterioramento del dialogo pubblico.[95]

La formazione dell'opinione pubblica può essere compressa anche in assenza di un'intenzione antidemocratica, grazie ai modelli economici della rete e alle dinamiche socio-informative che ne emergono. Infatti, la ricerca dei profitti pubblicitari induce le piattaforme a catturare l'attenzione degli utenti, e trattenerli proponendo loro contenuti che ne assecondino i gusti e concordino con le loro opinioni, profittando della propensione alla conferma (*confirmation bias*) che caratterizza la psicologia umana. Tale pratica può condurre alla polarizzazione e frammentazione della sfera pubblica,[96] oltre che alla proliferazione di notizie sensazionalistiche, non verificate o false (*fake news*). L'IA e i big data contribuiscono a questo fenomeno, mediante la generazione di messaggi persuasivi, e mediante la direzione di messaggi personalizzati, atti ad influenzare selettivamente i destinatari.

3.3 Dalla profilazione all'influenza e alla manipolazione

L'IA e i big data, in combinazione con la disponibilità di ampie risorse informatiche, hanno molto aumentato le opportunità di profilazione degli individui, vale a dire la pos-

sibilità di compiere inferenze —classificazioni, previsioni o decisioni— sulla base di dati riguardanti gli stessi. Ciò è in parte dovuto al fatto che l'automazione riduce i costi per la raccolta, l'archiviazione e l'elaborazione di informazioni, aprendo la strada a meccanismi di sorveglianza persistenti e pervasivi.

3.3.1 La raccolta massiva di dati

Grazie all'IA, tutti i dati personali possono essere utilizzati per analizzare, prevedere e influenzare il comportamento umano, un'opportunità che trasforma tali dati in merci dotate di valore. Informazioni che un tempo non erano raccolte o erano scartate (i cosiddetti "dati di scarto", *exhaust data*), sono oggi diventate una risorsa preziosa. Tutte le tracce del comportamento online possono essere utilizzate per addestrare un sistema di IA ad inferire informazioni sulle abitudini, gli interessi, la personalità, le condizioni di salute, e i tratti psicologici degli individui, inclusi gli stati emotivi, i valori, le opinioni politiche e morali, gli orientamenti sessuali.[97]

Grazie ai sensori che tracciano in modo crescente ogni attività umana online e nello spazio fisico, la raccolta e l'analisi dei dati (su dispositivi, sistemi, applicazioni e piattaforme online) hanno assunto connotati di assoluta pervasività e ubiquità. I dati raccolti vengono elaborati attraverso le tecnologie dei big data e dell'IA, cosicché gli individui sono soggetti a sorveglianza, influenza e manipolazione in molti più casi e in molti più contesti, sulla base di un insieme più ampio di caratteristiche personali (che spaziano dalle condizioni economiche e finanziarie allo stato di salute, al luogo di residenza, fino ad includere le scelte di vita personali di ciascuno, il comportamento online e offline, ecc.).

Questa dinamica contribuisce sempre più a plasmare l'economia globale, il flusso di idee e l'accesso alle informazioni.

3.3.2 La profilazione

I dati personali possono essere usati per la profilazione degli individui, cioè per classificarli, valutarli e prevederne i comportamenti. Il termine "profilazione" deriva da "profilare", che in origine significava tracciare una linea, e più in particolare i contorni di un oggetto. Questa è precisamente l'idea alla base della profilazione: espandere le informazioni e i dati disponibili su individui e gruppi, in modo da disegnarne —descriverne o anticiparne— tratti e propensioni.

Un sistema di profilazione stabilisce (prevede) che gli individui con determinate caratteristiche, C1, hanno una certa probabilità di possedere alcune caratteristiche aggiuntive, C2. Un sistema di questo tipo può predire, per esempio, che chi presenta certe caratteristiche genetiche ha una maggiore tendenza a sviluppare il cancro. Un altro sistema può invece predire che chi ha certa istruzione, svolge certi lavori, o appartiene a una certa etnia, ha una maggiore probabilità di adempiere ai propri debiti. In questi casi,

è possibile affermare che il sistema in esame ha profilato il gruppo di quanti possiedono le caratteristiche C1, aggiungendo un nuovo segmento di informazioni alla descrizione (al profilo) del gruppo, vale a dire la probabilità di possedere le caratteristiche aggiuntive C2.

Successivamente, indicando al sistema che un individuo possiede le caratteristiche C1, il sistema stesso sarà in grado di inferire che, con una certa probabilità, quell'individuo possederà anche le caratteristiche aggiuntive C2. Ciò può comportare effetti positivi o negativi per l'individuo, poiché in base alle caratteristiche che gli sono attribuite, l'individuo potrà essere trattato in modo a lui favorevole o invece sfavorevole. Per esempio, nel caso in cui la caratteristica inferita sia la maggiore probabilità di contrarre il cancro, la previsione del sistema potrà fornire la base per effettuare esami e terapie preventive, o invece per un aumento del premio assicurativo.

Una correlazione appresa può anche riguardare la propensione di un individuo a rispondere in determinati modi a certi stimoli. Ciò consente il passaggio dalla previsione all'influenzamento, che può consistere in un benevolo e lecito indirizzamento verso scelte vantaggiose per l'individuo stesso, o invece in forme di manipolazione illegale o immorale. Per esempio, le informazioni inferite dal sistema possono riguardare la propensione a rispondere positivamente a un certo trattamento terapeutico, con un conseguente miglioramento delle condizioni cliniche, o la propensione all'acquisto di un prodotto rispetto a un determinato annuncio pubblicitario o ad una certa variazione di prezzo, o la propensione a rispondere a un certo tipo di messaggio con un mutamento di preferenze o stati d'animo (per esempio, relativamente alle scelte politiche e alle preferenze di voto).

In talune circostanze, la profilazione comporta quindi la possibilità di influenzare e manipolare gli individui, innescando il comportamento desiderato.

Supponiamo per un esempio che un sistema possegga l'informazione che un individuo dotato di certe caratteristiche (per es. un giovane maschio, con carattere estroverso, amante dello sport, interessato al proprio aspetto fisico, ecc.), è propenso a rispondere a messaggi pubblicitari che presentano in un certo modo certi prodotti (per es., integratori alimentari). Il sistema sarà in grado di inferire che, con una certa probabilità, inviando quel determinato messaggio (la pubblicità di un integratore che aumenta la massa muscolare) a quello specifico individuo, egli risponderà adottando il comportamento previsto (acquistando il prodotto).

La nozione di profilazione appena presentata trova riscontro nella seguente, più elaborata, definizione, che introduce il concetto generale usato in ambito informatico e sociologico:

> La profilazione è una tecnica di trattamento (parzialmente) automatizzato di dati personali e/o non personali, finalizzata alla creazione di conoscenza predittiva mediante la scoperta di correlazioni tra i dati e la costruzione di profili, che possono essere poi utilizzati per assumere decisioni. Un profilo

> è un insieme di dati correlati che rappresentano un soggetto (individuale o collettivo). La costruzione di profili è il processo di scoperta di schemi ricorrenti e sconosciuti tra i dati, all'interno di grandi insiemi di dati, che possono essere utilizzati per creare profili. L'applicazione di profili consiste nell'identificazione e rappresentazione di uno specifico individuo o gruppo come corrispondente a un determinato profilo, e nel processo decisionale basato su tale identificazione e rappresentazione.[98]

La più specifica nozione delineata dall'articolo 4 del Regolamento per la Protezione dei dati (GDPR) collega la profilazione alle valutazioni delle decisioni relative agli individui, sulla base di dati personali, escludendo da tale nozione la costruzione di profili di gruppi di individui:

> [L]a "profilazione" [...] consiste in qualsiasi forma di trattamento automatizzato di dati personali che valuti gli aspetti personali relativi a una persona fisica, in particolare per analizzare o prevedere aspetti quali le prestazioni lavorative dell'interessato, la situazione economica, la salute, le preferenze o gli interessi personali, l'affidabilità o il comportamento, la posizione geografica o gli spostamenti, laddove ciò produca effetti giuridici che lo o la riguardano o che influiscono in modo significativo su lui o lei.

Anche quando un sistema automatizzato di valutazione e decisione —basato sulla profilazione— sia imparziale e in linea di principio volto a servire scopi benefici, esso può influire negativamente sugli interessati. La sorveglianza pervasiva, unita a forme persistenti di valutazione e influenza, comporta infatti pressioni che limitano l'autonomia personale e possono pregiudicare il benessere mentale.

3.4 Le decisioni algoritmiche: equità e discriminazione

Oltre ai rischi legati alle violazioni della privacy e della protezione dei dati, la profilazione crea nuovi rischi di stereotipizzazione, diseguaglianza e discriminazione a causa delle classificazioni e delle categorizzazioni su cui essa si basa. Essa può condurre a scelte che compromettono l'interesse dei singoli a un trattamento algoritmico equo e corretto, vale a dire, l'interesse a non essere soggetti a pregiudizi ingiustificati in seguito a elaborazioni automatiche.

La combinazione di big data e IA consente infatti di automatizzare i processi di decisione, anche in ambiti che richiedono scelte complesse, basate su numerosi fattori, in base a criteri non esattamente predeterminati. Ciò può migliorare la qualità delle decisioni pubbliche e private, ma comporta nuovi rischi.

La relazione tra profilazione e decisione è illustrata dalla Figura 3.1[99] che presenta uno schema semplificato del processo di decisione automatica basato sulla profilazione.

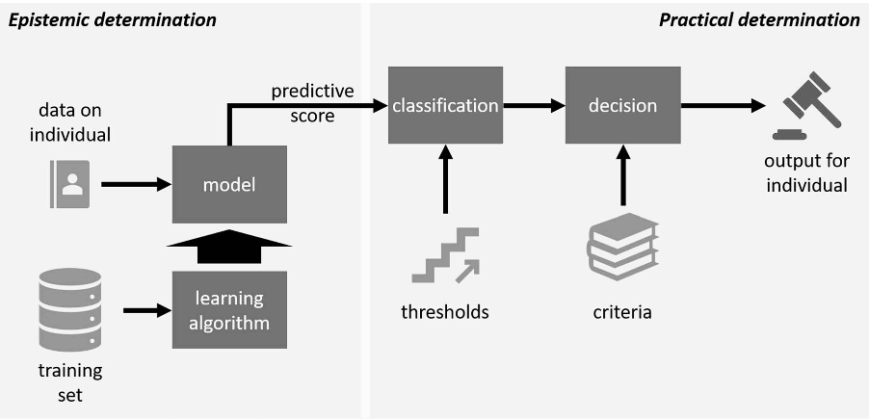

Figura 3.1: *Il processo della decisione automatica*

Il modello costruito dall'algoritmo di apprendimento automatico assegna agli individui un punteggio (score) che riflette la probabilità che l'individuo abbia la caratteristica predetta dal sistema (per es., che egli restituirà il credito che gli venisse concesso). Sulla base di tale punteggio l'individuo viene classificato (etichettato) in un certo modo (come un debitore affidabile o inaffidabile). Infine il sistema suggerisce o adotta la decisione collegata alla classificazione attribuita all'individuo (il mutuo è concesso a chi sia stato classificato come affidabile e negato a chi invece sia stato classificato come inaffidabile).

Si è aperto negli ultimi anni un ampio dibattito su prospettive e rischi delle decisioni algoritmiche. Alcuni studiosi hanno rilevato che in molti settori le previsioni e le decisioni algoritmiche, anche relative alla valutazione degli individui, possono essere più precise ed efficaci di quelle umane. I sistemi automatici possono evitare le propensioni all'errore proprie dell'uomo, in particolare nelle inferenze statistiche, così come i pregiudizi —etnici, sociali, di genere, ecc.— da cui spesso siamo affetti. Più in generale, le decisioni automatiche possono ridurre il "rumore" proprio delle decisioni umane (il fatto che casi simili possano avere decisioni diverse, senza una ragionevole spiegazione).[100] Si è osservato che in molti ambiti —dagli investimenti, al reclutamento di personale, alla concessione di libertà vigilata— le determinazioni algoritmiche risultano migliori, con riferimento ai criteri usuali, di quelle adottate anche da individui esperti.[101]

Altri hanno invece posto l'accento sulle possibilità di errore e discriminazione delle decisioni algoritmiche. È vero che solo in rari casi gli algoritmi assumeranno decisioni esplicitamente discriminatorie, attuando discriminazione diretta (*disparate treatment*), cioè basando le proprie previsioni su caratteristiche vietate, come razza, etnia o genere. Più spesso il risultato di una determinazione algoritmica comporterà una discriminazione indiretta (*disparate impact*), cioè avrà un impatto sfavorevole sproporzionato su individui appartenenti a certi gruppi, senza una giustificazione accettabile.

3.4.1 Le discriminazioni algoritmiche

I sistemi basati su metodi di apprendimento supervisionato imparano dagli esempi contenuti nel loro insieme di addestramento, e quindi tendono a riprodurre pregi e difetti dei comportamenti esemplificati, inclusa la propensione all'errore e al pregiudizio. Per esempio, un sistema per il reclutamento, addestrato su esempi di decisioni passate in cui certe categorie di persone (donne, giovani, anziani, disabili, minoranze, ecc.) sono state discriminate, riprodurrà la stessa logica. Il pregiudizio all'interno di un insieme di addestramento può essere presente anche quando le informazioni che costituiscono i dati di ingresso del sistema non comprendano caratteristiche la cui considerazione sia giuridicamente vietata, come l'etnia o il genere. Ciò può verificarsi ogni qual volta esista una correlazione tra caratteristiche discriminatorie e alcuni dati di input utilizzati dal sistema.

Supponiamo, ad esempio, che un responsabile delle risorse umane di un'azienda non abbia mai assunto candidati di una certa etnia a causa di un suo pregiudizio e che gli individui appartenenti a tale etnia abitino per lo più in certi quartieri della città. Un insieme di addestramento basato sulle decisioni di tale dirigente insegnerebbe al sistema a non selezionare gli individui residenti in quei quartieri, comportando il mancato accoglimento delle domande di assunzione provenienti da coloro che appartengano all'etnia discriminata.

In altri casi, un insieme di addestramento può essere sfavorevole a un certo gruppo, poiché il raggiungimento del risultato desiderato (ad esempio, certe prestazioni lavorative) è approssimato mediante un *proxy* (un elemento sostitutivo) che comporta un impatto discriminatorio indiretto su quel gruppo. Supponiamo, ad esempio, che le future prestazioni lavorative dei dipendenti (obiettivo di interesse nei processi di assunzione) siano valutate considerando unicamente la predizione concernente il numero di ore spese in ufficio. Tale criterio potrebbe comportare una valutazione peggiore delle donne rispetto agli uomini, qualora in passato le donne abbiano speso meno ore in ufficio, a causa di maggiori impegni familiari. Il sistema assocerebbe correttamente al genere femminile una minore propensione al lavoro straordinario, ma risulterebbe viziato dal collegamento di tale propensione ad una minore capacità lavorativa (senza considerare altri fattori, molto più significativi per la valutazione della capacità lavorativa).

In altri casi, errori e discriminazioni possono derivare da pregiudizi presenti nei dati usati nell'apprendimento automatico. Determinazioni inique e ingiuste possono derivare dall'uso di dati vantaggiosi applicabili ai soli membri di un certo gruppo (ad esempio, il fatto di aver frequentato certe scuole o Università, socialmente selettive).

L'ingiustizia può anche derivare dall'uso di dati basati su giudizi e valutazioni umane distorte, faziose, o parziali (ad esempio, le lettere di raccomandazione).

Infine, errori e discriminazioni possono risultare dal fatto che gli esempi nell'insieme di addestramento non riflettono la statistica della popolazione. Ciò può verificarsi soprattutto quando i membri di certi gruppi siano soggetti a maggiori controlli, o a valu-

tazioni meno benevole, e quindi abbiano maggiori probabilità di essere identificati quali autori di comportamenti indesiderati (sia che essi abbiano effettivamente tenuto quei comportamenti, sia che non li abbiano tenuti).

Supponiamo, per esempio, che nella valutazione di domande per ottenere la libertà condizionale, la presenza di precedenti penali a carico del reo abbia un peso sfavorevole, e che i membri di certi gruppi (ad esempio coloro che appartengono a una certa etnia o che professano un certo credo religioso) siano soggetti a controlli più rigorosi e stringenti, cosicché la loro attività criminale sia più spesso scoperta e conseguentemente condannata. Per i membri di tale gruppo, ciò comporterebbe generalmente una valutazione meno favorevole, a parità di propensione al crimine, rispetto ai membri di altri gruppi (i cui comportamenti criminali più spesso sfuggono ai controlli).

I membri di un gruppo possono anche essere svantaggiati dalla sottorappresentazione di quel gruppo nell'insieme di addestramento. Ciò può ridurre l'accuratezza delle previsioni per i membri di quel gruppo, e quindi diminuire l'attribuzione di caratteristiche desiderate o favorire l'attribuzione di caratteristiche indesiderate. Si consideri, ad esempio, il caso di un'azienda che ha abbia assunto poche donne in passato e che utilizzi i propri registri storici delle assunzioni come insieme di addestramento. Poiché le donne in passato non hanno avuto modo di rivestire certe posizioni e quindi di esemplificare le competenze desiderate in chi riveste tali posizioni, il sistema potrà non prevedere tali attitudini anche in nuove candidate dotate delle competenze richieste.

Si consideri invece il caso in cui un sistema per il riconoscimento facciale, usato per identificare gli autori di reati, si basi su un insieme di addestramento che contiene poche facce di persone appartenenti a una certa etnia. Ne potrebbe risultare una maggiore inaccuratezza delle identificazioni del sistema rispetto ai membri dell'etnia sottorappresentata. Di consequenza, i membri di quell'etnia potrebbero più spesso essere erroneamente identificati quali autori di reati, e quindi ingiustamente accusati.

3.4.2 Pervasività e contestazione delle decisioni algoritmiche

Per determinare l'impatto sociale delle decisioni algoritmiche, non è sufficiente confrontarle con decisioni umane sugli stessi temi, facendo riferimento a criteri quali l'accuratezza e l'equità. Bisogna altresì considerare la pervasività e l'economicità della predizione e decisione automatica, che può trovare applicazione anche in ambiti nei quali decisioni umane non sarebbero state possibili, o avrebbero avuto un costo esorbitante. Grazie all'enorme quantità di dati disponibili e all'efficienza delle elaborazioni automatiche, gli individui possono oggetto di molte più valutazioni e decisioni personalizzate, in un più ampio spettro di contesti. Di conseguenza, determinazioni potenzialmente errate, discriminatorie, o comunque inique (in tema di reclutamento, progressione di carriera, prestiti, premi assicurativi, ecc.) possono essere adottate in molti più casi e contesti, in base a un più ampio insieme di caratteristiche personali (dalla situazione economica, alle

condizioni di salute, alla residenza, alle scelte e vicende della vita, ai comportamenti online e offline, ecc.). Essere soggetti a una persistente osservazione e valutazione genera una pesante pressione psicologica e incide sull'autonomia personale.

Inoltre, la disponibilità di valutazioni automatiche —efficienti, economiche, e apparentemente oggettive, anche quando prive di una motivazione a noi comprensibile— può di per sé stessa avere conseguenze pregiudizievoli. Potrebbe indurre i decisori a trattare in modo iniquo le persone che si trovano in circostanze, non considerate dal sistema, di fragilità e bisogno. Inoltre, potrebbe indurli a respingere o ostacolare le contestazioni da parte degli interessati, poiché tali contestazioni, anche quando giustificate, interferiscono con l'operatività del sistema, causando costi e incertezze. Questo aspetto è stato sottolineato da molte voci critiche:

> Un algoritmo elabora un sacco di statistiche e ne estrae una probabilità che una certa persona potrà essere un cattivo dipendente, un mutuatario rischioso, un terrorista, o un insegnante scadente. Questa probabilità è distillata in un punteggio, che può sconvolgere la vita dell'interessato. E tuttavia, quando la persona reagisce, prove in contrario, meramente "evocatrici", semplicemente non funzionano. La contestazione deve essere inoppugnabile. Le vittime umane delle armi di distruzione matematica [...] sono tenute a uno standard di prova molto più alto rispetto agli stessi algoritmi.[102]

A queste critiche si è risposto osservando che se è vero che i sistemi algoritmici, e in particolare quelli basati sull'apprendimento automatico, possono riprodurre o anche esacerbare le iniquità esistenti, è anche vero che i processi algoritmici sono più controllabili di quelli sottesi alle decisioni umane, e possono essere migliorati e "ingegnerizzati" in modo da prevenire risultati iniqui.

> Mettendo in atto requisiti appropriati, l'uso degli algoritmi renderà possibile esaminare e valutare più facilmente l'intero processo di decisione, così da rendere molto più facile conoscere se si è verificata una discriminazione. Imponendo un nuovo livello di specificità, l'uso degli algoritmi mette in luce, e rende trasparenti, i compromessi tra valori in competizione. Gli algoritmi non sono solo una minaccia da regolare; accompagnati da corrette garanzie, essi hanno il potenziale per essere una forza positiva per l'equità.[103]

Non possiamo qui approfondire il tema della valutazione comparativa di decisioni umane e decisioni algoritmiche. Quello che è certo è che l'idea che solo le decisioni di routine —e non quelle che coinvolgono condizioni di incertezza, discrezionalità e valutazioni— possano essere affidate agli algoritmi è oggi superata. I sistemi di IA hanno dimostrato di poter operare con successo anche in ambiti in cui mancano criteri precisi e univoci, e quindi anche in settori tradizionalmente affidati all'intuizione umana, allenata con

l'esercizio (per esempio, nella diagnosi medica, negli investimenti finanziari, e nella concessione del credito).

L'alternativa ai processi decisionali automatizzati non è costituita da decisioni perfette ma da decisioni umane, con tutti i loro difetti e imperfezioni: anche un sistema algoritmico imperfetto potrebbe essere più giusto ed equo di un essere umano incline all'errore o al pregiudizio. Ciò non significa che in qualsiasi settore e ambito di attività la macchina prevalga sull'uomo. Quando è necessario affrontare questioni nuove, sviluppare soluzioni creative, cogliere attitudini, preferenze e interessi umani, la mente dell'uomo è insostituibile. La sfida per il futuro è trovare le migliori combinazioni tra intelligenza umana e artificiale, valutazioni umane e valutazioni automatizzate, integrandole fra loro, e tenendo conto delle potenzialità di ciascuna. Inoltre, l'IA può essere utilizzata per controllare le sue stesse applicazioni, così da individuare eventuali difetti nei meccanismi della decisione automatica, e aiutare nella predisposizione di contestazioni.

3.5 L'ecosistema della sorveglianza

L'uso dell'IA, in combinazione con le grandi masse di dati, ha trovato numerose applicazioni tanto nell'economia privata quanto nell'amministrazione pubblica. Alcuni autori hanno visto in modo positivo lo sviluppo di sistemi basati sulla raccolta massiva di informazioni, anche sui comportamenti individuali, osservando che l'integrazione di IA e big data consente maggiore efficienza, e fornisce nuovi mezzi di direzione e controllo sul comportamento sociale. Altri invece hanno sviluppato considerazioni critiche, con riferimento, in particolare, ai rischi di sorveglianza e manipolazione.

3.5.1 Le prospettive della fisica sociale

Al Varian, Chief Economist di Google dal 2002, osserva che, quando gli scambi economici —e più in generale, le interazioni sociali e le attività degli individui— sono mediati da computer, essi danno luogo a continue e dettagliate registrazioni di dati: il computer può osservare e verificare ogni aspetto dell'attività in cui è coinvolto.[104] I dati raccolti possono essere usati per analisi, valutazioni e feedback che consentono di personalizzare la relazione con gli interessati (come nella pubblicità personalizzata), ma anche per effettuare sperimentazioni (per esempio, valutando le risposte a cambiamenti nei prezzi o nei messaggi), e per guidare e controllare i comportamenti.

Diventano così possibili nuovi, più ampi ed efficienti, modelli di interazione economica e sociale, basati non più sull'incerta aspettativa del rispetto di norme sociali, etiche e giuridiche, o in vacillanti atteggiamenti di reciprocità e collaborazione, ma piuttosto nella possibilità di osservare ogni comportamento, di determinarne gli effetti, e di collegare ad esso sanzioni e ricompense. Tutti noi ci affidiamo a venditori di beni o for-

nitori di servizi con cui non abbiamo avuto alcun contatto personale, confidando nella piattaforma attraverso cui tali beni e servizi vengono forniti, e nei meccanismi di valutazione e votazione (*rating* e *scoring*), selezione ed esclusione attuati dalla piattaforma. Si pensi altresì a come sistemi basati sulla "catena di blocchi" (*blockchain*), un archivio immodificabile, replicato in tutti i nodi del sistema, in cui vengono registrate tutte le transazioni. Tali sistemi consentano di creare monete digitali, meccanismi contrattuali che si auto-eseguono (*smart contract*), e organizzazioni economiche digitali.[105]

Alex Pentland, che dirige lo Human Dynamics Lab presso il celebre MIT Media Lab, ha affermato che IA e big data offrono la prospettiva di una "fisica sociale", quale scienza in grado di capire e guidare la società. La disponibilità di grandi masse di dati e di metodi e risorse computazionali per elaborarli consentirebbe di realizzare finalmente il sogno di August Comte, cioè di ottenere una scienza sociale dotata di fondamenti teorico-matematici così come di capacità operative.

> Attraverso una migliore conoscenza di noi stessi, potremmo potenzialmente costruire un mondo senza guerra o crisi finanziarie, in cui le malattie infettive siano rapidamente individuate e fermate, in cui l'energia, l'acqua e le altre risorse non siano più sprecate, e in cui i governi siano parte della soluzione invece che parte del problema.[106]

3.5.2 Il capitalismo e lo Stato della sorveglianza

Alle prospettive di crescita economica e sociale offerte dall'integrazione di IA e big data si accompagno i rischi associati al "capitalismo della sorveglianza"[107] e allo "Stato della sorveglianza"[108].

Shoshana Zuboff identifica infatti nel capitalismo della sorveglianza il modello economico tipico della nostra epoca. Zuboff riprende il lavoro di Karl Polanyi[109], il quale aveva osservato che l'avvento del capitalismo —la "grande trasformazione"– è consistito nell'occupazione prima dell'economia e poi di altri ambiti dello spazio sociale da parte del mercato. In particolare, Polanyi affermava che il capitalismo trasforma in merci (prodotti da vendere sul mercato) anche entità che non sono state prodotte per il mercato: la terra (l'ambiente), il lavoro, e il denaro. Ne risultano tensioni distruttive per l'intera società, se le dinamiche del mercato non sono soggette a limiti e controlli, da parte del diritto, dalla politica e dei movimenti sociali (come quelli di lavoratori e consumatori).

Nel capitalismo della sorveglianza, afferma Zuboff, il dominio del mercato si estende all'esperienza umana: il comportamento degli individui viene registrato e analizzato. I dati, le previsioni e le conseguenti capacità di influenza, diventano una nuova merce.

> Il capitalismo della sorveglianza annette l'esperienza umana alla dinamica del mercato, cosicché essa rinasce come comportamento: la quarta "merce fittizia". Le prime tre merci fittizie —la terra, il lavoro e il denaro— sono

> state assoggettate alla legge. Nonostante queste leggi siano state imperfette, le istituzioni del diritto del lavoro, del diritto dell'ambiente e del diritto bancario sono quadri di regolazione volti a difendere la società (e la natura, la vita e lo scambio) dai peggiori eccessi del potere distruttivo del capitalismo allo stato grezzo. L'espropriazione dell'esperienza umana da parte del capitalismo della sorveglianza non ha incontrato impedimenti siffatti.[110]

Zuboff osserva che anche nel caso del capitalismo della sorveglianza le dinamiche del mercato, se lasciate a sé stesse, conducono a esiti distruttivi. Le persone sono soggette a manipolazione, sono private del controllo sul proprio futuro e di uno spazio in cui sviluppare la propria personalità. Le reti sociali di collaborazione vengono sostituite da meccanismi di incentivi e disincentivi che danno luogo a nuove forme di controllo e sfruttamento.

Si pensi alle dinamiche generate dalle piattaforme di Internet, che monitorando il comportamento degli utenti e analizzandolo mediante tecniche di IA sono in grado di influenzare le scelte degli utenti stessi. Si pensi altresì alle piattaforme per la fornitura di servizi —come Uber o Lyft nel campo dei trasporti— che registrano ogni attività svolta dai prestatori del servizio, le modalità in cui tali attività sono svolte, le reciproche valutazioni di prestatori e clienti, e associano ricompense e penalità ad ogni aspetto di tali attività, così da guidare il comportamento di ciascuno nel senso desiderato. Si tratta di un nuovo sistema di governo del comportamento umano (che sostituisce i tradizionali meccanismi del diritto e del contratto) con esiti economicamente efficienti, ma potenzialmente negativi per il benessere mentale e l'autonomia degli interessati.[111]. Mentre il capitalismo classico trovò correttivi sufficienti a limitarne gli esiti più distruttivi, secondo Zuboff mancano a tutt'oggi risposte adeguate ai nuovi rischi rappresentati dal capitalismo della sorveglianza.

Il capitalismo della sorveglianza trova un corrispettivo nella dimensione pubblica. Qui è emerso negli ultimi decenni il modello dello "Stato della sorveglianza", caratterizzato:

> dalla raccolta e dal confronto di informazioni sui cittadini, e dall'analisi di tali informazioni per identificare problemi, prevenire minacce potenziali, governare la popolazione, e fornire utili servizi sociali. Lo Stato della sorveglianza è un caso particolare dello Stato dell'informazione, uno stato che cerca di identificare e risolvere i problemi di governance attraverso la raccolta, il confronto, l'analisi e la produzione di informazioni".[112]

Anche in questo caso, ai possibili vantaggi relativi all'efficienza nella gestione delle attività pubbliche, al coordinamento dei comportamenti dei cittadini, alla prevenzione dei rischi, si accompagnano prospettive inquietanti, nuove forme di influenza e controllo volte ad asservire i cittadini rispetto agli scopi e ai valori di chi controlla l'infrastruttura di sorveglianza.[113]

3.5.3 Dalla pubblicità online alla manipolazione dell'opinione pubblica

Le tecniche per la sorveglianza, la profilazione, la persuasione e la decisione algoritmica sono state elaborate soprattutto in ambiti commerciali, in particolare nel contesto del commercio elettronico. La raccolta di dati sugli utenti, l'anticipazione delle loro attitudini e preferenze, la conseguente possibilità di influenzare il comportamento, hanno consentito di realizzare forme particolarmente efficaci di pubblicità mirata o micromirata rispetto alle caratteristiche dei singoli utenti. La possibilità di inviare pubblicità mirata, a propria volta, ha conferito un vantaggio competitivo alle piattaforme per il commercio elettronico, che in pochi anni hanno acquisito la fetta più importante del mercato pubblicitario, traendone enormi profitti.

Il modello di affari basato sull'invio di pubblicità mirata ha fatto sì che tutte le persone connesse online siano oggetto di penetrante sorveglianza e influenzamento algoritmico a fini pubblicitari. Inoltre, tale modello ha condotto le piattaforme a usare gli stessi metodi e tecnologie —l'invio di messaggi e notizie personalizzati— per trattenere gli utenti all'interno delle piattaforme, o dei siti, in modo che gli stessi utenti possano essere oggetto di messaggi pubblicitari.[114] Si è così creata una forma di comunicazione basata sull'invio dei contenuti che maggiormente inducono i destinatari a dedicare la propria attenzione alla piattaforma. Si tratta di regola dei contenuti che più corrispondono alle preferenze tematiche, ideologiche, politiche e culturali dei destinatari. Anche ai suggerimenti attinenti alle relazioni sociali da attivare sulle piattaforme si possono applicare metodi analoghi, cosicché ciascuno è condotto ad interagire con chi gli è più vicino nelle preferenze e attitudini. Il funzionamento delle piattaforme pertanto tende pertanto a favorire la polarizzazione delle opinioni, e la suddivisione dei cittadini in gruppi ideologicamente omogenei e non comunicanti.[115]

Inoltre, il modello della pubblicità online ha favorito la proliferazione dei contenuti che, come le cosiddette *fake news*, maggiormente attirano l'attenzione degli utenti, così da esporre gli stessi a messaggi pubblicitari pubblicitari diffusi insieme a quei contenuti. Il termine *fake news* identifica il fenomeno relativo alla generazione di contenuti distorti, fuorvianti, e/o falsi, generalmente "micro-mirati" e distribuiti online al fine di influenzare le opinioni di singoli individui e gruppi, profittando delle loro debolezze e insicurezze.

Non solo i produttori di *fake news* ma anche le piattaforme traggono vantaggio dai profitti pubblicitari derivanti dall'accesso a quei contenuti, il che disincentiva i controlli da parte delle piattaforme stesse.[116]

Si consideri, per esempio, il modello di affari alla base della proliferazione di *fake news* su Facebook. Tale modello comprende i seguenti momenti:

1. Gli individui interessati a diffondere *fake news* pubblicano le stesse nelle pagine di siti web specializzati nella distribuzione di *fake news*. Tali pagine contengono inserzioni pubblicitarie, accanto alle *fake news*.

2. Gli stessi individui acquistano spazi pubblicitari da Facebook, per distribuire agli utenti della rete sociale i link alle pagine contenenti le *fake news*. Facebook propone quei link nelle *newsfeed* (aggiornamenti alle notizie) degli utenti di Facebook (selezionati in base a vari criteri).
3. Facebook percepisce introiti pubblicitari in base al numero di click sui link verso *fake news* proposti ai propri utenti.
4. Gli utenti che cliccano sui link vengono indirizzati alle pagine corrispondenti sul sito di *fake news*, generando una registrazione per ogni pubblicità presentata loro durante la visita del sito.
5. Il sito contenente le *fake news* viene retribuito dagli inserzionisti sulla base delle registrazioni (di accessi alla pubblicità) generate dalle visite degli utenti, e i relativi profitti possono essere condivisi con chi ha prodotto le notizie.

A questo ciclo principale, basato su siti specializzati, si può aggiungere un ciclo secondario.

6. I produttori di *fake news* oltre a pubblicarle su siti dedicati a *fake news* (punto 1 sopra), possono pubblicare le stesse notizie sulla propria pagina Facebook, e sponsorizzare tale pagina, per ottenere delle adesioni da parte di altri utenti, i cosiddetti seguaci (*followers*).
7. Le *fake news* pubblicate nella pagina Facebook del produttore sono visualizzate automaticamente nelle newsfeed dei seguaci, che a loro volta condividono il link alla notizia con i propri contatti. In questo modo la notizia diventa virale, grazie alla ripetuta replicazione dei link verso la pagina Facebook che la contiene.
8. Se la pagina Facebook del produttore di *fake news* contiene inserzioni pubblicitarie, il produttore, a propria volta, ne trae profitto.

Sulla base dei meccanismi appena indicati, chi pubblica le *fake news* e gestisce i siti relativi trae proventi pubblicitari, cosicché un'attività dannosa per la società diventa vantaggiosa per chi la pone in essere. Le *fake news* tendono a diventare virali specialmente nei momenti di crisi sociale ed economica, quando gli individui cercano spiegazioni per i propri disagi e responsabili cui attribuirne la colpa:

> [O] gni volta che incombe una minaccia o si è verificato un evento spaventoso, è inevitabile che si diffondano voci incontrollate [...]. Nel periodo che segue una crisi vengono fatte molte congetture. Alcuni le ritengono plausibili, forse perché offrono una risposta all'indignazione e al desiderio di trovare un colpevole. Gli eventi tragici generano indignazione, e in un simile stato d'animo le persone accettano molto più facilmente le dicerie che giustificano le loro condizioni emotive, e sono inoltre più propense ad attribuire quegli eventi all'azione intenzionale.[117]

3.5.4 Profilazione e IA nella comunicazione politica

Come abbiamo visto nella sezione precedente, i meccanismi dominanti nell'economia della rete possono favorire usi dell'IA che hanno un impatto negativo sulla formazione dell'opinione pubblica e quindi sull'assetto democratico della società.

Infatti, la formazione di un'opinione pubblica consapevole presuppone il confronto delle diverse opinioni e la conoscenza dei fatti rilevanti. Invece, il meccanismo per l'invio di informazioni mirate e gradite al destinatario (la c.d. bolla del filtro, *filter bubble*) fa sì che ciascun individuo tendenzialmente non acceda a opinioni diverse dalla propria. Inoltre, il meccanismo delle *fake news* genera disinformazione, sfiducia nelle istituzioni e nei media tradizionali, e credenze incompatibili con la realtà dei fatti.

Le tecniche dell'IA possono essere utilizzate non solo nelle comunicazioni commerciali, ma anche nella comunicazione politica, influenzando le modalità del dibattito, la propaganda, e lo stesso esito delle campagne elettorali. Da un lato, l'IA può contribuire a fornire ai cittadini maggiori informazioni, meglio corrispondenti agli interessi, ai bisogni, e alle modalità cognitive di ciascuno, e può facilitare altresì l'instaurazione di interazioni, l'aggregazione delle opinioni, e lo sviluppo di iniziative collettive. Dall'altro lato però, la stessa IA può essere usata intenzionalmente a fini di disinformazione e manipolazione delle opinioni politiche e del comportamento elettorale. L'effetto delle tecnologie per l'influenzamento è accresciuto dalla sinergia con la polarizzazione e la disinformazione che risultano dai meccanismi economici della rete menzionati nella sezione precedente.[118]

I pericoli derivanti dall'uso scientifico delle tecniche per la disinformazione, l'influenza e la manipolazione a fini politici sono emersi con chiarezza nel caso di Cambridge Analytica, che attiene al tentativo di influenzare, con tecnologie di IA, le preferenze di voto dei cittadini durante le elezioni presidenziali americane del 2016 e probabilmente anche durante il referendum sulla Brexit.[119]

Il coinvolgimento di Cambridge Analytica nelle elezioni presidenziali americane può essere analizzato in quattro fasi principali,[120] di seguito descritte.

- *Fase* 1. I cittadini registrati come elettori negli Stati Uniti furono invitati a sottoporsi a un test della personalità, al fine di identificare il proprio profilo psicologico, in relazione ad aspetti come apertura mentale, coscienziosità e stabilità mentale. Per incentivare la partecipazione al test, disponibile online e basato su circa 120 domande, ai volontari fu promesso un premio consistente in una piccola somma di denaro (da due a cinque dollari). Ai partecipanti fu anche detto che i dati raccolti sarebbero stati utilizzati unicamente a fini di ricerca accademica. Circa 320.000 individui parteciparono al test. Per ricevere la ricompensa, a ciascun partecipante fu richiesto l'accesso al rispettivo profilo Facebook.
- *Fase* 2. L'accesso ai profili Facebook consentì di correlare le risposte di ciascun partecipante con le informazioni e le attività presenti sul profilo (per esempio *like*,

condivisioni e altri indicatori). Dopo aver ottenuto l'accesso ai profili dei partecipanti al test, Cambridge Analytica raccolse non solo i dati ivi contenuti ma anche le informazioni presenti sui profili dei loro amici (le persone alle cui pagine personali si ha accesso tramite Facebook) fino a coinvolgere complessivamente tra i 30 e i 50 milioni di individui.

- *Fase* 3. Una volta terminata la raccolta dei dati, Cambridge Analytica ebbe a disposizione due tipologie di dati personali da elaborare. Da un lato, i dati sui partecipanti al test, vale a dire le risposte fornite durante il questionari, più i rispettivi profili Facebook e dall'altro, le informazioni sui loro amici, costituite dai soli profili Facebook di questi ultimi. I dati dei partecipanti al questionario furono utilizzati da Cambridge Analytica come insieme di addestramento, per la profilazione di coloro che non avevano partecipato al questionario. Le informazioni estratte dai profili Facebook dei partecipanti (like, post, condivisioni, ecc.), quindi considerate come predittori e le risposte al questionario (e le attitudini psicologiche e le preferenze politiche correlate a tali risposte), come obiettivi da predire sulla base di quei predittori. Grazie a metodi di apprendimento automatico Cambridge Analytica poté quindi costruire un modello che correlasse i dati contenuti nei profili di Facebook individuali (contenuti, *like*, contatti, ecc.) con le attitudini psicologiche (personalità più o meno estroversa, emotiva, ecc.), e le preferenze politiche, così che i primi potessero fungere da predittori rispetto alle seconde.
- *Fase* 4: Sulla base di tale profilazione psicologica e politica, Cambridge Analytica identificò gli elettori indecisi, vale a dire coloro che avrebbero potuto cambiare le proprie preferenze di voto se debitamente sollecitati, e li sottopose a comunicazioni mirate atte ad innescare il cambiamento desiderato, senza che gli stessi fossero consapevoli dello scopo di tali comunicazioni.

Il caso di Cambridge Analytica mostra come la combinazione e l'analisi dei dati possano rivelare informazioni profondamente personali e specifiche sulle persone. Tali informazioni possono essere utilizzate per creare e inviare messaggi personalizzati tesi ad influenzare —anche facendo appello a fattori emotivi— scelte che idealmente dovrebbero essere deliberate, private e ponderate.

L'uso delle tecnologie di Internet e dell'IA all'interno della comunicazione politica trova ulteriore riscontro nella crescente presenza di bot politici nelle piattaforme online. Si tratta di software progettati per generare contenuti, condividere notizie e informazioni e interagire con gli utenti delle piattaforme. I bot generano una quantità crescente di traffico online e rappresentano una parte significativa dei profili attivi sulle piattaforme. Per esempio, si è affermato che solo su Twitter ci sarebbero circa 30 milioni di account che fanno capo a bot, anziché a persone.

I bot politici possono essere utilizzati anche per simulare una maggiore popolarità di personaggi politici, fingendo un'identità umana e registrandosi nelle reti sociali come

seguaci di quei personaggi. Possono inoltre diffondere notizie all'interno delle piattaforme, aumentando la replicazione di tali notizie, o disturbare la comunicazione politica degli avversari o dar vita a campagne diffamatorie. Per esempio, essi possono generare etichette (*hashtag*) prive di rispondenza alle notizie e alle immagini cui sono associate, o possono cercare notizie e dichiarazioni politiche al fine di negarle, generando così disinformazione e manipolando l'opinione pubblica. I bot politici sono generalmente caratterizzati da una forte polarizzazione ideologica, essendo finalizzati a promuovere una particolare opinione all'interno del dialogo politico.

L'uso di bot, in combinazione con le tecnologie dell'IA e dei big data può intensificare gli effetti negativi della cosiddetta propaganda politica computazionale, che combina tecnologie per la raccolta di informazioni, la profilazione, l'analisi e la generazione di contenti, e la decisione automatica.

3.5.5 Nuove dimensioni della sorveglianza. Il Sistema di Credito Sociale cinese

Nelle sezioni precedenti si è esaminato come l'IA possa essere usata in modo da disinformare e manipolare i cittadini. A questi rischi si affiancano quelli connessi all'uso delle stesse tecnologie per sorvegliare, valutare, e indirizzare i comportamenti dei cittadini. Mentre nel primo caso si manipolano le opinioni in modo da indurre comportamenti desiderati, nel secondo si influenza direttamente il comportamento mediante (micro) ricompense e sanzioni.

Un esempio paradigmatico è rappresentato da un'iniziativa del governo cinese, il cosiddetto Sistema di Credito Sociale (SCS). Tale sistema assegna ad ogni cittadino un punteggio che ne quantifica il valore sociale e ne oggettiva la reputazione.[121] Esso si basa sull'aggregazione e l'analisi di informazioni sulla sfera pubblica e privata di ciascuno. Tra gli aspetti considerati vi sono comportamenti economici (come il rispetto di obblighi legali, l'adempimento dei contratti), politici (come la partecipazione a movimenti e manifestazioni), giudiziari (il coinvolgimento in procedimenti passati e presenti, gli adempimenti conseguenti), e sociali (come la partecipazione a reti sociali e a rapporti interpersonali online e offline). A ogni nuovo comportamento rilevante —l'ambito dei comportamenti rilevanti è in linea di principio illimitato— il sistema assegna un punteggio positivo o negativo. Il totale dei punteggi costituisce il credito sociale di ciascun individuo, e ne determina l'accesso a servizi (come università, locazioni, e trasporti), posizioni lavorative, finanziamenti, ecc.

Forse il tipo di sorveglianza che più si avvicina al credito sociale, nell'esperienza occidentale è quella dei lavoratori delle piattaforme online, o comunque che operano in ambienti pervasi da tecnologie di controllo e sorveglianza. Si pensi ai lavoratori che, governati da piattaforme online, svolgono funzioni di trasporto di cose o persone: essi sono controllati e valutati in ogni loro spostamento, nei tempi di attività e inattività,

nelle risposte alle richieste di prestazioni, ecc. Si pensi inoltre ai lavoratori con funzioni esecutive in ambienti ad elevata e pervasiva automazione (per esempio, i magazzini di Amazon), che sono controllati minuziosamente nei tempi di lavoro e negli spostamenti.

Sempre più frequenti, inoltre sono le forme di sorveglianza cooperativa, detta anche co-veglianza. Pensiamo per esempio alla possibilità, offerta da numerose piattaforme, di fornire informazioni e valutazioni sui fornitori di prodotti o servizi. Più in generale, è diffusa la possibilità —per gli utenti di servizi, acquirenti di prodotti, fruitori di contenuti, ecc.— di comunicare la propria approvazione o disapprovazione rispetto a contenuti o attività.

Si potrebbe affermare che non vi siano differenze sostanziali tra gli effetti individuali e sociali del Sistema di Credito Sociale cinese e dei sistemi di punteggio utilizzati nei paesi occidentali, data la vicinanza delle tecnologie utilizzate e dei meccanismi psicosociali sfruttati per indurre modifiche comportamentali. In entrambi i casi, la sorveglianza pervasiva e la valutazione persistente possono innescare comportamenti di sottomissione, conformismo e adattamento alle regole, motivati unicamente dal desiderio di evitare sanzioni e ottenere ricompense. Ciò potrebbe sostituire ed eliminare l'autentica esperienza della normatività morale e giuridica. Un aspetto essenziale di tale esperienza è infatti la necessità di affrontare, secondo la propria coscienza, il conflitto tra l'interesse personale alla devianza e l'esigenza di adempiere agli imperativi morali o giuridici. La possibilità di un comportamento deviante rispetto alle norme sociali (senza la certezza assoluta di incorrere nella relativa sanzione) sarebbe preclusa da un sistema di sorveglianza e valutazione pervasivo.

Alla tesi appena esposta, si potrebbe obiettare che i sistemi di classificazione e valutazione non necessariamente inducono conformità e sottomissione piuttosto che creatività e autonomia. Se basati su fattori suscettibili di assumere una scala di diversi valori (per esempio, risultati scolastici, tempi di risposta, valutazioni degli utenti, ecc.), quegli stessi sistemi potrebbero invece incentivare l'eccellenza, sebbene secondo criteri prestabiliti. Nell'ambito del commercio online, per esempio, le valutazioni dei consumatori, se gestite in modo da garantirne l'autenticità e da prevenire abusi, possono contribuire a indirizzare utilmente il comportamento di venditori e consumatori.

L'assegnazione pervasiva di punteggi, tuttavia, produce inevitabilmente confronti, competizione, stress, e problemi relazionali. C'è il rischio che gli individui investano in modo eccessivo nel raggiungimento di punteggi elevati al solo fine di primeggiare. Poiché non tutti possono eccellere, alcuni rimarranno necessariamente indietro, e ciò può comportare non solo risultati negativi circoscritti —come licenziamenti o mancanza di opportunità– ma anche perdita di autostima e esclusione sociale. Inoltre, la difficoltà di migliorare il proprio punteggio (di riabilitazione o redenzione) può portare al disinvestimento e alla rinuncia.

Le considerazioni fatte fin qui si applicano a tutti i sistemi automatizzati di valutazione e punteggio su larga scala. Tuttavia, il Sistema di Credito Sociale cinese presenta

alcune peculiarità rispetto ai modelli occidentali.

In primo luogo, tale sistema è basato su un unico punteggio, ottenuto mediante l'aggregazione di valutazioni attinenti a diversi aspetti della vita di uno stesso individuo. Tale caratteristica del modello cinese lo distingue dai sistemi di valutazione occidentali, che guardano invece ai cittadini come "dividui", cioè come portatori di aspetti parziali della personalità.[122] I sistemi occidentali, infatti, si focalizzano su un particolare ambito della vita dell'individuo (per esempio, lo prendono in considerazione quale consumatore, debitore, giocatore, lavoratore, assicurato, paziente, medico, studioso, ecc.), e lo valutano rispetto a fattori pertinenti a tale ambito. Da questa prospettiva, si potrebbe affermare[123] che ciascun siffatto sistema di punteggio applica i criteri di una diversa "sfera di valutazione". Tale valutazione frammentata ha un impatto minore sull'autostima dei singoli così come sulla loro valutazione sociale, poiché ogni individuo può differenziare la propria identità sulla base di ciascun punteggio settoriale, il cui impatto è limitato al contesto corrispondente.

Quando invece le diverse valutazioni vengono aggregate in un unico punteggio, come nel caso del Sistema di Credito Sociale cinese, questo diviene l'unico indicatore del merito personale; pertanto, qualsiasi comportamento è in grado di innescare effetti sui ogni ambito della vita di un individuo, a causa del suo impatto sul punteggio complessivo.

Una seconda caratteristica del Sistema di Credito Sociale cinese riguarda il fatto che esso è gestito da un regime autoritario. Non vi è alcuna assicurazione che i meccanismi di classificazione e valutazione adottati in tale sistema — determinata da autorità politiche cinesi senza i vincoli che caratterizzano la rappresentanza democratica e lo stato di diritto— corrispondano a una nozione ragionevole di virtù civica, piuttosto che a obiettivi di conformismo e controllo sociale.

3.6 Etica per l'IA

L'IA costituisce forse la principale sfida alla quale l'umanità dovrà far fronte nei prossimi decenni. Si aprono grandi opportunità di progresso individuale e sociale, e si prospettano, al tempo stesso, gravi rischi, anche nel campo della formazione dell'opinione pubblica e del funzionamento dei processi democratici.

Da un lato, l'IA può contribuire all'informazione dei cittadini, allo sviluppo del dibattito pubblico, alla nascita di nuove forme di aggregazione, alla formazione di opinioni razionali, basate su dati di fatto e risultati scientifici. D'altro lato, essa può invece incidere negativamente sul dibattito democratico, sulle elezioni e sul funzionamento delle istituzioni.

Nelle pagine precedenti ci siamo soffermati sui possibili impatti negativi dell'IA. In particolare, abbiamo esaminato i rischi della profilazione e delle decisioni automatiche, della polarizzazione, della disinformazione, della manipolazione e infine della sorveglianza. L'analisi realistica dei pericoli presenti e l'anticipazione di quelli futuri

non deve condurre, tuttavia, al rifiuto generalizzato delle tecnologie dell'IA. Non solo tali tecnologie possono portare enormi benefici, ma la loro diffusione a livello globale è spinta da dinamiche economiche e sociali che non possono essere fermate. Bisogna invece indirizzare l'uso di tali tecnologie verso risultati benefici, e prevenire i possibili esiti negativi con adeguate misure politiche, giuridiche e tecnologiche.

Nel corso degli ultimi anni, mentre si diffondevano le applicazioni di IA e crescevano le preoccupazioni sui loro possibili impatti, ha preso avvio un dibattito via via più intenso sull'etica e il diritto dell'IA.

3.6.1 L'etica dell'IA

L'etica dell'IA studia come sviluppare e utilizzare in modo "buono" le tecnologie dell'IA, come assicurare che la loro natura e il loro impiego corrisponda a norme e valori etici. Essa può essere vista come un aspetto di una più generale etica dell'informatica, intesa come:

> il ramo dell'etica che studia e valuta i problemi morali che riguardano i dati (inclusa la generazione, la registrazione, la cura, la disseminazione, la condivisione e l'uso), gli algoritmi (inclusa l'IA, gli agenti artificiali, l'apprendimento automatico e i robot) e le pratiche corrispondenti (inclusa l'innovazione responsabile, le norme sull'hacking e i codici delle professioni), al fine di formulare e promuovere soluzioni moralmente buone (per es., condotte corrette o valori corretti).[124]

Al dibattito sull'etica dell'IA hanno preso parte gli studiosi della materia, ma anche le imprese del digitale, le istituzioni politiche, le organizzazioni della società civile, le religioni.[125] Tale dibattito ha condotto alla redazione di numerosi documenti promossi e sottoscritti da soggetti pubblici e privati.

L'analisi comparativa dei documenti sull'etica dell'IA mostra una convergenza globale sui valori della trasparenza, non-maleficenza, responsabilità e privacy. Anche dignità e solidarietà sono menzionate frequentemente. Questa convergenza non esclude che rispetto a scelte concrete (per esempio, se ammettere o non l'identificazione facciale negli spazi pubblici) possano esserci differenze importanti tra diverse culture o diverse categorie di attori (per es., imprese, governi, società civile, ecc.) Nell'impossibilità di dar conto di tale dibattito,[126] ci limiteremo ad illustrare alcune iniziative dell'Unione Europea in materia.

3.6.2 I principi

Promuovere pratiche eticamente positive nell'uso dell'IA significa assicurare che lo sviluppo e l'impiego dell'IA abbia luogo in un contesto socio-tecnico —inclusivo di tecnologie, capacità umane, strutture organizzative e norme etiche e giuridiche— nel quale

interessi individuali e valori sociali siano rispettati e promossi. A tal fine, non è sufficiente far riferimento a specifiche norme etiche e giuridiche oggi in vigore, che possono essere inadeguate rispetto ai nuovi problemi dell'IA, e comunque non danno indicazioni sufficienti ai cittadini e alle istituzioni pubbliche e private.

È quindi necessario richiamarsi ai principi più astratti e fondamentali, che includono diritti fondamentali e valori sociali, sia al livello etico che a quello giuridico. Una sintesi di alto livello della cornice etica per l'IA è fornita dal documento AI4People,[127] che presenta come segue le opportunità fornite dall'IA e i rischi corrispondenti:

- abilitare l'autorealizzazione umana, senza svalutare le capacità umane,
- promuovere la capacità di azione umana, senza rimuovere la responsabilità umana,
- coltivare la coesione sociale, senza erodere l'autodeterminazione umana.

Lo High-Level Expert Group on Artificial Intelligence ha recentemente pubblicato un insieme di linee guide etiche per IA affidabile o meglio "degna di fiducia" (*trustworthy AI*).[128] La cornice etica e giuridica dell'IA dovrebbe riflettere i seguenti principi:

- Il rispetto per l'autonomia umana. Gli umani che interagiscono con l'IA dovrebbero mantenere piena ed effettiva autodeterminazione. l'IA non dovrebbe ingiustificatamente subordinare, coartare, ingannare, manipolare, condizionare o massificare gli esseri umani, ma dovrebbe essere progettata per aumentare, integrare o rafforzare le abilità umane, di tipo cognitivo, sociale o culturale.
- Prevenzione del danno. Dovrebbe essere garantita la protezione della dignità umana, così come l'integrità fisica e mentale. Secondo questo principio, i sistemi di IA e gli ambienti nei quali essi operano devono essere sicuri e protetti, tali sistemi non debbono causare o esacerbare danni o avere altri impatti avversi sugli esseri umani.
- Equità (*fairness*), nella dimensione sostanziale e in quella procedurale. La dimensione sostanziale implica un impegno per assicurare una distribuzione eguale e giusta sia dei benefici che dei costi, e per assicurare che individui e gruppi siano liberi da pregiudizi, discriminazione e stigmatizzazione. La dimensione procedurale implica l'abilità di contestare le decisioni adottate da sistemi di IA e dagli esseri umani che li gestiscono, e di cercare un rimedio effettivo contro tali decisioni.
- Spiegabilità. I processi algoritmici debbono essere trasparenti, le capacità e gli scopi dei sistemi di IA debbono essere comunicati apertamente e le decisioni debbono essere spiegabili a quanti siano toccati direttamente o indirettamente da esse.

Secondo lo stesso High Level Expert Group, al fine di attuare e realizzare un'IA affidabile, debbono essere soddisfatti sette requisiti , in base ai principi appena indicati:

- capacità di agire e supervisione umana, incluso il rispetto dei diritti fondamentali;
- robustezza tecnica e sicurezza, inclusa la resilienza ad attacchi e la protezione, piani di riserva e sicurezza generale, accuratezza, affidabilità e riproducibilità;
- privacy e governo dei dati, incluso il rispetto per la privacy, la qualità, l'integrità dei dati, e l'accesso ai dati;
- trasparenza, che include tracciabilità, spiegabilità e comunicazione;
- diversità, non-discriminazione ed equità, inclusa la prevenzione di pregiudizi iniqui (*unfair bias*), accessibilità e progettazione universali (in modo che tutti possano utilizzare sistemi e servizi), partecipazione di tutti i soggetti interessati;
- benessere sociale e ambientale, incluse la sostenibilità e la protezione dell'ambiente, l'impatto sociale, la società e la democrazia;
- responsabilità (*accountability*), che include la verificabilità (*auditability*), la minimizzazione e la segnalazione degli impatti negativi, dei compromessi delle riparazioni.

L'attuazione di questi requisiti deve riguardare l'intero ciclo di vita di un sistema di IA, come richiesto da ciascuna specifica applicazione.

3.6.3 Tre interessi da tenere in considerazione

Non è possibile esaminare nel dettaglio i valori e i rimedi appena illustrati. Ci si limiterà a considerare tre interessi sui quale l'IA può incidere in modo significativo.

Innanzitutto, c'è l'interesse alla protezione dei dati, cioè all'uso solo legittimo e proporzionato dei dati personali. Questo interesse è compromesso in un ambiente (online e offline) nel quale ogni comportamento sia registrato, e le relative informazioni siano utilizzate per estrarre ulteriori conoscenze sugli individui, al di fuori del loro controllo, e per elaborare tali conoscenze in modi potenzialmente contrari agli interessi degli stessi.

Il trattamento di dati personali attraverso sistemi di IA può anche compromettere l'interesse a un trattamento algoritmico equo e corretto, ovvero l'interesse a non essere soggetti a pregiudizi ingiustificati in seguito a elaborazioni automatiche, come osservato nelle sezioni precedenti

La possibilità di trattamenti algoritmici iniqui e scorretti, così come il bisogno di mantenere il controllo sui propri dati e di comprendere (e possibilmente contestare) le ragioni per le determinazioni che riguardano ciascuno, origina un interesse alla trasparenza/spiegabilità algoritmica. In altre parole, l'interessato ha l'esigenza di capire come e perché sia stata data una certa risposta o sia stata presa una certa decisione, così da "comprendere il processo di decisione dell'IA e poter chiedere conto di esso".[129] In primo luogo, questa esigenza riguarda le decisioni più importanti, da parte di poteri pubblici

e privati (l'accesso a impieghi o ad altre posizioni, la concessione di prestiti, l'allocazione di benefici, o l'imposizione di sanzioni). Tuttavia, la trasparenza/spiegabilità dovrebbe essere garantita anche quando, sulla base di profilazione, gli individui siano oggetto di una serie di micro-decisioni che individualmente considerate non sono particolarmente importanti, ma che, nel loro complesso, sono in grado di incidere significativamente su di essi.

Sia che si tratti di singole decisioni significative o di sequenze di micro-decisioni basate su profilazione, l'autonomia individuale è compromessa quando gli individui interagiscono con sistemi informatici che si presentano come scatole nere,[130] i cui meccanismi di funzionamento rimangono inaccessibili, e le cui decisioni restano inspiegabili e quindi non contestabili. A questo riguardo si apre però una grave tensione, tuttora irrisolta, tra la trasparenza/spiegabilità e l'efficienza delle decisioni automatizzate: in molti ambiti i sistemi più efficaci (in particolare, le reti per l'apprendimento profondo) sono anche quelli il cui comportamento risulta più opaco.[131]

I sistemi di IA, avendo accesso a grandi masse di dati su ognuno di noi, e su quanti sono simili a noi, sono in grado di utilizzare tali informazioni per suscitare comportamenti desiderati, per scopi che potremmo non condividere, possibilmente violando le aspettative fiduciarie riposte su chi controlla i sistemi in questione.[132] Pertanto, vi è un interesse a non essere fuorviati o manipolati dai sistemi di IA, ma anche a poter fare affidamento su di essi, sapendo che non approfitteranno della nostra esposizione e delle nostre debolezze. Una fiducia ragionevole nell'IA è necessaria per evitare che le limitate capacità cognitive dell'uomo siano sprecate nello sforzo di respingere le insidie dei sistemi intelligenti (che nell'interazione con il singolo possono avvalersi di enormi capacità di calcolo con costi marginali vicini allo zero).

Per esempio, le scelte dei consumatori possono essere guidate da piattaforme digitali che rendono certe scelte meno accessibili o dirigono le limitate capacità cognitive dei consumatori verso risultati di cui gli stessi potrebbero pentirsi, o verso scelte che non avrebbero adottato se fossero stati meglio informati. I consumatori si trovano infatti in una posizione di debolezza di fronte a persuasori automatici che hanno accesso ad enormi quantità di conoscenza, che possono dispiegare un illimitato potere computazionale, e possono conformare il contesto delle azioni delle persone (le interfacce presentate e le opzioni offerte) e le informazioni disponibili alle stesse. Inoltre, l'esigenza di catturare gli utenti delle piattaforme può generare fenomeni conosciuti come bolle del filtro (*filter bubble*) e camere di risonanze (*echo chamber*).[133] Come si è già osservato, le informazioni inviate agli utenti sono spesso selezionate sulla base di quanto tali informazioni possano catturare l'attenzione delle persone, indurli a restare sulla piattaforma e stimolarne comportamenti desiderati. Gli stessi dati raccolti per la pubblicità personalizzata (per es., ricerche online, pagine web visitate, attività online, come like e connessioni sociali) possono anche essere usati per influenzare le opinioni e le scelte politiche, come evidenziato nel recente caso di Cambridge Analytica.

I cittadini e i consumatori hanno anche un interesse ad una corretta concorrenza algoritmica, cioè a non essere soggetti ad abusi resi possibili dallo sfruttamento di posizioni dominanti sul mercato, che risultano dal controllo esclusivo di grandi masse di dati e dall'accesso privilegiato a tecnologie avanzate. Questi squilibri riguardano direttamente le imprese concorrenti, ma influiscono negativamente anche sui consumatori: le imprese dominanti possono limitare le scelte dei consumatori o imporre loro condizioni sfavorevoli. Recentemente, sulla scia di questa preoccupazione l'autorità per la concorrenza tedesca (*Bundeskartellamt*) ha imposto un'elevata sanzione a Facebook, per aver abusato della propria posizione dominante nelle reti sociali: Facebook imponeva ai propri utenti di accettare la gestione unificata dei dati personali raccolti mediante diversi servizi controllati da Facebook stessa.

3.7 Diritto per l'IA

L'IA può promuovere o invece pregiudicare i diritti fondamentali e i valori sociali. Con riferimento ai diritti garantiti dalla Carta dei diritti fondamentali dell'Unione Europea, possiamo ricordare i seguenti: diritto alla vita privata (Art. 7) e alla protezione dei dati (Art. 7 e 8), dignità (Art. 1), diritto alla libertà e sicurezza, (Art. 6), libertà di pensiero, coscienza e religione (Art. 10), libertà di espressione e d'informazione (Art. 11), libertà di riunione e di associazione (Art. 12), libertà delle arti e delle scienze (Art. 13), diritto all'istruzione (Art. 14), libertà professionale e diritto di lavorare (Art. 15), libertà d'impresa (Art. 16), diritto di proprietà (Art. 17), diritto di asilo (Art. 18), uguaglianza davanti alla legge (Art. 20), non discriminazione (Art. 21), parità tra uomini e donne (Art. 23), diritti del bambino (Art. 24), condizioni di lavoro giuste ed eque (Art. 31), sicurezza sociale e assistenza sociale (Art. 34), protezione della salute (Art. 35), tutela dell'ambiente (Art. 37), protezione dei consumatori (Art. 38), diritto a una buona amministrazione (Art. 41), diritto a un ricorso effettivo e a un giudice imparziale (Art. 47).[134]

Oltre ai diritti individuali sono in gioco valori sociali riconosciuti dalle costituzioni nazionali e dai trattati dell'Unione Europea, come la democrazia, la pace, il benessere, la concorrenza, il dialogo sociale, l'efficienza, lo sviluppo di scienza, arte e cultura, la sicurezza interna ed internazionale.

L'IA, avendo un impatto così ampio sulla vita individuale e sociale, è già oggi disciplinata da diversi settori del diritto. Si tratta, in particolare, della protezione dei dati, la protezione dei consumatori e il diritto della concorrenza. Come osservava il Garante europeo della protezione dei dati, nell'Opinione 8/18, "c'è sinergia tra i tre regimi. La protezione dei dati e la protezione dei consumatori condividono l'obiettivo di correggere gli squilibri nel potere basato sull'informazione e sul mercato, e con il diritto della concorrenza, contribuiscono ad assicurare che le persone siano trattate equamente".

Altri settori del diritto sono implicati nell'IA: il diritto del lavoro per le nuove forme di controllo sui lavoratori; il diritto amministrativo per opportunità e rischi nell'uso dell'IA nelle decisioni amministrative; la responsabilità civile per i danni causati da sistemi governati dall'IA; il diritto dei contratti, rispetto all'uso dell'IA nella predisposizione di contratti e nell'adempimento delle obbligazioni contrattuali; la disciplina della propaganda politica e delle elezioni, per l'uso dell'IA nelle campagne politiche; il diritto militare per l'uso dell'IA nei conflitti armati, e così via.

Nelle pagine seguenti si considerano alcune normative che affrontano direttamente il tema dell'IA: il Regolamento sulla protezione dei dati, la Proposta di Regolamento sull'IA, la Proposta di Direttiva sulla responsabilità extracontrattuale per i danni causati da sistemi di AI. Inoltre si introducono due temi sui quali si è già formato un ampio dibattito politico e giuridico: la disciplina dei robot e delle armi intelligenti. Infine si propongono alcune considerazioni sulla possibilità di usare l'AI a sostegno della tutela giuridica delle persone.[135]

3.7.1 L'IA nel Regolamento sulla protezione dei dati

Il Regolamento Generale sulla Protezione dei Dati, il cosiddetto GDPR (General Data Protection Regulation), non affronta direttamente il tema dell'IA, ma tuttavia contiene norme rilevanti in tema di informazione, profilazione e decisione automatica [136]

Il GDPR non vieta la profilazione (come definita dall'Art. 4, num. 1), ma richiede che essa abbia una base giuridica (Art. 6), e che si fondi su "procedure matematiche o statistiche appropriate" (Considerando 71), in conformità agli orientamenti indicati dal Comitato europeo per la protezione dei dati (Considerando 72).

Al fine di determinare la liceità della profilazione, anche in presenza di una base giuridica, le linee guida elaborate dal Gruppo di Lavoro Articolo 29 (Opinione 2016/679) richiedono che siano presi in considerazione i seguenti fattori:

- il livello di dettaglio e la completezza del profilo (se questo descrive solo aspetti parziali della persona interessata, o ne ricostruisce un quadro più completo);
- l'impatto della profilazione sull'interessato; e
- le misure di sicurezza volte ad assicurare equità, non discriminazione e accuratezza nel processo di profilazione.

Il GDPR sancisce, invece, un generale divieto di sottoporre gli individui a processi decisionali completamente automatizzati, compresa la profilazione, in grado di produrre effetti giuridici o di incidere in modo significativo sull'interessato (Art. 22).[137] Tuttavia vi sono ampie eccezioni a tale divieto. Le decisioni automatiche sono consentite qualora il loro uso: (i) sia necessario per la conclusione o l'esecuzione di un contratto tra l'interessato e il titolare, (ii) sia autorizzato da una legge o un regolamento, o (iii) sia basato sul consenso esplicito dell'interessato.

Inoltre, nelle ipotesi di profilazione e processi decisionali automatizzati, il GDPR garantisce all'interessato il diritto di esserne informato (Art. 13, 14, 22 e Considerando 71 del GDPR). In particolare, in capo al titolare del trattamento sussiste l'obbligo di informare gli interessati circa (i) le modalità e le finalità della profilazione, (ii) la sua logica e (iii) le sue conseguenze.

Il diritto di informazione è accompagnato dal diritto di opporsi alla profilazione (articolo 21), di richiedere la cancellazione dei propri dati e del proprio profilo (articolo 17), e di contestare le decisioni automatizzate (Art. 22(3)).

3.7.2 La proposta di un regolamento sull'IA

La Commissione Europea ha recentemente presentato la proposta di un regolamento (provvedimento direttamente applicabile) dell'Unione teso a disciplinare in modo organico l'IA (Legge sull'IA).[138]

Come abbiamo già osservato (Sezione 1.1.3), il Regolamento adotta un concetto estremamente ampio di IA, che si estende a tutte le applicazioni che adottino "approcci di apprendimento automatico", "approcci basati sulla logica e approcci basati sulla conoscenza" e "approcci statistici, stima bayesiana metodi di ricerca e ottimizzazione".

Il Regolamento adotta una prospettiva "basata sul rischio": il suo obiettivo non è quello di assicurare ai singoli nuovi rimedi giuridici contro i pregiudizi che derivino da usi vietati dell'IA, ma piuttosto di prevenire quei pregiudizi (*a*) regolando sviluppo, distribuzione e uso di sistemi di IA, e (*b*) predisponendo standard e controlli, la cui definizione e verifica è affidata a strutture private e pubbliche. Il Regolamento adotta quindi una prospettiva di tipo amministrativo-preventivo, simile a quella solitamente usata nella disciplina dei prodotti alimentari, dei dispositivi medici, o della sicurezza sul lavoro.

La disciplina prevista dal Regolamento si fonda su una classificazione delle applicazioni di IA in diverse categorie di rischio, come risulta dalla Figura 3.2.

Le applicazioni che comportano un rischio inaccettabile per individui e società sono vietate dall'Art. 5 del regolamento. Si tratta delle seguenti categorie:

- sistemi che utilizzano "tecniche subliminali che agiscono senza che una persona ne sia consapevole al fine di distorcerne materialmente il comportamento in un modo che provochi o possa provocare a tale persona o a un'altra persona un danno fisico o psicologico";
- sistemi che sfruttano "le vulnerabilità di uno specifico gruppo di persone, dovute all'età o alla disabilità fisica o mentale, al fine di distorcere materialmente il comportamento di una persona che appartiene a tale gruppo in un modo che provochi o possa provocare a tale persona o a un'altra persona un danno fisico o psicologico";
- sistemi impiegati da pubbliche amministrazioni "ai fini della valutazione o della classificazione dell'affidabilità delle persone fisiche" mediante l'attribuzione di un

"punteggio sociale" che possa condurre a un trattamento sfavorevole (a) in contesti diversi da quelli che hanno originato i dati oppure (b) in modo ingiustificato o sproporzionato;

- sistemi di identificazione biometrica remota "in tempo reale" in spazi accessibili al pubblico usati a fini di attività di contrasto dell'illegalità, a meno che non si tratti della ricerca di vittime di reati, della prevenzione di gravi minacce, o della ricerca di autori di gravi reati.

Quindi il Regolamento, nel vietare solo questi usi dell'IA, assume che gli altri usi dell'IA siano invece in linea di principio leciti, sempre che non violino altre norme giuridiche. Non sono mancate le voci che hanno auspicato un'estensione dei divieti di cui all'Art. 5. In particolare, si è criticati i seguenti aspetti: la limitazione dei divieti di cui ai punti (1) e (2) solo ai danni fisici o psicologici, con l'esclusione dei danni economici e morali; la limitazione del divieto di cui al punto (3) alle attività della pubblica amministrazione; le deroghe al divieto dei sistemi di identificazione biometrica di cui al punto (4), che consentono sia messa in funzione un'infrastruttura per la sorveglianza, potenzialmente utilizzabile anche ad altri scopi. A queste critiche si è opposta l'esigenza che il legislatore intervenga con cautela rispetto all'IA, settore nuovo e in rapida evoluzione.

L'Art. 53 del Regolamento introduce alcuni obblighi di informazione, volti a far sì che le persone non siano ingannate o manipolate da sistemi di IA;

- le persone che interagiscono con un sistema di IA debbono essere informate che il loro interlocutore è un sistema artificiale anziché un essere umano, tranne che ciò risulti evidente dalle circostanze e dal contesto di utilizzo;[139]
- le persone esposte a un sistema di riconoscimento delle emozioni o a un sistema di categorizzazione biometrica debbono essere informate sul funzionamento di tale sistema;
- si deve rendere noto che un contenuto è stato generato o manipolato artificialmente, qualora quel contenuto possa apparire "falsamente autentico o veritiero".

Le previsioni più importanti del Regolamento sono quelle rivolte alle applicazioni ad alto rischio, che sono lecite a condizione che rispettino le prescrizioni che le riguardano, contenute nel Titolo III. Si tratta dei seguenti sistemi (Art. 6):

- i sistemi di IA che siano disciplinati da normative su prodotti ad elevato rischio (le normative di armonizzazione elencate nell'allegato II) e per i quali sia prevista una valutazione della conformità da parte di terzi;
- i sistemi di IA elencati all'allegato III del regolamento (Identificazione e categorizzazione biometrica; Gestione e funzionamento delle infrastrutture critiche; Istruzione e formazione professionale; Accesso a prestazioni e servizi pubblici e

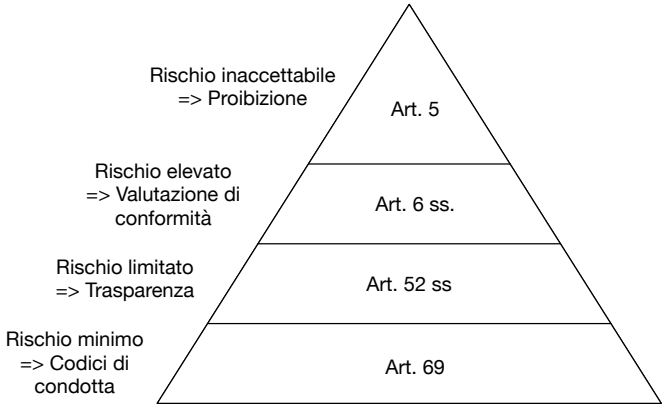

Figura 3.2: *Il regolamento sull'IA*

a servizi privati essenziali e fruizione degli stessi; Attività di contrasto; Gestione della migrazione, dell'asilo e del controllo delle frontiere; Amministrazione della giustizia e processi democratici).

Chi fornisca o impieghi un sistema ad alto rischio deve garantire che il sistema rispetti i seguenti requisiti:

- istituire, attuare, documentare e mantenere un sistema di gestione dei rischi (Art. 9);
- adottare criteri di qualità per i dati usati nell'addestramento (Art. 10);
- predisporre documentazione tecnica che dimostri il rispetto dei requisiti applicabili (Art. 11);
- consentire la registrazione automatica degli eventi ("log") durante il loro funzionamento (Art. 12);
- garantire che il loro funzionamento sia sufficientemente trasparente da consentire agli utenti di interpretare l'output del sistema e utilizzarlo adeguatamente (Art. 13);
- poter essere efficacemente supervisionati da persone fisiche (Art. 14);
- conseguire un adeguato livello di accuratezza, robustezza e cibersicurezza (Art. 15);

Anche le disposizioni sui sistemi ad alto rischio sono oggetto di un ampio dibattito: alcuni ritengono che i requisiti elencati non siano sufficienti ad assicurare l'affidabilità dei sistemi di IA; altre invece li ritengono eccessivamente restrittivi e tali da compromettere la competitività dell'industria europea.

Il regolamento prevede, infine, una struttura istituzionale volta ad assicurarne l'applicazione, struttura che si sviluppa in due direzioni.

Innanzitutto, è previsto che la valutazione dei sistemi di IA sia affidata a "organismi di valutazione della conformità," le cui capacità debbono essere previamente valutate da "autorità di notifica", designate dagli Stati membri (Art. 30 e 31). Tali organismi emettono certificati di conformità che consentono l'impiego dei sistemi ad alto rischio.

In secondo luogo, ciascuno Stato membro dovrà istituire o designare autorità nazionali competenti al fine di garantire l'applicazione e l'attuazione del Regolamento (Art. 59). Al vertice di tali autorità sarà nominata un'Autorità Nazionale di Controllo, con funzioni di autorità di notifica e di autorità di vigilanza del mercato.

Le Autorità Nazionali di Controllo, e il Garante Europeo per la Protezione dei Dati formano il Comitato Europeo per l'IA, che fornisce assistenza alla Commissione e alle Autorità Nazionali di Controllo, e ne coordina le attività.

3.7.3 La Proposta di Direttiva in tema di responsabilità extracontrattuale e IA

La Commissione Europea ha recentemente presentato una Proposta di Direttiva sulla responsabilità extracontrattuale per i danni causati da sistemi di IA.[140] La Direttiva affronta due temi principali:

- l'acquisizione di elementi di prova attinenti al funzionamento di sistemi di IA ad alto rischio, come identificati dal Regolamento sull'IA (Art. 3);
- l'onere della prova per la responsabilità extracontrattuale per danni causati da sistemi di IA (Art. 4).

L'Art. 3 mira ad agevolare l'attore nell'acquisizione di elementi di prova, compito estremamente difficile, data la complessità e opacità dei sistemi di IA, rispetto ai quali di regola il convenuto ha esclusivo controllo e superiori conoscenze.

Il comma 1, statuisce che i giudici nazionali debbono avere il potere di ordinare —ai fornitori e utilizzatori di sistemi di IA ad alto rischio che si sospetta abbiano cagionato danni— la divulgazione di elementi di prova. L'esercizio del potere presuppone: (*a*) la richiesta di un un attore in giudizio che abbia compiuto "ogni sforzo proporzionato" per ottenere gli elementi di prova (Art. 3, comma 2), o (*b*) la richiesta di un potenziale attore che avendo richiesto tali elementi, si sia visto opporre un rifiuto, e adduca fatti e prove sufficienti a sostenere la plausibilità della domanda di risarcimento del danno.La divulgazione degli elementi di prova deve limitarsi "a quanto necessario e proporzionato per sostenere una domanda o potenziale domanda di risarcimento del danno" (Art. 3, comma 4).

L'Art. 4 alleggerisce l'onere della prova dell'attore (il preteso danneggiato), qualora il convenuto abbia violato un obbligo di diligenza. Il comma 1 stabilisce che gli

organi giurisdizionali nazionali debbono presumere il nesso di causalità tra la colpa del convenuto e il cattivo funzionamento del sistema di AI (consistente nella produzione di un certo output o anche nella mancata produzione di un certo output) date le seguenti condizioni:

(a) l'attore ha dimostrato la violazione di obbligo di diligenza "direttamente inteso a proteggere dal danno verificatosi";

(b) è ragionevolmente probabile che il comportamento colposo abbia influito sul funzionamento del sistema;

(c) l'attore ha dimostrato che il danno è stato causato dal funzionamento del sistema.

Si tratta di una presunzione relativa, che il convenuto può confutare dimostrando che il danno non è dovuto a sua colpa.

3.7.4 Il contrasto alla disinformazione

Recentemente, sono state adottate alcune iniziative volte a contrastare la disinformazione, che sempre più spesso viene realizzata grazie all'uso di tecnologie di IA per la generazione e distribuzione di contenuti. Un passo in questa direzione può ravvisarsi nella recente comunicazione della Commissione Europea sulla lotta alla disinformazione online, volta ad intensificare gli sforzi per contrastare tale fenomeno.[141] In particolare, ci si propone di migliorare la capacità di rilevamento e analisi dell'esposizione dei cittadini alla disinformazione, la cooperazione e la capacità di elaborare risposte comuni a tali minacce in collaborazione con le piattaforme online e l'industria, e la resistenza della società mediante campagne e strumenti di sensibilizzazione. La Commissione ha altresì proposto numerose azioni tra cui il sostegno al giornalismo di qualità, mediante aiuti di stato da parte degli Stati membri, e il coinvolgimento attivo delle piattaforme online e dell'industria pubblicitaria.[142] È stato elaborato nel 2018 un codice di condotta sulla disinformazione (*Code of Practice on Disinformation*), sottoscritto da importanti rappresentanti delle piattaforme e reti sociali online, di cui è stata presentata nel 2022 una nuova versione ampliata (*Strenghtened Code of Practice on Disinformation*). Il codice identifica una serie di azioni che i firmatari devono mettere in atto, tra cui un sistema di valutazione delle fonti e della qualità dei contenuti. Il codice richiede inoltre l'uso di strumenti tecnologici per l'identificazione e la chiusura di profili falsi, per facilitare la diffusione di informazioni pertinenti, autentiche, accurate e autorevoli, per ottenere una maggiore trasparenza. Tali strumenti dovranno consentire agli utenti di avere informazioni sull'identità di chi ha inviato un messaggio o ne ha promosso la trasmissione, e di ottenere altresì indicatori dell'affidabilità delle fonti di informazione. L'IA può contribuire all'individuazione di testi e immagini che riportano notizie false, artatamente presentate come vere e autentiche (le cosiddette *fake news*), grazie alle tecnologie per

l'analisi automatica del linguaggio parlato e scritto, delle immagini, dei filmati e dei flussi di informazione.[143]

3.7.5 La disciplina dei robot

Oggi il mondo è popolato di milioni di robot che operano nei campi più diversi: l'industria, i servizi (per esempio, pulizia), i trasporti, l'ambito militare, la sicurezza, la ricerca (per es., esplorazioni sottomarine), l'istruzione, la medicina, l'intrattenimento, l'assistenza.

La diffusione dei robot ha determinato nuovi problemi giuridici, se non l'emergere di un nuovo ambito del diritto. La peculiarità giuridica dei robot discende dalla combinazione di due aspetti: (a) le loro capacità cognitive, che li rende capaci di svolgere con autonomia compiti complessi, anche senza controllo umano, e (b) la loro dimensione fisica, che li rende capaci di causare danni materiali anche gravi. Nella misura in cui è dotato di autonomia, il robot stesso determina quali comportamenti adottare, in base alle sue competenze cognitive, ai compiti che gli sono assegnati, e al contesto in cui attuarli.[144] Chi usa un robot (e più in generale, una tecnologia di IA) delega allo stesso il compimento di certe attività cognitive (per esempio, il riconoscimento delle condizioni sulla strada e le conseguenti determinazioni circa la guida): il delegante, avendo rinunciato a compiere egli stesso tali funzioni, non ne conoscerà l'esito. Quindi, il robot può tenere comportamenti non previsti, e non ragionevolmente prevedibili da parte di chi lo utilizza. Il problema è particolarmente significativo qualora il robot sia in grado di apprendere dalla propria esperienza, sviluppando competenza e attitudini che non gli erano state attribuite dai suoi creatori. Ci si è quindi chiesti in quale modo si debba concepire e qualificare giuridicamente il rapporto tra l'utilizzatore e il robot: secondo il modello dei rapporti di titolarità o controllo tra persone e cose (tra proprietario, custode, o utilizzatore e la cosa posseduta, custodita, o utilizzata), o invece secondo il modello dei rapporti di direzione o vigilanza tra persone (tra il committente, datore di lavoro, genitore, tutore, o precettore e il commesso, dipendente, o individuo soggetta a potestà, tutela o vigilanza).

Un secondo problema attiene al fatto che diversi soggetti contribuiscono a determinare il funzionamento di un robot: il progettista che l'ha disegnato, il produttore che l'ha costruito, il programmatore-adattatore che l'ha fornito di contenuti e attitudini ulteriori, l'istruttore che gli ha insegnato come svolgere certe attività, e infine chi l'ha utilizzato in un certo contesto, assegnandogli certi compiti, da svolgere secondo certe istruzioni. Nel caso di un malfunzionamento può essere molto difficile distinguere i contributi di ciascuno.

Il tema maggiormente trattato nell'ambito del diritto dei robot attiene alla responsabilità per i danni causati dagli stessi. Sono ormai numerosi i casi in cui robot sono stati coinvolti in gravi incidenti anche mortali. Spesso si tratta di incidenti sul lavoro, ma non mancano casi di decessi dovuti al malfunzionamento di robot in campo me-

dico e militare, e recentemente si sono verificati i primi incidenti mortali in cui sono coinvolte automobili o aerei senza pilota.[145] Ci si è chiesti[146] se la disciplina della responsabilità civile per i danni causati da robot, nell'ambito della loro azione autonoma, sia da concepirsi come la responsabilità di una persona per i danni causati dalla cosa che possiede o custodisce (di regola, una responsabilità oggettiva) o invece come una più limitata responsabilità per fatto altrui (di regola suscettibile di esonero nel caso di provata impossibilità di impedire il fatto). Un'attenuazione della responsabilità per il danno da robot, rispetto a quella per il danno da cose, si è argomentata considerando da un lato l'impossibilità di prevedere l'azione autonoma del robot, e dall'altro l'esigenza di stimolare l'impiego dei robot (rassicurando gli utilizzatori rispetto a responsabilità non prevedibili). Alcuni sostenitori di una limitazione della responsabilità per il danno da robot hanno, peraltro, affermato la necessità di garantire una tutela ai danneggiati, ricorrendo a forme di responsabilità oggettiva limitata da massimali e garantita da assicurazioni obbligatorie.[147]

La dottrina più recente ha affrontato altresì il tema della responsabilità penale per l'attività dei robot. Il problema principale attiene al fatto che nell'esercizio della loro autonomia i robot possono compiere azioni che costituirebbero reati se fossero compiute da persone (danneggiamento, percosse, traffico di stupefacenti, omicidio, etc.). In alcuni casi è possibile individuare una partecipazione di un essere umano, che quindi può essere chiamato a rispondere personalmente (ad esempio, per aver richiesto al robot il compimento dell'attività criminale), ma in altri casi può non esservi alcuna persona che possegga lo stato mentale (dolo o colpa) richiesto per il reato in questione. Ci si è così interrogati sulla possibilità di considerare il robot stesso responsabile penalmente,[148] o sulle possibili tecniche per allocare una responsabilità penale per negligenza o una responsabilità civile a progettisti e utilizzatori.[149] Un secondo importante tema attiene all'attività contrattuale compiuta da automi intelligenti.[150] Tale tema riguarda i robot dotati di dimensione fisica, ma ancor più gli agenti software o *softbot*, cioè i sistemi informatici capaci di attività autonoma in ambiti virtuali. Tali sistemi sono utilizzati soprattutto su Internet, dove operatori e utenti ricorrono ad agenti software per cogliere e sfruttare le possibilità offerte dallo sconfinato, diversificato e dinamico ambiente virtuale del ciberspazio. Gli agenti software possono essere usati in particolare nella raccolta di informazioni (ricerca di dati, offerte concernenti beni e servizi, o possibili partner commerciali) e nell'attività contrattuale (trattativa, formulazione di proposte, valutazione della loro convenienza, conclusione di contratti). Essi possono utilizzare tecniche di IA per accedere alle informazioni utili (ad esempio, le opportunità offerte dal mercato), accertare le circostanze rilevanti per le loro azioni (per esempio, la convenienza comparativa delle diverse offerte), e agire conseguentemente (effettuando l'acquisto più conveniente). Agenti software sono utilizzati largamente nei mercati dei titoli azionari e obbligazionari, in particolare nel c.d. *high frequency trading* (negoziazione ad alta frequenza), cioè per scambi che si svolgono in frazioni di secondi.

Un profilo giuridico particolarmente interessante è quello della qualificazione dell'attività giuridica realizzata dall'agente nell'interesse del suo utilizzatore, eseguendo il compito affidatogli dall'utilizzatore stesso. Si è affermato che si potrebbero forse applicare per analogia le norme sulla rappresentanza qualora l'utilizzatore deleghi al suo aiutante elettronico lo svolgimento di un compito da realizzare con autonomia (compito che includa attività cognitive, come quelle necessarie per individuare la controparte di un contratto e determinare il contenuto del contratto stesso). Conseguentemente, già in base al diritto vigente, gli effetti giuridici ricadrebbero sull'utilizzatore, ma gli stati psicologici rilevanti (la volontà, la conoscenza, l'errore, ecc.) sarebbero quelli dell'agente digitale (per quanto riguarda gli elementi del contratto determinati dall'agente stesso).[151]

Un tema importante del diritto dei robot è la tutela della privacy di utilizzatori e terzi.[152] I robot, tramite i propri sensori, attingono dallo spazio fisico dati personali d'ogni tipo (immagini, suoni, ecc.), cui si aggiungono quelli forniti dall'utilizzatore interagendo con il robot. Quei dati possono essere memorizzati indefinitamente, elaborati e trasmessi. Le possibilità di raccolta di dati personali, anche sensibili, si accresce per i robot mobili in grado di visitare diversi ambienti, e per quelli destinati a svolgere funzioni di cura e assistenza. Ci si chiede come l'attività di elaborazione di dati personali dei robot possa essere limitata ai trattamenti consentiti dalla disciplina della protezione dei dati, dato che i robot agiscono autonomamente per il perseguimento degli scopi loro assegnati, e raccolgono dall'ambiente le informazioni utili per realizzare quegli scopi. Sembra che i robot possano rispettare la privacy di utenti e terzi solo se il rispetto della privacy diventa un principio interno al loro stesso funzionamento, cioè se essi sono stati progettati secondo modelli di *privacy by design* (privacy fin dalla progettazione). Un pieno risultato potrebbe aversi solo quando i robot stessi fossero agenti normativi, capaci di rispondere alle norme giuridiche (espresse in un linguaggio comprensibile ai robot stessi) astenendosi dai comportamenti vietati. Il tema del rispetto del diritto da parte dei robot si collega quindi a quello della rappresentazione formale delle norme giuridiche (Sezione 4.1).

Infine, ci si è interrogati se in futuro si potrà attribuire ai robot una personalità giuridica, cioè l'attitudine ad avere posizioni giuridiche (diritti e doveri) proprie. Potrebbe trattarsi innanzitutto di una forma di mera autonomia patrimoniale. Il robot (o il softbot) potrebbe essere assimilato a una società commerciale, e dotato di un proprio patrimonio,[153] e potrebbe quindi essere direttamente responsabile per i danni commessi e le obbligazioni da esso assunte nell'esercizio della propria autonomia. Ciò potrebbe facilitare l'uso dei robot riducendo il rischio degli utilizzatori, ma garantendo comunque una tutela alle terze parti che interagiscono con il robot. C'è anche chi si è chiesto se in futuro si dovrebbe attribuire ai robot —sempre più intelligenti, e dotati di una propria dimensione valutativa, volitiva— una personalità giuridica più ampia, inclusiva anche di alcuni diritti della personalità, e in particolare di una protezione delle propria integrità fisica (hardware) e psichica (software).[154]

3.7.6 Le armi intelligenti

Molte applicazioni dell'IA sono state costruite per scopi militari. Può trattarsi di applicazioni software, per la raccolta di dati e l'estrazione di informazioni (intelligence) o per la difesa o l'attacco nella guerra informatica a distanza (*cyberwar*). Può anche trattarsi di sistemi robotici armati, guidati dall'IA nell'uso della forza fisica (cinetica), anche letale. Si parla in particolare di "armi autonome" per designare i dispositivi in grado di operare in un teatro di guerra senza essere dirette, in tutto o in parte, dall'uomo. Pensiamo ai droni (aerei senza pilota) dotati di missili o bombe; alle "bombe intelligenti" in grado di procedere da sole verso il loro bersaglio, ai veicoli militari terrestri armati in grado di spostarsi e far fuoco senza equipaggio a bordo.

Il grado di autonomia di questi dispositivi è variabile: di regola essi sono guidati e controllati a distanza, ma possono in alcuni casi prendere l'iniziativa, specialmente quando il controllo a distanza non sia disponibile o quando sia necessaria una reazione rapida. L'evoluzione tecnologica va certamente nel senso di una loro crescente autonomia, che giunge in alcuni casi alla capacità di attivarsi autonomamente nel perseguimento, e in alcuni casi nella stessa individuazione, di obiettivi contro i quali esercitare forza letale. Per esempio, un drone potrebbe individuare da solo il soggetto (per esempio, il supposto terrorista) da colpire sulla base delle informazioni ricevute (per esempio, un'immagine del volto dello stesso), valutare se il rischio di danni collaterali sia accettabile, e procedere ad ingaggiare il bersaglio.

Molti auspicano la proibizione, o almeno una moratoria, dell'uso di armi robotiche che agiscano in modo del tutto autonomo, in particolare nell'uso di forza letale contro esseri umani. Si afferma che l'uso di queste armi condurrebbe a nuove gravi violazioni del diritto umanitario, la disciplina che governa la condotta nei conflitti armati.[155] In particolare si è osservato che le armi autonome,[156] almeno allo stato della tecnica, non sarebbero in grado di applicare i principi fondamentali del diritto umanitario, cioè i principi di distinzione (il divieto di rivolgere le armi contro obiettivi civili) e di proporzionalità (l'esigenza di limitare al massimo i danni collaterali, in particolare quelli che riguardano i civili). Si è osservato altresì che l'uso delle armi autonome potrebbe condurre ad una nuova corsa alle armi, con esiti potenzialmente distruttivi[157] e in ogni caso favorire l'uso della forza soprattutto da parte delle nazioni che godono di preminenza tecnologica (la c.d. guerra asimmetrica). Altri autori, tuttavia, hanno obiettato che non è possibile fermare la corsa verso armi sempre più intelligenti, poiché le tecnologie dell'IA usate nelle armi sono impiegate anche a scopi civili. Inoltre, si è osservato che l'uso della forza da parte di sistemi autonomi potrebbe accrescere il rispetto del diritto, poiché quei sistemi non sarebbero soggetti alle passioni (paura, rabbia, vendetta, crudeltà) che conducono ad atrocità negli scenari di guerra, e potrebbero essere costruiti in modo da garantire la legalità del loro comportamento.[158]

3.7.7 Tecnologie di IA per l'etica e il diritto

Per indirizzare l'IA verso il bene degli individui e della società è necessaria un'adeguata cornice normativa, comprensiva di indicazioni etiche e norme giuridiche, principi generali e dettagliate discipline di settore. Tuttavia, si può dubitare che la specificazione norme e principi, e la loro attuazione da parte dei poteri pubblici sia sufficiente.

Infatti, le tecnologie dell'IA e dei big data operano in settori già caratterizzati da grandi squilibri di potere, che esse contribuiscono ad accentuare. Tali tecnologie, infatti, creano nuove conoscenze (capacità di analisi e previsione) e poteri (capacità di controllo e direzione) e le mettono a disposizione di chi governa le stesse tecnologie, cioè delle grandi imprese e dei poteri pubblici che cooperano con esse. Non solo l'individuo isolato, ma anche le sue organizzazioni, prive di analoghe risorse, si trovano in una crescente posizione di svantaggio, nella quale diventa difficile avvalersi delle tutele giuridiche astrattamente disponibili.

Un possibile, seppur parziale, rimedio si può individuare stabilendo un parallelismo tra le dinamiche di potere sottese allo sviluppo dell'IA e quelle relative alla società industriale e alla società dei consumi di massa. In entrambi i casi, un limite agli eccessi del mercato, cui faceva riferimento Polanyi, si trovò in movimenti sociali, quali quello dei lavoratori e dei consumatori. Come osservava l'economista Ken Galbraith, per assicurare un'adeguata protezione ai cittadini, non sono sufficienti gli strumenti normativi e la loro attuazione da parte di organi pubblici, ma sono altresì necessari contro-poteri o poteri compensativi (*countervailing power*) della società civile.[159] I cittadini e le loro organizzazioni possono individuare abusi, informare il pubblico, promuovere l'applicazione delle norme, ed esercitare forme di pressione collettiva.

Per essere efficace, tuttavia, un contropotere deve disporre di mezzi adeguati a quelli a disposizione del potere cui si oppone. Nell'era dell'IA l'esercizio di contropoteri da parte della società civile presuppone che anche la società civile sia in grado di avvalersi dell'IA. Solo se i cittadini e le loro organizzazioni saranno in grado di utilizzare l'IA a loro vantaggio potranno resistere e rispondere alle imprese e ai governi il cui potere è sostenuto dall'IA. Nel rendere efficace l'azione dei cittadini, l'IA può favorire una cittadinanza attiva, essenziale valore democratico, in opposizione al modello del cittadino manipolabile con micro-ricompense e micro-sanzioni (come nel caso del Sistema di Credito Sociale cinese).

Alcuni esempi di tecnologie che conferiscono potere agli individui sono già presenti, come i sistemi per il blocco della pubblicità o i più tradizionali software antispam e antifrode. A tutela dei consumatori, si stanno sviluppando nuovi sistemi per riassumere e combinare recensioni, comparare i prezzi, e consentire ai consumatori di coordinare le proprie azioni. Tali strumenti potrebbero contribuire a nuovi mercati nei quali i consumatori si avvalgano di agenti digitali per negoziare, formare coalizioni, ed effettuare scelte di acquisto comuni. Anche i metodi per l'estrazione e l'analisi dei dati possono essere utili alla società civile, ad esempio, per identificare pratiche discriminatorie, in

settori quali la concessione del credito, l'assunzione al lavoro, la concessione di benefici sociali.

L'apprendimento automatico e le tecnologie per l'elaborazione del linguaggio naturale possono essere altresì utilizzati per identificare casi in cui vengano raccolti dati inutili o eccessivi. Infine, tali tecnologie possono essere usate per l'analisi e la validazione del contenuto di documenti testuali. In particolare, esse possono essere impiegate per individuare contenuti illeciti o omissioni di contenuti obbligatori, come vedremo e con riferimento al sistema Claudette (Sezione 4.2.1).

Capitolo 4

L'IA: applicazioni giuridiche

Nel presente capitolo si illustrano i principali indirizzi per la creazione di applicazioni di intelligenza artificiale in ambito giuridico. Dapprima si presentano i modelli "logici", basati sulla rappresentazione esplicita, comprensibile all'uomo, della conoscenza e sulla sua elaborazione mediante metodi di ragionamento. Si esaminano quindi i sistemi basati sull'apprendimento automatico.[160]

4.1 Conoscenza e ragionamento

Non è possibile illustrare in questa sede i diversi metodi del ragionamento logico (o, meglio, dei diversi ragionamenti consentiti da diversi modelli logici), né il modo in cui quei metodi possono essere applicati da sistemi informatici. Ai nostri fini è sufficiente illustrare il tipo di ragionamento più frequente in ambito giuridico, cioè l'applicazione di regole. Per "regola" intendiamo qualsiasi enunciato *condizionale*, che collega un *antecedente* e un *conseguente* consentendo di inferire il secondo dal primo. Questo generalissimo concetto di regola non si limita quindi agli enunciati deontici (che qualificano come obbligatorio, vietato o permesso un certo comportamento, come per esempio "è vietato fumare"), ma si fonda sulla connessione inferenziale tra antecedente e conseguente.

Quindi per noi sono regole tanto l'enunciato "se una persona guida l'automobile, allora deve avere con sé la patente di guida", quanto l'enunciato "se una persona è nata da un cittadino italiano, allora quella persona è cittadino italiano."

Ricordiamo che le regole possono essere espresse tanto nella forma SE *antecedente* ALLORA *conseguente*, quanto nella forma *conseguente* SE *antecedente*, forma che adotteremo di regola negli esempi seguenti. Il concetto di regola come enunciato condizionale (piuttosto che come obbligo o prescrizione) corrisponde all'idea diffusa che le regole giuridiche colleghino una fattispecie e una conseguenza giuridica: la fattispecie astratta è l'antecedente della regola e la conseguenza giuridica astratta ne è il conseguente. L'antecedente può consistere di un solo elemento, o invece di una congiunzione di elementi. L'esempio seguente (che fa riferimento alla regola generale sulla responsabi-

lità civile, di cui all'Art. 2043 cc.) illustra l'inferenza mediante la quale viene applicata una regola:

1. *x* deve risarcire *y* SE
   (a) *x* causa intenzionalmente un danno a *y*
   (b) il danno a *y* è ingiusto
2. *Lucia* causa intenzionalmente un danno a *Renzo*
3. il danno a *Renzo* è ingiusto
   PERTANTO
4. *Lucia* deve risarcire *Renzo*

Nell'esempio, gli enunciati (2) e (3) riportano la fattispecie concreta del caso, che realizza (specifica) la fattispecie astratta della regola (le condizioni 1(a) e 1(b)). La corrispondenza tra fattispecie astratta e fattispecie concreta risulta dal fatto che la fattispecie concreta si ottiene dalla fattispecie astratta sostituendo uniformemente le variabili *x* e *y* con le costanti (i nomi di individui) *Lucia* e *Renzo*. Così, "*x* causa intenzionalmente un danno a *y* " si trasforma in "*Lucia* causa intenzionalmente un danno a *Renzo*". Sulla base di tale corrispondenza possiamo inferire l'effetto giuridico concreto risultante dalla medesima sostituzione, applicata al conseguente della regola. Nell'esempio, si tratta della conclusione concreta "*Lucia* deve risarcire *Renzo*", che realizza (specifica) la conclusione astratta "*x* deve risarcire *y* ". In termini giuridici, abbiamo sussunto la fattispecie concreta (i fatti del caso) alla fattispecie astratta (l'antecedente), e quindi abbiamo derivato il corrispondente effetto giuridico concreto (la specificazione del conseguente).

Si noti che la regola dell'esempio potrebbe essere estesa mediante regole ulteriori, come quella secondo cui vi è un danno ingiusto se il danno consiste nella lesione della proprietà.

1. il danno a *y* è ingiusto SE
   (a) il danno a *y* consiste nella lesione di una proprietà di *y*

Grazie a questa regola il fatto "il danno a *Renzo* è ingiusto" può essere sostituito dal fatto "il danno a *Renzo* consiste nella lesione di una proprietà di *Renzo*", poiché il primo fatto diventa una conseguenza del secondo, grazie alla nuova regola.

Come osservato, mediante regole possiamo esprimere anche norme classificatorie (frammenti di ontologie giuridiche), come le seguenti, che riproducono parzialmente gli articoli 1 e 2 della legge sull'IVA. Come illustrano gli esempi, la scrittura delle regole può essere facilitata ammettendo l'uso anche del connettivo "O" (anziché scrivere due regole *A* SE *B* e *A* SE *C*, scriviamo *A* SE *B* O *C*.

1. x è imponibile ai fini IVA SE
   (a) i. x è una cessione di beni O
   ii. x è una prestazione di servizi
   E
   (b) x è effettuata a titolo oneroso E
   (c) x è effettuata nel territorio dello Stato E
   (d) i. x è effettuata nell'esercizio di un'impresa O
   ii. x è effettuata nell'esercizio di un'arte o professione.
2. x è una cessione di beni SE
   (a) i. x importa trasferimento delle proprietà O
   ii. x importa costituzione o trasferimento di diritti reali di godimento.

4.1.1 I sistemi basati su regole nel diritto

I sistemi basati su regole (*rule-based system*) costituiscono il tipo più semplice e diffuso di sistema basato sulla conoscenza (*knowledge-based system*). Tali sistemi contengono una base di conoscenza costituita da regole, e un motore di inferenza, che applica tali regole ai dati di fatto attinenti a casi concreti. Le informazioni attinenti ai casi possono essere fornite dall'utilizzatore umano del sistema o estratte da archivi informatici.

Sistemi basati su regole giuridiche si realizzarono fin dagli anni Settanta,[161] ma la ricerca che ha maggiormente influenzato il dibattito teorico e le soluzioni applicative fu realizzata presso l'Imperial College di Londra tra la fine degli anni '70 e l'inizio degli anni '80. Un gruppo di studiosi di formazione informatica—coordinato da Robert Kowalski e Marek Sergot, due tra i massimi esperti nella logica computazionale—applicò le nuove tecniche della programmazione logica alla rappresentazione di norme giuridiche e all'esecuzione di inferenze corrispondenti. La ricerca dell'Imperial College mostrò come le norme giuridiche possano essere rappresentate in forma di regole, e possano essere applicate secondo modelli logici efficienti, intuitivi e rigorosi.[162] Negli anni successivi accanto a ricerche sperimentali volte a cogliere ulteriori aspetti del ragionamento giuridico (come il ragionamento basato sui casi, l'argomentazione, ecc.) non mancarono i tentativi di realizzare sistemi basati su regole che trovassero immediato impiego nell'ambito delle attività giuridiche, uscendo dallo stadio prototipale.

La semplicità degli esempi appena riportati nella sezione precedente non deve far pensare che un sistema basato su regole giuridiche sia utile solo nei casi semplici (così da dare all'utilizzatore un aiuto non diverso da quello che potrebbe essere fornito da un manuale di istruzioni). L'importanza di tali sistemi deriva dalla loro capacità di applicare un numero elevato di regole (vi sono sistemi che ne contengono decine di migliaia), tenendo conto dei collegamenti tra le stesse (come quando il conseguente di una regola rappresenti la precondizione per l'applicazione di altre regole o indichi un'eccezione ad

altre regole). Essi possono così supplire ai limiti della memoria, dell'attenzione, e della capacità combinatoria dell'uomo.

Un importante stimolo allo sviluppo di sistemi basati sulla conoscenza può venire dall'adozione di standard per esprimere regole giuridiche (come il Rule Modelling Language - RuleML[163]), così da rendere interoperabili e riutilizzabili le basi di regole realizzate per diversi sistemi.

4.1.2 La scrittura di regole in linguaggio quasi-naturale

Come si è osservato, i sistemi basati sulla conoscenza si basano sulla logica, ma la comprensione delle formule logiche può risultare difficile, soprattutto per chi non abbia familiarità con i linguaggi formali. Alcuni ambienti per lo sviluppo di sistemi basati sulla conoscenza facilitano la redazione e l'esame delle basi di conoscenza, offrendo forme espressive più vicine al linguaggio naturale rispetto a quelle adottate nella logica matematica e nei linguaggi di programmazione logica più diffusi.

Un recente progetto in questa direzione è *Logical English* (un prototipo del sistema è disponibile online, con sorgente aperta).[164] Questo sistema accetta formulazioni vicine al linguaggio naturale e provvede automaticamente a trasformarle in programmi logici, in Prolog o ASP (*Answer set programming*). Ecco nel seguito, come potremmo rappresentare nella versione italiana (Italiano Logico) del sistema Logical English un frammento della disciplina della cittadinanza italiana.

una persona A ha la cittadinanza italiana
    se una persona B è genitore di A
    e B ha la cittadinanza italiana.

una persona A è genitore di una persona B
    se A è il padre di B
    o A è la madre di B.

una persona A ha la cittadinanza italiana
    se A è natx in italia
    e non risulta che
        A segue la cittadinanza del genitore straniero
        o entrambi i genitori di A sono apolidi.

una persona A segue la cittadinanza del genitore straniero
    se una persona B è il genitore di A
    e B è cittadinx di uno stato straniero
    e A acquista la cittadinanza di B.

Assumiamo di fornire al sistema uno scenario come il seguente:

Felice è padre di Giuseppe.
Tatiana è madre di Giuseppe.
Felice ha la cittadinanza italiana.

La rappresentazione appena illustrata consente al sistema di concludere che Giuseppe è cittadino italiano.

4.1.3 L'uso dei sistemi basati sulla conoscenza giuridica

I risultati operativi ottenuti in ambito giuridico mediante sistemi basati sulla conoscenza furono inizialmente inferiori alle grandi attese suscitate da questa tecnologia.[165]

I primi prototipi realizzati non condussero allo sviluppo di sistemi funzionanti, furono rifiutati dai funzionari destinati ad applicarli, furono abbandonati dopo una breve sperimentazione, o non si ritenne di procedere al loro aggiornamento quando le loro basi di conoscenza diventavano obsolete in seguito a modifiche normative.

Negli ultimi anni, tuttavia, queste iniziali delusioni sono state seguite da alcuni importanti successi, in particolare nell'ambito della pubblica amministrazione. La ragione fondamentale di tale cambiamento è da ritrovarsi nell'avvento di Internet, che ha modificato l'uso delle applicazioni informatiche e in particolare dei sistemi basati sulla conoscenza. Grazie a Internet, un sistema disponibile online può essere utilizzato da un numero illimitato di utenti e sue funzioni possono essere integrate con quelle offerte da altri sistemi nell'ambito di architetture distribuite. Ciò significa che il sistema finalizzato a una particolare funzione (per esempio, destinato a determinare il diritto a ricevere certe prestazioni) può essere utilizzato (possibilmente con modalità diverse) sia da parte dei cittadini, sia da parte dei funzionari pubblici competenti, nei diversi uffici che gestiscono la normativa corrispondente. Inoltre, il sistema può essere integrato con altri sistemi informatici (come per esempio, banche di dati, archivi di documenti e formulari, ecc.) in modo tale che le funzioni svolte da ciascuno di tali sistemi siano "orchestrate" per realizzare una funzionalità complessiva.

L'uso di sistemi basati sulla conoscenza può rappresentare uno sviluppo "naturale" dell'uso di Internet nel settore pubblico. Le pubbliche amministrazioni e, più in generale, i soggetti che forniscono servizi pubblici, già negli anni Novanta hanno iniziato a usare Internet per fornire informazioni ai cittadini. Inizialmente, si è trattato di materiale di presentazione dell'ente e delle sue funzioni, cui si accompagna la possibilità di "scaricare" documenti di vario genere (testi normativi, manuali di istruzioni, moduli e formulari). Un passo avanti si è realizzato quando l'utente è stato abilitato a estrarre dagli archivi elettronici della pubblica amministrazione i dati che lo riguardano e a effettuare l'aggiornamento degli stessi. Il passo successivo è consistito nell'applicare automaticamente le regole associate ai documenti giudici, così da agevolare lo svolgimento delle

relative attività amministrative. Si sono così realizzati sistemi che svolgono un ruolo attivo nei processi amministrativi "determinativi", cioè volti a stabilire quali obblighi e diritti spettino ai cittadini in base alla legge, e in particolare a verificare se il cittadino possegga i requisiti per accedere a un certo servizio pubblico e a quali condizioni possa avervi accesso. La tecnologia dei sistemi basati sulla conoscenza può consentire di rendere efficace e controllabile l'automazione dei processi determinativi, grazie alla rappresentazione esplicita delle regole e dei processi della loro applicazione.[166]

I sistemi basati sulla conoscenza non sono però esperti automatici in grado di sostituirsi al funzionario pubblico, ma piuttosto strumenti che forniscono un aiuto intelligente al richiedente del servizio o al funzionario che lo gestisce. Si tratta di un aiuto che può integrarsi con funzionalità informatiche ulteriori (la ricerca di documenti, la fornitura di moduli, la loro predisposizione automatica) e soprattutto con le attività dell'utente e del funzionario stesso.

Le esperienze delle pubbliche amministrazioni che più ampiamente hanno utilizzato sistemi basati sulla conoscenza (quella australiana, inglese e olandese) indicano come questa tecnologia possa avere un impatto positivo non solo sull'applicazione amministrativa di norme giuridiche, ma anche sulla formulazione delle norme stesse. Al fine di predisporre una base di conoscenza bisogna infatti tradurre i testi giuridici in regole tanto precise da essere automaticamente applicabili. Emergono così le incoerenze e le lacune della normativa da applicare (e le possibilità di riformularne i contenuti con maggiore precisione e chiarezza), e da ciò si possono trarre utili indicazioni per il legislatore, che può essere invitato a porre rimedio ai difetti delle sue statuizioni.

Si è talvolta osservato che l'uso di sistemi basati sulla conoscenza nella pubblica amministrazione può condurre a un'applicazione rigida e iniqua del diritto, sorda alle esigenze del caso concreto, a una forma estrema di legalismo che vincolerebbe l'attuazione del diritto non solo al testo legislativo ma anche alla particolare interpretazione dello stesso che si è deciso di inserire nel sistema informatico. Queste critiche possono trovare risposta non tanto nel richiamarsi a superiori esigenze di certezza ed efficienza (che renderebbero sopportabili rigidità e iniquità), ma piuttosto nel ribadire che i sistemi basati sulla conoscenza non sminuiscono necessariamente né il ruolo del richiedente un servizio pubblico, né quello del funzionario competente. Al contrario, l'uso di tali sistemi può accrescere le sfere di iniziativa e di autonomia informata del cittadino e del pubblico funzionario incaricato di provvedere a un'attività regolata dal diritto. Ciò può avvenire quando l'uso del sistema sia agevole e controllabile, esistano procedure aperte ed efficaci per rivedere la base di conoscenza quando essa conduca a conclusioni inaccettabili, siano rispettati gli ambiti nei quali l'apprezzamento non predeterminato (o predeterminato solo in parte) delle caratteristiche del caso concreto possa dare risultati migliori dell'applicazione di regole predefinite.

Bisogna, invece, progettare modalità di accesso e gestione che consentano di integrare nel modo migliore attività umana ed elaborazione (e memoria) informatica, in modo

che la valorizzazione dell'iniziativa del cittadino e della competenza dell'amministrazione si combini con i vantaggi che l'informatica può dare sotto i profili dell'informazione, dell'efficienza e della certezza. I sistemi basati sulla conoscenza giuridica possono essere utilmente impiegati nella pubblica amministrazione solo grazie all'integrazione delle seguenti attività:

- l'attività dell'esperto informatico-giuridico, che si occupa non solo dell'inserimento, ma anche della correzione e revisione delle regole (nel caso di errori o della necessità di tener conto di nuove informazioni), assicurando che la base di conoscenza non contenga errori di diritto;
- l'attività del cittadino o del funzionario, che inseriscono i dati sui casi concreti, sulla base dell'esame delle circostanze di fatto e dell'interpretazione delle regole da applicare;
- l'attività del sistema, che si occupa della derivazione delle conseguenze deducibili dalle regole in esso comprese e dei dati di fatto a esso forniti o cui abbia accesso.

Per esempio, assumiamo che la concessione di un certo beneficio sia condizionata al possesso di un reddito inferiore a €10.000, e che il sistema chieda alla persona interessata "Hai un reddito inferiore €10.000?". Per rispondere correttamente al quesito si dovrà tener conto sia dei fatti sia dell'interpretazione del requisito reddituale come previsto dalla norma (potrebbe trattarsi del reddito individuale o del reddito familiare, incluse o escluse le prestazioni sociali, ecc.). Quindi, se il reddito individuale di Tizio è €9.000 e il suo reddito familiare è €15.000, la risposta "Sì, ho un reddito inferiore a €10.000" è corretta se "reddito" significa reddito individuale; è invece errata se "reddito" significa reddito familiare. Peraltro, un buon sistema basato sulla conoscenza, in casi come questo, non dovrebbe limitarsi a porre un quesito così ambiguo (almeno per chi non conosca la normativa da applicare) ma dovrebbe fornire indicazioni utili a dare una risposta appropriata.

4.1.4 I limiti dell'applicazione di regole

I sistemi che applicano automaticamente regole preesistenti, rappresentate nella loro base di conoscenza, trovano le maggiori applicazioni, come abbiamo visto, nell'ambito della pubblica amministrazione. Essi non sono in grado di affiancare il giurista nelle attività che formano la parte centrale del suo pensiero, che consiste nell'affrontare casi controversi, nell'identificare e valutare diversi argomenti, nel contribuire alla formazione (o concretizzazione) del diritto.

Tali sistemi potrebbero svolgere per intero il ragionamento giuridico (e in particolare il ragionamento giudiziario) solo se risultasse vera, presa alla lettera, la famosa frase di Charles Montesquieu [1689-1755]:

> I giudici della nazione non sono [...] che la bocca che pronuncia le parole della legge; esseri inanimati che non ne possono moderare né la forza né il rigore.[167]

Il pensiero giuridico si ridurrebbe allora al classico modello del cosiddetto sillogismo giudiziale:

- le soluzioni giuridicamente corrette di ogni caso concreto sarebbero derivabili da un insieme di premesse comprendente norme giuridiche generali (la legge) e asserti fattuali specifici (i fatti del caso concreto), e
- tale derivazione consisterebbe in una deduzione (qualora le premesse normative e fattuali fossero vere/valide saremmo assolutamente sicuri che anche le conclusioni sarebbero vere/valide).

Alla riduzione del ragionamento giuridico al modello sillogistico, come è noto, si oppongono, però, alcune fondamentali caratteristiche del diritto e, conseguentemente, alcuni aspetti essenziali del ragionamento del giurista: le fonti del diritto spesso non forniscono norme sufficienti a disciplinare tutti gli aspetti del caso (vi sono lacune), comprendono norme che si contraddicono (vi sono antinomie) o il cui contenuto rimane indeterminato (vi sono ambiguità e vaghezze), includono casi particolari (precedenti), da cui si possano trarre indicazioni solo mediante analogie o prospettando spiegazioni e generalizzazioni.[168] Inoltre, in certi casi può essere necessario andare al di là dell'applicazione di regole e precedenti, e considerare i valori individuali e sociali perseguiti dal diritto, e i modi in cui massimizzarne la realizzazione, nel rispetto delle aspettative dei cittadini e delle funzioni che competono alle diverse istituzioni giuridiche e politiche.[169]

Secondo alcuni, l'uso di sistemi basati su regole in abito giuridico presupporrebbe l'accettazione del modello sillogistico, e quindi dei suoi presupposti: l'idea che il diritto si ridurrebbe a un insieme di regole e che il ragionamento giuridico consisterebbe nell'applicazione "meccanica" (deduttiva) di tali regole.[170] Questo errore teorico determinerebbe l'inutilità pratica di tali sistemi: non rappresentando fedelmente il diritto e il ragionamento giuridico, essi sarebbero inutili o anzi dannosi, conducendo necessariamente a risultati scorretti o ingannevoli.

La tesi appena esposta non sembra condivisibile. È vero che i sistemi basati sull'applicazione deduttiva del diritto si limitano a cogliere un aspetto limitato del ragionamento giuridico, ma ciò non postula un errore teorico e non ne determina l'inutilità pratica. Infatti, le limitazioni di questi sistemi non ne escludono l'utile impiego, nella piena consapevolezza di tali limitazioni. In un modello simbiotico di interazione uomo-macchina è possibile affidare alla macchina la registrazione delle regole, la loro applicazione logica e il prelievo dei dati fattuali già disponibili in forma elettronica, e affidare invece all'uomo la formulazione delle regole, il controllo sul funzionamento del sistema e l'inserimento e la qualificazione giuridica di nuovi dati fattuali.[171]

4.1.5 La dialettica giuridica: il ragionamento defeasible

Le ricerche degli ultimi anni hanno evidenziato come si possa andare al di là dei limitati risultati conseguibili nell'ambito del modello deduttivo, operando in due direzioni complementari:

- la costruzione di precisi modelli dei ragionamenti giuridici non-deduttivi, la creazione, cioè, di una nuova logica giuridica formale, più estesa dei tradizionali modelli deduttivi;[172]
- lo sviluppo, sulla base di tali modelli, di software che possano agevolare la ricerca e l'elaborazione di informazioni giuridiche.

Lo sviluppo delle logiche defeasible si basa sull'idea che sia possibile e utile predisporre strumenti logici e informatici capaci di cogliere la dialettica del ragionamento giuridico, superando i limiti del ragionamento sillogistico. Infatti, quando la soluzione a un problema giuridico controverso è presentata quale risultato di una deduzione da regole giuridiche date, la spiegazione risulta monca: non indica perché i testi sono stati interpretati in un certo modo, perché altri testi non sono stati presi in considerazione, quali principi sono stati attuati e quali tralasciati, quali eccezioni si sono accolte o respinte, quali obiettivi sono stati perseguiti e quali sono rimasti insoddisfatti. Chi intenda giustificare una decisione giudiziale o una soluzione dottrinale solo mediante un sillogismo, è costretto a tacitare, a rimuovere (in quanto inesprimibile nel mezzo logico scelto), proprio il nucleo del processo raziocinativo-argomentativo che ha condotto a quel risultato, nucleo che rimane inarticolato, se non inconsapevole. Di conseguenza, la giustificazione stessa risulta monca e insufficiente.[173]

Come abbiamo visto (Sezione 2.1.2), un'inferenza defeasible è un argomento che si impone alla nostra ragione, ma solo in modo provvisorio, cioè solo a condizione che non emergano eccezioni, contro-esempi, argomenti contrari di importanza preminente. Pertanto, il fatto che si accolgano le premesse dell'argomento che conduce a una certa conclusione può essere insufficiente a giustificare tale conclusione: bisogna indicare quali controargomenti siano stati considerati, e perché l'argomento prescelto abbia prevalso su di essi. Per illustrare la dialettica di argomenti contrapposti dobbiamo quindi considerare non solo la struttura degli argomenti, ma anche i modi in cui essi possono essere attaccati da contro-argomenti.

La struttura interna degli argomenti defeasible. La struttura fondamentale degli argomenti basati su regole defeasible è rappresenta, in forma generale, nei grafi proposti dal Stephen Toulmin,[174] nei quali si distinguono i dati di partenza dell'argomento (*data*) e la regola generale (*warrant*), che consente di passare dai dati alla conclusione che si vuole sostenere (*claim*). Questa struttura è riproposta nel modello qui presentato, che si basa sull'assunto che gli argomenti defeasibile possano essere ricondotti alla

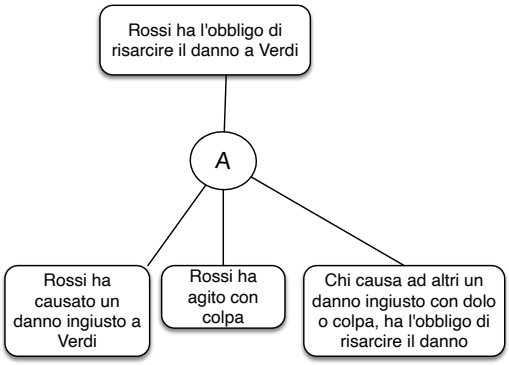

Figura 4.1: *Argomento*

forma del cosiddetto *modus-ponens defeasible*. Sia data una combinazione di fatti specifici A_c (la cosiddetta fattispecie concreta), e una regola generale defeasible "SE A_a allora (normalmente o presuntivamente) B_a", dove A_a (la cosiddetta fattispecie astratta) è un'espressione generale che sussume A_g. Ne possiamo inferire B_c (l'effetto giuridico concreto corrispondente all'effetto giuridico astratto B_a.

Per esempio, nell'argomento *A* della Figura 4.1, i dati sono costituiti dal fatto che Rossi ha causato un danno ingiusto a Verdi agendo con colpa, la conclusione è rappresentata dall'obbligo risarcitorio a carico di Rossi, e il nesso tra fatti e conclusione è fornito dalla regola generale che ogni fatto doloso o colposo che causi ad altri un danno ingiusto obbliga l'autore al risarcimento.

Possiamo dare a questi argomenti una forma logica. Per esempio, l'argomento *A* potrebbe essere riformulato nel seguente:

1. x deve risarcire y SE
   (a) x ha causato a y un danno ingiusto E
      i. x ha agito con dolo O
      ii. x ha agito con colpa
2. *Rossi* ha causato a *Verdi* un danno ingiusto
3. *Rossi* ha agito con colpa
   PERTANTO
4. *Rossi* deve risarcire *Verdi*

Sulla base di questa rappresentazione possiamo precisare il rapporto tra regola generale defeasible e fatti particolari: la regola generale contiene delle variabili (x e y) e la sussunzione della fattispecie concreta nella fattispecie astratta avviene sostituendo, in modo uniforme, le variabili con i nomi di persone o cose particolari (*Rossi* e *Verdi*).[175]

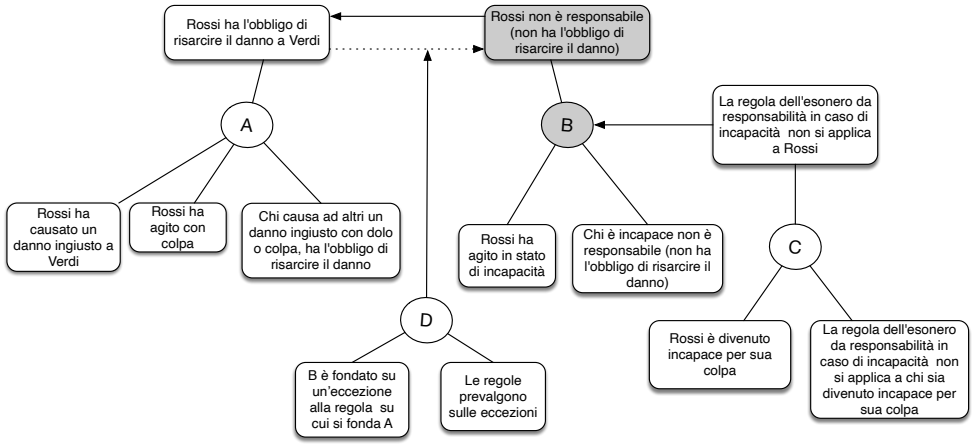

Figura 4.2: *Argomenti e controargomenti*

Gli attacchi tra argomenti. Possiamo distinguere due modi in cui un argomento defeasible α può essere attaccato. Il primo è detto *rebutting*,[176] termine che forse possiamo tradurre con "contraddire", e consiste nell'opporre ad α un contro-argomento β che contraddice la conclusione di α: la conclusione di β nega, o comunque è incompatibile con la conclusione di α. Il secondo è detto *undercutting*,[177] che forse possiamo tradurre con *recidere*, e consiste nell'opporre ad α un contro-argomento β che nega la forza argomentativa di α. Più esattamente, β afferma che le premesse di *A* sono inidonee a fondarne la conclusione (in particolare, β può contestare l'applicabilità della regola su cui si fonda, in α, il passaggio da premesse a conclusioni).

Per esempio, l'argomento *A* della Figura 4.2 è contraddetto (*rebutted*) dall'argomento *B* secondo il quale Rossi non è responsabile essendo incapace di intendere e di volere nel momento dell'incidente. L'argomento *B* della Figura 4.2 (Rossi non è responsabile, in forza della sua incapacità) viene invece reciso (*undercut*) dall'argomento *C* secondo il quale l'incapacità di Rossi non è rilevante (non lo scusa) poiché Rossi provocò colpevolmente il suo stato di incapacità, per esempio, assumendo bevande alcoliche. Il fatto che Rossi abbia bevuto troppo non è un'autonoma ragione per la quale egli debba essere ritenuto responsabile (si è responsabili per aver causato ad altri un danno ingiusto, non per la propria ubriachezza). Si tratta invece di una ragione per disapplicare la regola secondo la quale chi è incapace non è responsabile o, detto in altro modo, per impedire (recidere) il passaggio dall'incapacità alla non-colpevolezza.

Come valutare argomenti in conflitto. L'accoglimento delle premesse (i dati e la regola) di un argomento defeasible non è sufficiente a garantire l'accettabilità della conclusione di tale argomento. A tal fine, dobbiamo altresì considerare se esista-

no contro-argomenti che contraddicano o recidano il nostro argomento. Solo se tutti i contro-argomenti possono essere superati alla luce di considerazioni ulteriori, la conclusione dell'argomento risulta giustificata. Da ciò discendono due aspetti importanti dell'interazione dialettica tra argomenti contrapposti.

Il primo aspetto consiste nella possibilità di ristabilire (*reinstate*) un argomento, attaccando gli argomenti che potrebbero sconfiggerlo. Un argomento può essere difeso non solo producendo argomenti ulteriori che vanno nella stessa direzione (che ne ribadiscono la conclusione) ma anche, e soprattutto, cercando di demolire i contro-argomenti che a esso si oppongono (in modo da ristabilire l'argomento stesso). Per esempio, l'argomento *A* appena sopra (Rossi è responsabile avendo danneggiato colpevolmente Verdi) può essere difeso dall'attacco del contro-argomento *B* (Rossi non è responsabile essendosi trovato in stato di incapacità), attaccando *B* con il nuovo contro-contro-argomento *C* (l'incapacità di Rossi è irrilevante, essendosela procurata Rossi stesso, ingerendo sostanze alcoliche). *C*, nell'invalidare *B*, fa in modo che A riacquisti la propria forza (*C* ristabilisce *A*), come illustrato nella Figura 4.2.[178]

Il secondo aspetto, invece, consiste nella possibilità di sostenere un argomento *A*, contraddetto (*rebutted*) da un contro-argomento *B*, adducendo ulteriori argomenti che indichino perché *A* debba prevalere su *B*. Nell'esempio della Figura 4.2, *B* prevale su *A* trattandosi di un'eccezione alla regola stabilita da *A*. Tuttavia, l'attacco di *B* non inficia lo status IN di *A*, poiché *B* è stato eliminato da *C*.

Gli argomenti, alla luce delle relazioni con gli altri argomenti presenti nel medesimo contesto argomentativo possono essere suddivisi in tre classi, sulla base in una valutazione che tiene in considerazione tutti e soli gli argomenti presenti in quel contesto:

- argomenti IN, che debbono essere accettati;
- argomenti OUT, che debbono essere respinti;
- argomenti UND, sui quali permane l'incertezza.

Si consideri che questa valutazione è sempre defeasible, nel senso che lo status di tali argomenti potrebbe cambiare: ogni argomento potrebbe assumere un diverso status se altri argomenti fossero introdotti nel contesto (un argomento IN potrebbe diventare OUT o UND; un argomento OUT, IN or UND; un argomento UND, IN o OUT).

Nelle Figure 4.2 e 4.3 lo stato dialettico di un argomento è indicato dal colore dello stesso: gli argomenti IN (validi, alla luce dell'informazione riportata nel diagramma) sono in chiaro, gli argomenti OUT (invalidi, alla luce dell'informazione disponibile) sono ombreggiati. Una simile notazione è stata adottata per gli attacchi tra argomenti, denotati da frecce: la freccia continua indica un attacco IN (efficace), la freccia tratteggiata indica un attacco OUT (inefficace).

La Figura 4.3 illustra come sia possibile sostenere un argomento *A*, contraddetto (*rebutted*) da un contro-argomento *B*, adducendo ulteriori argomenti che indicano perché *A* debba prevalere su *B*. Assumiamo che nella causa di divorzio tra Rossi e Verdi, Rossi

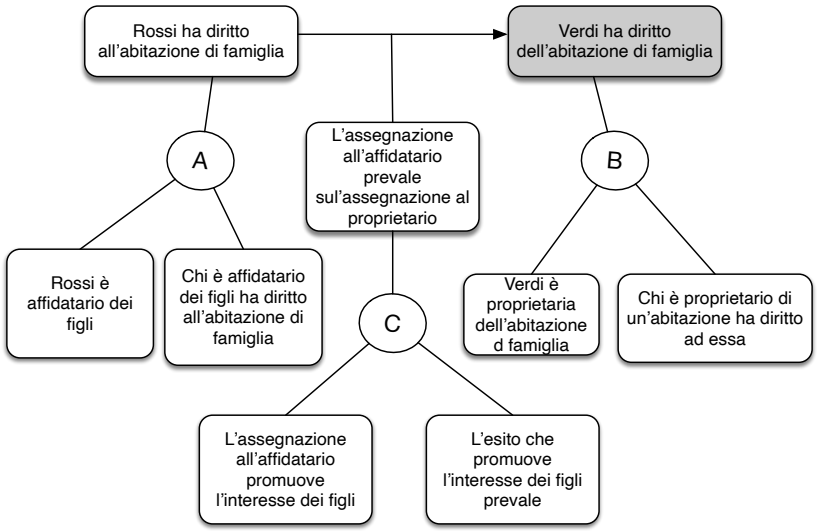

Figura 4.3: *Argomenti e controargomenti*

affermi (argomento *A*) il proprio diritto all'uso dell'abitazione della famiglia, essendo affidatario dei figli, e che Verdi (argomento *B*) opponga il proprio diritto a possedere tale abitazione, in quanto proprietaria della stessa. A questo punto, il contrasto tra i due argomenti può forse essere risolto affermando (argomento *C*) la prevalenza dell'argomento di Rossi, poiché tale argomento corrisponde al preminente interesse dei figli. L'esito del conflitto tra *A* e *B* è indeterminato quando si considerino solo *A* e *B*. Esso acquista determinatezza alla luce della valutazione comparativa motivata dall'ulteriore argomento *C*.[179]

Grazie a questa valutazione comparativa possiamo escludere che *B* riesca ad attaccare con efficacia *A*: l'attacco di *A* contro *B* resta IN, mentre l'attacco di *B* contro *A* è OUT.

È possibile fornire una definizione logica per determinare con precisione lo stato di ciascun elemento (sia esso un argomento o un attacco) in un grafo di argomenti:

- un elemento *A* è IN, se nessun argomento IN propone un attacco IN contro *A*;
- un elemento *A* è OUT, se un argomento IN propone un attacco IN contro *A*;
- un elemento è UND (il suo stato è indeterminato), se non è né IN né OUT.

Il lettore può verificare che seguendo queste regole è possibile assegnare agli argomenti le valutazioni rappresentate nelle Figure 4.2 e 4.3. Per esempio, consideriamo la Figura 4.2. L'argomento *C* è IN non avendo alcun attacco, così come è IN l'attacco di *C* contro *B*. Di conseguenza *B* è OUT e *A* (non essendo attaccata da argomenti IN) è IN.

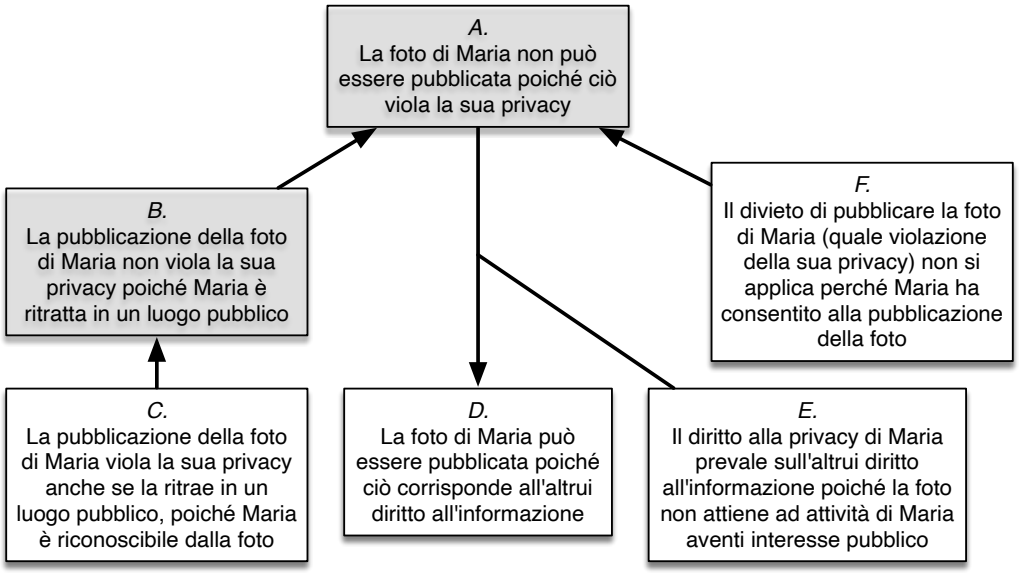

Figura 4.4: *Dialettica di argomenti: la privacy*

Una dialettica argomentativa a più livelli è illustrata nella Figura 4.4, che fornisce una rappresentazione meno analitica dei singoli argomenti, senza esplicitarne la struttura interna (la divisione tra premesse e conclusione). Ripercorriamo brevemente questo esempio. S'immagini che la foto di una persona, chiamiamola Maria, sia stata pubblicata in un giornale e che Maria chieda lumi al suo avvocato sui possibili rimedi giuridici (l'esempio non intende essere giuridicamente accurato, ma si limita a illustrare, nel modo più semplice, la logica qui proposta).

Il primo argomento considerato dal legale è il seguente:

A : La foto di Maria non può essere pubblicata, poiché ciò viola la sua privacy.

Da *A*, considerato isolatamente, possiamo trarre la conclusione che la foto di Maria non poteva essere resa pubblica (la sua pubblicazione viola i diritti di Maria). Tuttavia, l'avvocato deve considerare anche i possibili contro-argomenti, come *B*:

B : La pubblicazione della foto di Maria non viola la sua privacy, poiché la ritrae in un luogo pubblico.

Come abbiamo osservato, per determinare se un argomento giuridico sia giustificato, non basta guardare a quell'argomento isolatamente. Al contrario, bisogna collocare l'argomento nel contesto di tutti gli argomenti rilevanti, cioè nel *contesto argomentativo*

(*argumentation framework*) del caso: alla luce di nuovi argomenti, argomenti che apparivano giustificati (IN) possono cessare di esserlo, mentre argomenti che apparivano sopraffatti da argomenti in contrario (e quindi erano OUT), possono riacquistare forza.

Per esempio, se A fosse stato l'unico argomento rilevante proposto nel caso di Maria, e nessuna obiezione (contro-argomento) si fosse potuta sollevare contro di esso, avremmo dovuto accettare la sua conclusione (A sarebbe stato IN). Invece nella cornice argomentativa che unisce A e il contro-argomento B, A non è più giustificato: B prevale su di A e quindi A è OUT (Maria non riesce a dimostrare che la foto non poteva essere pubblicata).

Tuttavia, l'avvocato di Maria, riflettendo ancora sulla causa, trova un argomento contro B:

C : La foto di Maria viola la sua privacy, anche se la ritrae in un luogo pubblico, poiché Maria è riconoscibile nella foto.

L'argomento C prevale contro B (essendo più specifico) e quindi priva B di rilevanza nel caso in esame: in base a C, B è OUT, e quindi che il suo attacco contro A diventa inefficace. Di conseguenza, nel contesto argomentativo costituito degli argomenti A, B e C (dove A è attaccato da B, ma C attacca B e prevale su di esso) risulta che A è nuovamente IN (Maria dimostra che la sua foto non poteva essere pubblicata).

Vi è però un altro possibile argomento a favore della pubblicazione della foto: essa realizza l'interesse del pubblico (assumiamo che Maria rivesta un ruolo pubblico).

D : La foto di Maria può essere pubblicata poiché ciò realizza l'interesse del pubblico.

L'argomento D porta un nuovo attacco contro A. Nel contesto argomentativo A, B, C, e D, risulta che A è nuovamente OUT (Maria non riesce a dimostrare che la sua foto non poteva essere pubblicata). A questo punto, l'avvocato perplesso ricorre a un meta-ragionamento. Egli sviluppa cioè un nuovo argomento (E) che adduce ragioni a favore della prevalenza di uno degli argomenti in conflitto:

E : Il diritto alla privacy prevale sull'interesse del pubblico alla conoscenza, poiché la foto di Maria non attiene ad attività aventi interesse pubblico.

Grazie all'aiuto fornitogli dal nuovo argomento E, A appare ora superiore a D. Nel grafo ciò risulta dal fatto che la freccia che denota l'attacco efficace vada da A verso B, piuttosto che da B verso A. Di conseguenza, è ora E ad essere OUT, in quanto attaccato da A. Nel contesto argomentativo costituito dagli argomenti A, B, C, D, E —dove A e D si attaccano reciprocamente, e E afferma che A è superiore a D— A risulta nuovamente IN (Maria dimostra che la sua foto non poteva essere pubblicata).

Il ragionamento dell'avvocato, purtroppo, non è ancora finito. Nel colloquio con la cliente emerge che questa aveva consentito alla pubblicazione della propria foto. Abbiamo così il nuovo argomento F che attacca A:

F : Il divieto di pubblicare la foto di Maria (in quanto attinente alla sua privacy) non si applica poiché Maria ha consentito alla pubblicazione della foto.

L'attacco di F fa sì che A debba essere respinto: nel contesto argomentativo rappresentato dagli argomenti A, B, C, D, E, e F, risulta che quindi A è OUT (quindi Maria non riesce a dimostrare che la sua foto non poteva essere pubblicata). D, privato dell'attacco di A, è tornato ad essere *IN*.

L'argomentazione potrebbe continuare indefinitamente (per esempio, si potrebbe sostenere che il consenso di Maria era stato revocato, cosa che rende inapplicabile la regola del consenso), ma quanto detto finora può dare sufficiente supporto alla tesi che il ragionamento giuridico consiste nella dialettica tra argomenti e contro-argomenti, e che tale dialettica fa sì che il ragionamento giuridico sia defeasible: le conclusioni che appaiono giustificate secondo certi argomenti possono essere inficiate da argomenti ulteriori.[180]

Le logiche degli argomenti. Questo pur sommario esempio dovrebbe essere sufficiente a illustrare l'idea fondamentale che caratterizza i tentativi di costruire modelli formali della dialettica giuridica. Si tratta di affrontare situazioni nelle quali vi sono numerosi argomenti in gioco. Alcuni stanno combattendo per la supremazia, altri danno sostegno ad alcuni degli argomenti in conflitto (per esempio, fornendo ragioni per le quali essi dovrebbero prevalere sui propri oppositori), altri negano l'applicabilità di altri argomenti, e così via. Il compito di una logica dialettica non è solo dirci quali argomenti siano costruibili utilizzando premesse date, ma anche di determinare quali argomenti emergano vincitori (giustificati) dallo scontro con i loro contro-argomenti, quali siano sconfitti, e quali siano difendibili (non sconfitti, ma neppure vincitori).

Logiche degli argomenti, come quella appena presentata, sono state usate per realizzare sistemi informatici che, anziché limitarsi a fornire una risposta univoca ai quesiti loro proposti, elaborino giustificazioni per la soluzione di punti controversi, suggeriscano argomenti possibili, valutino lo stato degli argomenti alla luce dell'architettura argomentativa complessiva risultante dalle informazioni fornite al sistema (gli argomenti, i contro-argomenti e i meta-argomenti costruibili con tali informazioni e le loro relazioni).[181] La realizzazione di tali sistemi richiede nuovi linguaggi per la rappresentazione della conoscenza, sufficientemente espressivi da cogliere le strutture fondamentali della conoscenza giuridica (le regole, i diritti, i casi, i principi, i valori, ecc.) e nuovi metodi di inferenza, che riproducano i passi tipici del ragionamento giuridico (l'applicazione di regole, il riferimento ai precedenti, il ragionamento teleologico, ecc.).

Uno sviluppo ulteriore consiste nella realizzazione di sistemi tesi ad agevolare le discussioni giuridiche (cooperative o conflittuali) indicando alle parti, in ogni momento della loro interazione, quale sia lo stato di ogni argomento, e quali nuovi argomenti possano essere rilevanti per l'oggetto della discussione. Inoltre, tali sistemi mirano a organizzare le informazioni fornite nel corso del dialogo in un'architettura di argomenti e contro-argomenti (ragioni e contro-ragioni), dove ogni argomento sia collegato agli

argomenti che sostiene o attacca. Chi entra nella discussione può quindi accedere più facilmente al punto del dibattito che lo interessa maggiormente, o rispetto al quale egli voglia fornire un contributo. Tra le applicazioni, ricordo i sistemi intesi a promuovere il dialogo democratico (la discussione di temi politici e amministrativi), o a facilitare la soluzione di controversie mediante conciliazione. Questi sistemi si scontrano però con la difficoltà di tradurre in strutture logiche uniformi (necessarie affinché il sistema possa organizzare le argomentazioni fornite dalle parti), le molteplici e complesse forme linguistiche in cui si svolge il dibattito giuridico. Pertanto, ancor oggi sono disponibili solo applicazioni prototipali.[182]

4.1.6 Il ragionamento basato sui casi

Quale esempio di sistema per il ragionamento che opera sulla base di rappresentazioni di casi, anziché di regole, possiamo ricordare il sistema HYPO, applicato in particolare nell'ambito della violazione dei segreti industriali.[183] La base di conoscenza del sistema è costituita da un insieme di precedenti. Ciascun precedente riguarda un caso nel quale il convenuto abbia fatto uso di informazioni sulle attività dell'attore.

La rappresentazione dei casi. Ogni caso è descritto (annotato) con i seguenti dati:

- il suo esito, che può essere a favore dell'attore *a*, cioè l'accertamento di una violazione, o invece a favore della difesa del convenuto *c*, cioè l'accertamento che non vi è stata violazione;
- un insieme di fattori, cioè di circostanze a favore o contro l'uno o l'altro esito.

Presentiamo i fattori che sono rilevanti per la decisione dei casi che considereremo, prima quelli a favore dell'attore e poi quelli a favore del convenuto:

- F4(a) il convenuto si era impegnato a non diffondere/utilizzare l'informazione
- F6(a) l'attore aveva adottato specifiche misure di sicurezza intese a mantenere segreta l'informazione
- F7(a) il convenuto ha assunto un ex-dipendente dell'attore
- F15(a) il prodotto cui afferiva l'informazione era l'unico di quel tipo presente sul mercato
- F18(a) il prodotto sviluppato dal convenuto è identico a quello dell'attore
- F21 (a) il convenuto sapeva che l'informazione era confidenziale
- F1(c) durante le negoziazioni l'informazione era stata comunicata al convenuto
- F10(c) l'informazione era stata comunicata a terzi

- F16(c) l'informazione era ottenibile mediante l'esame del prodotto (*reverse engineering*)

I fattori debbono essere individuati con un'analisi giuridica dell'ambito considerato, possibilmente adiuvata da strumenti automatici.

Consideriamo due precedenti, il caso Rossi e il caso Verdi:

- nel caso Rossi, deciso a favore del convenuto, il convenuto aveva assunto un ex dipendente dell'attore (F7(a)) l'informazione in questione (che l'attore afferma essere un segreto industriale) era stata comunicata a terzi (F10(c)) ed era ottenibile mediante *reverse engineering* (F16(c));
- nel caso Verdi, deciso a favore dell'attore, il convenuto aveva assunto un ex dipendente dell'attore (F7(a)), il convenuto sapeva che l'informazione era confidenziale (F21(a)), e l'informazione era ottenibile mediante reverse engineering (F16(c)).

Infine nel nuovo caso, il caso Gialli, l'attore ha adottato misure di sicurezza (F6 (a)), il prodotto cui afferisce l'informazione è l'unico di quel tipo presente sul mercato F15 (a), il convenuto sapeva che l'informazione era confidenziale, (F21 (a)) durante le negoziazioni l'informazione era stata comunicata al convenuto (F1(c)), l'informazione era ottenibile mediante reverse engineering (F16 (c)).

I casi sono rappresentati dai fattori che si applicano a essi, uniti alle relative decisioni (A, a favore dell'attore o C, a favore del convenuto). Il nuovo caso, il caso Gialli, è anch'esso rappresentato mediante i fattori ad esso applicabili, ma la decisione, ancora sconosciuta, è rappresentata da un punto di domanda.

- Caso Rossi. Fattori: F7(a), F10(c), F16(c). Decisione: C
- Caso Verdi. Fattori: F7(a), F21(a), F16(c). Decisione: A
- Caso Gialli. Fattori: F6(a), F15(a), F21(a), F1(c), F16(c). Decisione: ?

La Figura 4.5 indica i rapporti tra i tre casi, specificando quali fattori essi condividano.

Analogie e distinzioni. Sulla base della rappresentazione dei casi presentata nel paragrafo precedente, sono possibili alcune inferenze analogiche, fornite automaticamente dai sistemi Hypo e Cato.[184]

Per esempio, l'attore può affermare che il caso Gialli dovrebbe essere deciso a suo favore (A), poiché esso è simile al caso Verdi (che ha decisione A): in entrambi i casi l'informazione era ottenibile mediante l'esame del prodotto (*reverse engineering*, F1(c)), e ciò giustifica una decisione a favore dell'attore non solo nel caso Verdi, ma anche nel caso Gialli.

Il convenuto può contestare questa analogia in due modi. Innanzitutto, egli può effettuare una distinzione, cioè affermare che il caso Gialli presenta due importanti differenze

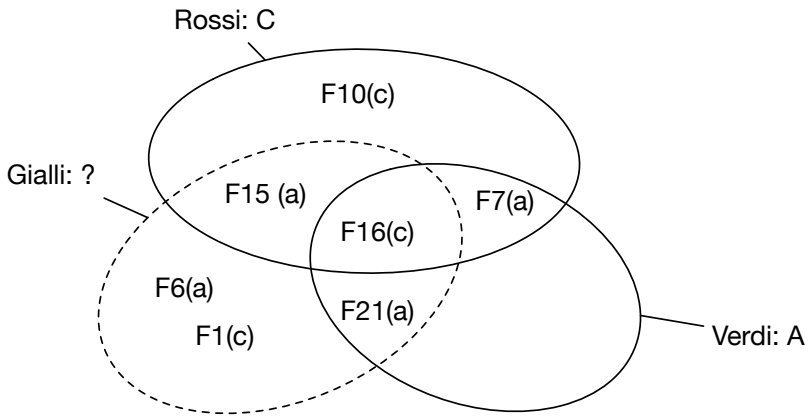

Figura 4.5: *Casi e fattori*

rispetto al caso Verdi, che favoriscono una decisione a favore del convenuto, invece che a favore dell'attore.

- nel caso Gialli è presente un elemento a favore del convenuto che non era presente nel caso Verdi: durante le negoziazioni l'informazione era stata comunicata al convenuto (F1(c))
- nel caso Gialli manca un elemento a favore dell'attore, elemento che era presente nel caso Verdi: il convenuto aveva assunto un ex dipendente dell'attore (F7(a)).

Oltre ad effettuare una distinzione, il convenuto può citare a suo favore il caso Verdi, che condivide con il caso Gialli l'elemento consistente nel fatto che l'informazione era ottenibile mediante *reverse engineering* (F16(c)).

Secondo il modello del ragionamento per fattori, oltre a usare i casi per trarne analogie, possiamo trarne indicazioni vincolanti (se si vogliono rispettare i precedenti). Consideriamo un ulteriore nuovo caso, il caso Neri, che include tutti i fattori del caso Verdi, e in più contiene F1, un ulteriore fattore a favore del convenuto (Figura 4.6):

- Caso Neri. Fattori: F7(a), F1(c), F10(c), F16(c),. Decisione: C

Sulla base di questa analisi il caso Neri deve a maggior ragione (*a fortiori*) essere deciso a fattore del convenuto. Infatti, in base al caso Rossi possiamo dire che i fattori F10(c) e F10(c) prevalgono sul fattore F7(a), giustificando una decisione a favore del convenuto. Di conseguenza, la combinazione di F1(c), F10(c) e F16(c) (nella quale F10(c) e F16(c) sono rafforzati dall'aggiunta di F1(c)) a maggior ragione prevarrà

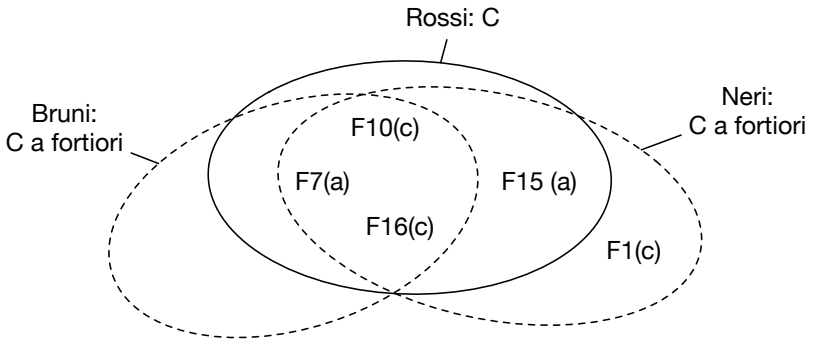

Figura 4.6: *Casi e fattori: ragionamento* a fortiori

su F7(a), imponendo la medesima conclusione. Analogo ragionamento porterebbe alla stessa conclusione nel caso Bruni, che contiene i fattori del caso Rossi, ad eccezione del fattore F7(a), a favore dell'attore.

Il modello del ragionamento per fattori è stato sviluppato in numerosi contributi, nei quali si sono identificati i rapporti logici tra fattori che si collocano a diversi livelli di astrazione,[185] la connessione tra ragionamento basato sui casi e modelli per l'argomentazione,[186] la definizione logica dei casi in cui un precedente impone una certa decisione,[187] fattori scalabili (suscettibili di essere presenti in diverse quantità).[188]

4.2 L'apprendimento automatico nel diritto

Fino a tempi molto recenti gran parte delle ricerche di intelligenza artificiale in ambito giuridico erano basate sull'idea che la realizzazione di applicazioni intelligenti giuridiche dovesse fondarsi su rappresentazioni formali della conoscenza, come negli indirizzi di ricerca illustrati nelle pagine precedenti.[189] L'enorme successo dei metodi per l'apprendimento automatico —in particolare di quelli subsimbolici, come le reti neurali— in altri settori ha condotto gli studiosi di informatica giuridica e gli sviluppatori di tecnologie giuridiche a rivedere questa opinione. Oggi sono sempre più (sono forse ormai maggioritari) i contributi che invece utilizzano metodi di apprendimento automatico, applicati alla sempre più grande quantità di materiali giuridici disponibili in formato digitale.

Nelle pagine seguenti si presenta dapprima Claudette, un sistema basato sull'apprendimento automatico, per poi passare ad esaminare il tema dell'uso di tecnologie per l'apprendimento automatico nel campo della giustizia predittiva, e infine proporre al-

cune considerazioni sulle prospettive di tali tecnologie in ambito giuridico, e sulla loro integrazione con tecnologie per la rappresentazione della conoscenza.

4.2.1 Claudette, un sistema per l'analisi di documenti giuridici

Il sistema Claudette si propone di applicare metodi di apprendimento automatico per identificare e classificare le clausole (potenzialmente) abusive e illegali, contenute in contratti e informative privacy.

La nascita di Claudette è stata motivata dalla constatazione della divergenza tra discipline normative e pratiche commerciali online, e quindi dall'esigenza di elaborare nuove forme di tutela giuridica, anche grazie all'uso consapevole dell'IA. Esiste, infatti, un ampio corpus di norme europee e nazionali in materia di protezione dei consumatori e tutela dei dati personali, e vi sono autorità pubbliche preposte al controllo e all'applicazione della legge, cui si uniscono gli organismi associativi e le iniziative della società civile. Ciononostante, i contratti delle piattaforme online spesso contengono clausole che appaiono, spesso in modo evidente, non conformi al diritto.

Nelle pagine seguenti, si presenterà l'applicazione del sistema Claudette ai contratti per l'utilizzo di servizi online (termini di servizio - *terms of service*), illustrando l'analisi giuridica svolta, la metodologia adottata per l'addestramento del sistema, e i risultati ottenuti. La metodologia di Claudette è stata peraltro applicata anche alle informative privacy.[190]

Claudette per l'analisi dei termini di servizio. Il principale riferimento normativo per l'applicazione di Claudette ai termini di servizio è dato dalla Direttiva 93/13/CEE sui contratti stipulati con i consumatori, che vieta l'utilizzo di «clausole contrattuali abusive» nei contratti unilateralmente redatti da produttori e intermediari. L'Art. 3 della Direttiva 93/13 stabilisce che una clausola è abusiva se: (*a*) non è stata individualmente negoziata e (*b*) anche se adottata in buona fede, determina uno squilibrio significativo nei diritti e negli obblighi delle parti, a scapito del consumatore. Tale definizione è integrata da un allegato contenente un elenco indicativo e non esaustivo di clausole abusive, e da numerose sentenze della Corte di Giustizia.

Per poter automatizzare l'identificazione e la classificazione delle clausole abusive sono stati raccolti più di 100 contratti per la fornitura online di beni e servizi in lingua inglese, predisposti dai più importanti operatori online per numero di utenti, rilevanza geografica e durata dall'istituzione del servizio (inclusi ad esempio Amazon, eBay, Dropbox, Facebook, Google, Microsoft, Netflix, Spotify, Twitter, Yahoo, YouTube, ecc.).

Sono state identificate otto categorie di clausole da valutare (quelle che più frequentemente presentano profili di vessatorietà), e per ciascuna di esse è stato definita un'abbreviazione: giurisdizione (j), legge applicabile (law), limitazione di responsabilità (ltd),

modifiche unilaterali (ch), risoluzione unilaterale (ter), arbitrato (a), consenso implicito mediante utilizzo del servizio (use), e rimozione di contenuti (cr). Per ciascuna delle otto categorie sono stati definiti criteri per determinarne la qualità giuridica, cioè se esse siano da considerarsi legittime, possibilmente vessatorie, chiaramente vessatorie.

I contratti sono stati analizzati da esperti giuridici, che hanno identificato, in ciascuno di essi, le clausole attinenti alle otto categorie, valutandone la qualità giuridica. Ogni clausola siffatta è stata così etichettata mediante un'espressione che ne indica la categoria e la qualità giuridica (1, legittima; 2, potenzialmente vessatoria; 3, chiaramente vessatoria). Per esempio, l'etichetta j3 indica che la clausola verte sulla giurisdizione (j) ed è chiaramente vessatoria (3). Usando la convenzione propria del meta-linguaggio di marcatura XML, l'etichetta tra parentesi uncinate precede la clausola, e la stessa etichetta, preceduta dalla barra "/" la chiude.

Per esempio, le clausole sulla giurisdizione sono state ritenute chiaramente legittime quando garantiscono ai consumatori il diritto di agire in giudizio nel luogo in cui gli stessi risiedono. Al contrario, le clausole che stabiliscono la giurisdizione in un luogo diverso dalla residenza del consumatore sono state considerate chiaramente abusive. Per esempio, la clausola seguente, tratta dai termini di servizio offerte da Shazam, è etichettata in quest'ultimo modo, poiché conferisce giurisdizione esclusiva ai giuridici di Inghilterra e Galles:

> <j1>You and we agree to submit to the personal and exclusive jurisdiction of the courts of England and Wales</j1>

Le clausole sulla legge applicabile sono state considerate chiaramente legittime quando rinviano alla legge del luogo di residenza del consumatore. In tutti gli altri casi, sono state classificate come potenzialmente abusive. È questo il caso della seguente clausola tratta dai termini di servizio di Facebook, che sancisce l'applicabilità esclusiva delle leggi della California:

> <j1>The laws of the State of California will govern this Statement, as well as any claim that might arise between you and us, without regard to conflict of law provisions</j1>

Le clausole di limitazione della responsabilità si sono considerate chiaramente legittime quando sanciscono in modo esplicito la piena responsabilità del prestatore. Le clausole che riducono, limitano o escludono la responsabilità sono state invece marcate come potenzialmente vessatorie se relative ad ampie categorie di danni o cause di danni. Si veda la seguente clausola tratta dai termini di servizio di Instagram, che esclude la responsabilità per la modifica, la sospensione o l'interruzione del sevizio, così come per la perdita di contenuto.

> <j1>Instagram will not be liable to you for any modification, suspension, or discontinuation of the Services, or the loss of any Content,</j2>

Invece, le clausole finalizzate a ridurre, limitare o escludere la responsabilità nelle ipotesi di lesioni fisiche, danni intenzionali e negligenza grave, sono state considerate chiaramente abusive. In un modo simile si sono etichettate le clausole rientranti nelle rimanenti categorie (risoluzione unilaterale (ter), arbitrato (a), consenso implicito mediante utilizzo del servizio (use), e rimozione di contenuti (cr)).

La metodologia per l'addestramento del sistema. Il sistema Claudette è stato addestrato a riconoscere le clausole (potenzialmente) abusive mediante metodi di apprendimento automatico supervisionato.

Le clausole marcate nell'insieme di addestramento indicano al sistema come a certe forme testuali corrispondano certe classificazioni giuridiche, cioè la determinazione della categoria cui appartiene la clausola e la valutazione della sua vessatorietà.

Come si è visto nella Sezione 2.2, gli algoritmi per l'apprendimento "insegnano" al sistema a effettuare determinazioni e valutazioni analoghe a quelle specificate nell'insieme di addestramento. Tali algoritmi generano un "modello" che fornisce al sistema criteri e meccanismi per la classificazione delle clausole, modello costruito sulla base di esempi nell'insieme di addestramento. Potremmo anche dire che il sistema, sulla base dell'addestramento, genera un proprio concetto di vessatorietà, che poi applica alla classificazione di nuove clausole. Tale concetto è costruito in modo che, nell'applicarlo, il sistema giunga alle stesse conclusioni cui sono giunti i giuristi che hanno predisposto l'insieme di addestramento, rispetto ai casi contenuti in tale insieme. Il concetto implicito nel sistema è però astratto (implica una generalizzazione). Infatti il sistema stesso classifica secondo quel concetto (come vessatorie o non-vessatorie) anche frasi nuove, non comprese nell'insieme di addestramento. Anche rispetto a tali frasi le conclusioni raggiunte dal sistema dovrebbero avvicinarsi a quelle cui giungerebbe un giurista esperto. Applicando il modello così costruito il sistema si propone di realizzare due obiettivi:

- categorizzare le clausole di nuovi contratti, cioè determinarne l'oggetto, per esempio, se tali clausole vertano su giurisdizione, arbitrato, responsabilità, ecc.;
- valutare le clausole, cioè determinare se esse siano vessatorie.

Per realizzare questi obiettivi sono necessarie due fasi. La prima consiste nella trasformazione delle clausole in strutture di dati che possano essere utilizzate per l'addestramento del sistema. A tal fine è necessario innanzitutto procedere all'analisi lessicale, che consiste nel suddividere il testo negli elementi rilevanti (tipicamente parole e segni di punteggiatura). La fase successiva è quella dell'analisi sintattica, che individua le

	il	fornitore	è	responsabile	per	i	danni	non	risarcisce	l'	utente	sempre
a	1	1	1	1	1	1	1	0	0	0	0	0
b	1	1	0	0	0	1	1	1	1	0	0	0
c	0	0	1	0	0	0	0	0	0	1	1	1

Figura 4.7: *Rappresentazione di frasi mediante "borse di parole"*

strutture sintattiche in cui tali elementi sono combinati. Per esempio, la frase "il provider non è responsabile per i danni", può essere suddivisa negli elementi seguenti: [il], [provider], [non], [è], [responsabile], [per], [i], [danni]. Nella fase dell'analisi sintattica si può stabilire come quegli elementi si combinano nel formare strutture sintattiche (e.g., in componenti nominali, composti da un nome con relativi aggettivi e attributi e componenti verbali, composte da verbi e relativi avverbi e complementi, i quali a loro volta possono essere componenti nominali).

Qui ci limiamo a introdurre il modello più semplice per la rappresentazione di contenuti testuali, che prescinde dalla struttura sintattica della frase. In questo modello, detto "borsa di parole"(*bag of words*), ogni frase è rappresentata specificando quali parole vi compaiano. Si considerino per esempio le frasi seguenti

a: il fornitore è responsabile per i danni

b: il fornitore non risarcisce i danni

c: l'utente è sempre responsabile

Innanzitutto, estraiamo la lista delle parole, che riportiamo, nella sequenza in cui compaiono nelle frasi, a partire dalla prima.

	1	2	3	4	5	6	7	8	9	10	11
il	fornitore	è	responsabile	per	i	danni	non	risarcisce	l'	utente	sempre

A questo punto possiamo rappresentare ogni frase specificando quali parole vi compaiono (assumiamo che interessi solo sapere se una parola compaia o no nella frase, senza che rilevi il numero di volte in cui vi compare). Una rappresentazione semplice si può ottenere mediante vettori (sequenze) di numeri binari: nel nostro esempio, può essere rappresentata mediante una sequenza di 10 bit. Usando questa tecnica, le tre frasi appena riportante possono essere rappresentate come indicato nella Figura 4.7:

Si noti che il bit nella posizione i-esima nella sequenza è 1 se il lemma i-esimo è presente nella frase rappresentata, altrimenti è 0. Per esempio, la frase (c), "l'utente è sempre responsabile" è rappresentata dal vettore [001000000111]. In questo vettore la prima cifra, che corrisponde alla parola "il" ha valore 0 perché "il" non compare nella frase (c); invece la terza cifra, che corrisponde alla parola "è", ha valore 1, perché tale parola compare nella frase.

Nello sviluppo di Claudette il metodo della borsa di parole ha dato buoni risultati, anche se taluni miglioramenti si sono ottenuti considerando anche le struttura sintattiche. È importante ricordare che la conoscenza linguistica utilizzata dal sistema è limitata a quanto è espresso dalle strutture usate per rappresentare tale conoscenza. Nel caso della borsa di parole, tale conoscenza consiste esclusivamente nella specificazione di quali parole siano presenti o assenti nelle frase considerata, come indicato dai corrispondenti vettori di numeri binari.

Nell'addestramento del sistema, gli esempi sono stati forniti dalle frasi etichettate, o più esattamente dalla loro rappresentazione vettoriale. Per illustrare questo aspetto non consideriamo il training set usato da Claudette, ma prendiamo in considerazione il semplice esempio della Figura 4.7, e assumiamo di avere un dataset nel quale compaiano solo le parole usate nell'esempio, e di usare solo la rappresentazione mediante borsa di parole. Usando quel dataset, la frase (c), che verte sulla limitazione della responsabilità del fornitore darebbe luogo ad una rappresentazione del tipo ([0,0,1,0,0,0,0,0,0,1,1,1], ltd,3 dove

- il vettore: [0,0,1,0,0,0,0,0,0,1,1,1] indica le caratteristiche della clausola, cioè la presenza o assenza delle parole del vocabolario (secondo la rappresentazione adottata nella Figura 4.7);
- il codice ltd indica la categoria della clausola (limitazione di responsabilità) e il codice 3 specifica la relativa qualità giuridica (chiaramente vessatoria).

L'apprendimento del sistema consisterà nel costruire un modello che correli la presenza o l'assenza di combinazioni di parole (come rappresentate dai numeri 1 e 0 nei vettori) nelle frasi da valutare alle corrispondenti categorizzazioni e valutazioni giuridiche. Sulla base di tale modello, data una nuova frase (cioè il vettore che indica la presenza o assenza in essa delle parole del vocabolario), il sistema ne indicherà (predirà) categoria e valutazione giuridica.

In Claudette si sono utilizzati diversi metodi per costruire un modello di questo tipo. Oltre alle reti neurali (vedi Sezione 2.2.5) si è usato il modello della cosiddetta macchina a vettori di supporto (*Support Vector Machine* - SVM).

Secondo questo modello, il sistema impara ad assegnare dei pesi alle caratteristiche in considerazione, in modo che venga massimizzata la distanza tra gli esempi positivi e gli esempi negativi, cioè tra le linee parallele (i vettori di supporto) che toccano gli esempi positivi e negativi più vicini.[191] Il modello che ne risulterebbe può essere rappresentato graficamente (in modo semplificato) dalla Figura 4.8, nella quale una linea (più in generale un iperpiano) equidistante dai due vettori divide il piano (più in generale, lo spazio) in due settori. Gli elementi risultati positivi nell'insieme di addestramento (nel nostro caso, le frasi vessatorie) sono collocati a sinistra della linea e gli elementi negativi a destra. Ogni nuova frase viene classificata come positiva o negativa a seconda che, in base alle parole che contiene, venga a collocarsi nell'uno o nell'altro settore.

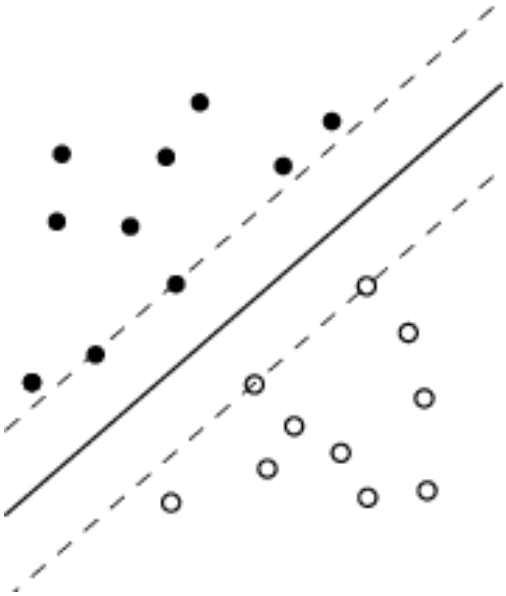

Figura 4.8: *Macchina a vettori di supporto*

La valutazione delle prestazioni del sistema. Un aspetto essenziale nei sistemi basati sull'apprendimento automatico concerne la valutazione delle loro prestazioni. A tale fine è necessario raffrontare i risultati forniti dal sistema, rispetto a un certo campione di casi, con i risultati corretti per quei casi.

Nel caso di Claudette, addestrato sulla base dell'insieme di addestramento costituito dalle indicazioni di giuristi esperti (sulla categorizzazione delle clausole e la loro qualità giuridica) il criterio di valutazione sarà rappresentato da quelle stesse indicazioni. Quindi, si sono sottoposti a Claudette casi non inclusi nel suo insieme di addestramento, e si è verificato se le risposte del sistema coincidessero con quelle che esperti giuridici avevano fornito a quegli stessi casi.[192]

Sulla base di tale confronto, le prestazioni del sistema si sono misurate con riferimento alle metriche classiche usate nella classificazione automatica: richiamo, precisione, e media armonica tra precisione e richiamo. La precisione è la valutazione della correttezza delle risposte positive del sistema, cioè dei casi in cui Claudette classifica correttamente una clausola come abusiva. Essa consiste, più esattamente, nel rapporto tra le clausole correttamente classificate come abusive (i cosiddetti veri positivi, VP), e l'insieme di tutte le clausole classificate come abusive, che include tanto i veri positivi quanto i falsi positivi (le clausole erroneamente classificate come abusive, FP):

$$precisione = \frac{VP}{VP + FP}$$

Il richiamo invece valuta la completezza delle risposte positive del sistema, cioè il rapporto tra le clausole correttamente classificate come abusive (VP) e l'insieme di tutte le clausole abusive presenti nell'insieme di addestramento, incluse quelle non riconosciute dal sistema (la somma dei veri positivi e dei falsi negativi, FN, le clausole abusive erroneamente considerate non-abusive):

$$richiamo = \frac{VP}{VP+FN}$$

Gli esperimenti condotti mostrano un valore superiore all'80% sia tanto la precisione quanto per il richiamo. Si sono ottenuti ottimi risultati sia usando reti neurali, sia ricorrendo alle macchine a vettori di supporto, sia combinando le due tecnologie.

Un prototipo del sistema è disponibile e liberamente utilizzabile al seguente indirizzo: `http://claudette.eui.eu/`. L'utente ha la possibilità di inserire un qualsiasi contratto online (in inglese), sottoponendolo all'esame di Claudette. Il sistema include anche un modulo che segnala in modo automatico i cambiamenti nei più importanti contratti online. Uno sviluppo recente del sistema Claudette riguarda l'uso di motivazioni giuridiche, cioè dell'indicazione delle ragioni per le quali una clausola può essere abusiva. Tali motivazioni sono state elaborate mediante reti di memoria (*Memory Network*), al fine di potenziare le prestazioni del sistema.[193] Grazie all'uso delle motivazioni, precisione e richiamo sono aumentate considerevolmente nelle sperimentazioni effettuate. In futuro ci si propone di usare le motivazioni anche per fornire all'utente una spiegazione semplice e intuitiva del perché una certa clausola è stata classificata dal sistema come (potenzialmente) abusiva.

L'applicazione disponibile online è ancora prototipale, ma ha incontrato il gradimento di diversi utilizzatori. Il sistema è infatti stato usato sia da consumatori interessati alla valutazione dei propri contratti, sia da studenti di giurisprudenza nell'ambito di corsi dedicati al diritto dei consumatori.

Si stanno sperimentando nuovi sviluppi, tra i quali l'identificazione delle omissioni di contenuti obbligatori, e la trattazione di documenti in linguaggi diversi dall'inglese, in particolare in tedesco e in italiano.

4.2.2 La previsione di fattori e la previsione basata su fattori

Nel caso di Claudette, il sistema correla direttamente le strutture testuali nei documenti giuridici e il risultato giuridico da predire, cioè la categoria e qualità giuridica della clausola. Il sistema svolge quindi un'analisi giuridica: determina se una certa formulazione testuale esprime un contenuto di un certo tipo e se esso sia conforme alla legge.

È possibile utilizzare tecniche per l'apprendimento automatico anche in altro modo, cioè per effettuare predizioni sulla base di una descrizione strutturata degli esempi, piuttosto che sulla base della forma linguistica degli stessi. In questo caso, bisogna associare

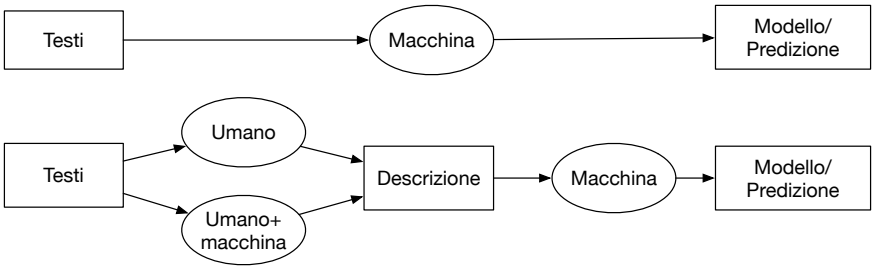

Figura 4.9: *Apprendimento (semi-) automatico*

al documento un insieme di descrittori, che ne esprimono le caratteristiche rilevanti. I descrittori sono in generale forniti da esperti umani, che possono però avvalersi di supporti automatici.

Questa soluzione può essere preferibile quando la predizione da effettuare riguardi testi complessi, nei quali è difficile, anche per una rete profonda, individuare connessioni significative tra strutture testuali e l'obiettivo della predizione (per esempio, quando si voglia prevedere l'esito di un giudizio sulla base di una descrizione dei fatti della causa). Inoltre, l'individuazione degli aspetti giuridicamente rilevanti è utile (e spesso necessaria) al fine di fornire una spiegazione significativa. Le diverse opzioni sono raffigurate nella Figura 4.9.

Quale esempio di modello che si fonda su descrizioni strutturate, aggiunte al documento in linguaggio naturale, possiamo ricordare la rappresentazione dei casi mediante fattori (Sezione 4.1.6. Se, da un lato, la specificazione dei fattori che caratterizzano i casi consente di ragionare sulla base di caratteristiche giuridicamente significative, dall'altro lato, l'impiego esteso di questo modello si scontra con il "collo di bottiglia" della rappresentazione della conoscenza, sia nella creazione del sistema, sia nel suo successivo impiego:

- l'assegnazione di fattori a tutti i precedenti da prendere in considerazione comporta un notevole lavoro giuridico (e conseguenti incertezze e idiosincrasie)
- l'utente che sottoponga una nuova situazione al sistema ha l'onere di specificare quali siano i fattori che compaiono in essa.

Una naturale estensione di questo indirizzo consiste nell'usare tecniche per l'elaborazione del linguaggio naturale e per l'apprendimento automatico, al fine di associare fattori (più in generale, descrittori) ai testi giuridici. A tal fine bisogna addestrare il sistema in modo che esso apprenda ad associare a certe formulazioni testuali la presenza di fattori corrispondenti.

Bisogna quindi predisporre un insieme di addestramento nel quale ciascuna unità di testo sia associata a fattori che ne indichino il rilievo giuridico (rispetto al risultato da

prevedere). Per esempio, la frase "Rossi assunse Gialli, dipendente di Verdi, nel 2021" potrà essere etichettata con il fattore F7, che indica l'assunzione, da parte del convenuto, di un dipendente del preteso titolare del segreto industriale. Il sistema, da un grande numero di frasi etichettate in questo modo, apprenderà ad associare a nuovi documenti i fattori corrispondenti, sulla base del lessico e della sintassi di quei documenti.[194]

Un'interessante metodologia è stata adottata in un recente progetto brasiliano, nel quale è stato sviluppato un sistema basato sull'apprendimento automatico al fine di facilitare la soluzione transattiva delle controversie tra viaggiatori e vettori in tema di ritardi e perdite.

In questo progetto, tecniche per l'apprendimento non-supervisionato sono state usate per raggruppare i precedenti in diversi insiemi, distinti in base alla loro tematica (come evincibile dal contenuto testuale dei precedenti), e per estrarne i termini più significativi. I risultati di questa analisi sono stati rielaborati da giuristi esperti, che hanno raffinato i risultati dell'identificazione automatica, eliminando termini non significativi, riducendo i sinonimi ad un unico termine, e aggiungendo ulteriori descrittori sulla base di un'analisi giuridica.

Si è quindi creato un insieme di addestramento, composto di precedenti, ciascuno caratterizzato dai descrittori ad esso applicabili. Questo insieme è stato usato in duplice modo. Da un lato, mediante una rete neurale si è creato un modello opaco, che fornisce predizioni rispetto a nuovi casi. Dall'altro lato, si sono estratte automaticamente regole di associazione, che connettono descrittori ed esiti. Mediante tali regole, raffinate e integrate da esperti giuristi, si è creato un modello "trasparente" in grado di fornire spiegazioni ai risultati forniti dal modello opaco.[195]

4.3 La giustizia predittiva

Nelle pagine seguenti si svilupperanno alcune considerazioni sulla possibilità di realizzare, mediante apprendimento automatico, funzioni di giustizia predittiva, nel senso di sistemi capaci di predire aspetti rilevanti di decisioni giuridiche.[196]

4.3.1 Le predizioni giuridiche

Nell'ambito dell'apprendimento automatico il termine predizione (come l'aggettivo predittivo) vengono usati in un senso estremamente ampio, così da applicarsi a qualsiasi inferenza intesa a espandere le informazioni disponibili su un certo problema, inferenze che possono riguardare non solo il futuro, ma anche il passato e il presente.

Quindi la predizione può coprire attività molto diverse da una prospettiva teorico-giuridica: la ricerca documentale (predire quali dati siano più rilevanti per chi formula un quesito, e quindi fornirli in risposta allo stesso), la creazione automatica di sommari

o massime di sentenze (predire, e quindi formulare, un estratto che corrisponda al testo)[197], l'associazione di descrittori al testo (predire, e quindi aggiungere, parole chiave o classificazioni a documenti), l'anticipazione di condotte future di imputati (predire se l'imputato commetterà nuovi reati, se turberà le prove o fuggirà in seguito alla libertà vigilata) o parti (predire quale accordo le parti stipulerebbero consensualmente, se dovessero accordarsi per le implicazioni finanziarie di una separazione o un divorzio), il supporto alla creazione di documenti (predire quali frammenti di testi passati possono essere utili per la redazione di un nuovo documento, per esempio, una sentenza), il supporto allo studio empirico del diritto (predire correlazioni, tendenze, attitudini), e altro ancora.

Limitando l'analisi alle predizioni che riguardino eventi rilevanti (anziché meri elementi informativi), possiamo tracciare le distinzioni seguenti.

Predizione su decisioni giudiziarie o su altri eventi. La predizione può riguardare (i) il risultato finale di un caso giudiziario o (ii) altro evento o circostanza che possa contribuire a determinare quel risultato. Nella prima ipotesi si tratta di anticipare il contenuto della statuizione del giudice. Nella seconda ipotesi si tratta di predire altri eventi rilevanti per la decisione, tipicamente, il comportamento futuro di una parte. Per esempio, presso i tribunali statunitensi sono già in funzione sistemi che quantificano il rischio di recidiva o di violazioni della libertà vigilata. Si tratta di previsioni che influiscono sulla determinazione della sentenza, fornendo al giudice indicazioni rilevanti per quantificare la pena da irrogare o da inasprire per il rischio di recidiva in concreto. Tali sistemi, quali ausili informativi del giudice, hanno sollevato importanti questioni etiche o giuridiche, con riferimento all'accuratezza delle loro previsioni, alla loro imparzialità, all'assenza di discriminazioni, al rispetto del principio del giudizio individuale.[198]

Predizioni su eventi presenti o passati. La predizione può riguardare (i) un evento futuro (per es., la decisione di un caso nuovo), o (ii) un evento passato (per es., il risultato di un caso già deciso). Nella prima ipotesi la predizione si configura come una previsione. Il caso paradigmatico consiste nell'anticipazione dell'esito di una controversia, sulla base di elementi disponibili prima della decisione (quali gli atti di parte rispetto alla sentenza, o la sentenza di primo grado rispetto alla decisione in appello). Nella seconda ipotesi la previsione dell'esito della controversia si fonda su elementi disponibili solo dopo che la controversia è stata decisa. Per esempio, un sistema potrebbe "predire" l'esito di una controversia sulla base di una porzione della sentenza che ha deciso quella controversia (la porzione che descrive i fatti della causa, o quella che presenta l'argomentazione giuridica). Questa funzione è stata realizzata in alcune applicazioni che vertono sulle decisioni della Corte europea dei diritti dell'uomo: sulla base di porzioni della sentenza che decide una controversia si è classificato l'esito di quella stessa controversia, indicando se consista nel riconoscere o nell'escludere la violazione di uno

dei diritti stabiliti dalla Convenzione.[199] In questi casi, quindi, si predice un aspetto (l'esito) di un evento passato (la controversia già decisa), sulla base di informazioni sullo stesso evento (per es., la motivazione in fatto o in diritto della decisione) disponibili solo dopo di esso. I risultati ottenuti con tali sistemi possono essere utili per arricchire un archivio di sentenze con le classificazioni ottenute, o per futuri sviluppi nella direzione della vera previsione.

Predizioni fondate su ragioni giuridiche o paragiuridiche. La predizione può essere fondata su (i) elementi normativamente rilevanti del caso in esame, che il giudice potrebbe considerare nel giustificare la propria decisione o (ii) elementi normativamente irrilevanti, che il decisore non dovrebbe prendere in considerazione. Nella prima ipotesi, il sistema automatico fonda le proprie predizioni su elementi giuridicamente significativi, e in particolare sugli aspetti della causa che giustificano l'una o l'altra decisione, secondo il diritto. Tali aspetti possono essere descritti nel linguaggio naturale, o in modo strutturato (elencando le caratteristiche o i fattori rilevanti, nella loro combinazione). Nella seconda, invece, il sistema basa la propria predizione su aspetti che, pur statisticamente correlati all'esito delle controversie, non forniscono una giustificazione giuridicamente rilevante. Per esempio, le decisioni dei giudici della Corte Suprema statunitense si sono previste sulla base di informazioni quali l'oggetto della causa, la procedura che ha condotto alla decisione, l'orientamento politico e le esperienze professionali dei giudici, ad esclusione di ogni informazione sulle ragioni di merito addotte dalle parti.[200] Allo stesso modo, è disponibile un sistema che prevede l'esito di casi in tema di brevetti sulla base delle caratteristiche delle parti, degli avvocati e dei giudici.[201] Sistemi di questo tipo possono essere utili alle parti interessate ad anticipare l'esito di una controversia, ma ovviamente lo sono molto meno per il giudice che deve valutare il merito della causa che è chiamato a decidere.

Predizioni spiegate o opache. Il sistema può (i) essere capace di fornire una spiegazione del suo output, adducendo ragioni comprensibili, o (ii) rimanere opaco, limitandosi a dare la propria indicazione sull'esito finale della controversia. Nella prima ipotesi, il sistema accompagna la propria predizione con l'indicazione degli aspetti giuridicamente rilevanti su cui essa si basa. Abbiamo visto come i sistemi basati su alberi di decisione possano fornire spiegazioni ripercorrendo il cammino che dalla radice dell'albero ha condotto alla decisione. Allo stesso modo, un sistema che si fondi su regole di associazione, può spiegare il risultato presentando le regole usate per raggiungerlo. Nella seconda ipotesi, esemplificata dalle reti neurali (e in particolare da quelle profonde), manca una spiegazione, poiché l'esibizione del processo mediante il quale il sistema, grazie all'attivazione dei neuroni, è giunto alla decisione è privo di significato per gli esseri umani.

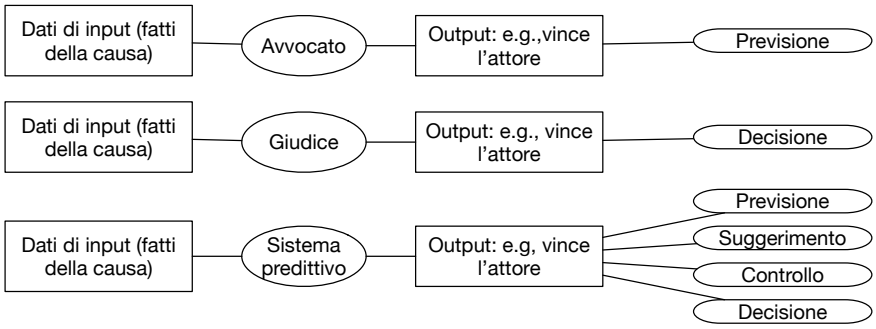

Figura 4.10: *Funzioni della determinazione dell'esito delle controversie*

Predizioni per giudici e per avvocati/parti. Nel caso prototipico di predizioni concernenti l'esito futuro di una controversia, è importante sottolineare come le indicazioni del sistema sono suscettibili di essere utilizzate in diverso modo a seconda di chi faccia uso di tali indicazioni, e del modo in cui il loro uso è regolato.

La Figura 4.10 indica come gli avvocati, i giudici, e i sistemi automatici formulano gli asserti che riguardano circa l'esito della causa. Giudizi aventi il medesimo contenuto (per esempio, "vince la causa l'attore al quale è riconosciuto un risarcimento pari all'intero ammontare del danno") possono però avere un significato pragmatico assai diverso.

Il prototipico giudizio dell'avvocato o della parte sull'esito di una controversia consiste di una previsione di fatto, che sarà confermata o smentita dal successivo comportamento del giudice. La previsione di come potrebbe essere decisa una futura controversia determina le aspettative e, quindi, il comportamento delle parti, anche prima dell'insorgere della controversia, mentre poi determina il loro comportamento processuale (in particolare la scelta se addivenire a una transazione).

Il prototipico giudizio del giudicante ha un effetto istituzionale. Si tratta di una statuizione che determina le posizioni giuridiche oggetto della controversia tra le parti, diventa incontestabile con il passaggio in giudicato e dà alla parte vittoriosa il potere di ottenere l'esecuzione.

La predizione fornita da un sistema automatico può avere un effetto diverso a seconda del contesto in cui la predizione viene formulata, e dalle attitudini di chi lo riceve.

Tale predizione sarà usata dagli avvocati e dalle parti in modo analogo a quello in cui esse avrebbero usato la predizione formulata dall'avvocato stesso, o da un consulente giuridico cui si fosse richiesto un parere, cioè come un elemento di fatto, di cui tener conto per definire le proprie strategie processuali.

La stessa predizione potrà essere considerata dal giudice quale suggerimento di cui

valutare criticamente la correttezza giuridica, al fine di decidere se uniformarsi o invece discostarsene. Il giudice potrebbe anche scegliere di guardare anche alla previsione del sistema solo dopo essersi formato una propria opinione sul caso. In questa ipotesi, la predizione del sistema fungerebbe da meccanismo di controllo: segnale di conferma, qualora il suggerimento del sistema coincida con la conclusione cui era giunto il giudice; segnale d'allarme, nel caso contrario.

Predizioni dotate o prive di valore giuridico. La predizione del sistema potrebbe essere dotata di valore giuridico, così da risultare una decisione vincolante per le parti, almeno quando queste non decidano di contestarla. Non risulta che esistano a tutt'oggi sistemi che operino in questo modo; a tal fine sarebbero necessarie norme giuridiche che attribuissero tale efficacia all'output prodotto dal sistema. Le ragioni contro questo utilizzo dei sistemi predittivi sembrano prevalere su quelle favorevoli, anche rispetto ai casi più semplici, aventi natura bagatellare. Ricordiamo che un sistema giudicante dotato di intelligenza artificiale, sulla base delle tecnologie oggi disponibili, non avrebbe alcuna consapevolezza dei fatti naturali o sociali su cui verte il proprio giudizio, non avrebbe alcuna percezione della loro rilevanza individuale e sociale. Opererebbe in modo "cieco" sulla base di rapporti sintattici tra input ed output, che consisterebbero:

- nell'inferenza logica che collega premesse (fatti del caso e regole) e conclusioni, nel caso di sistemi basati sulla conoscenza;
- nella correlazione statistica tra casi e predizioni, come colta dal modello costruito in base agli esempi, nel caso di sistemi basati sull'apprendimento automatico.

Le ragioni contro i giudizi automatici sono più forti rispetto ai sistemi opachi, cioè incapaci di indicare le ragioni delle indicazioni fornite, poiché l'assenza di spiegazioni impedisce un reale controllo critico, basato su una valutazione giuridica di norme e fatti. Un oracolo giuridico automatico che non fornisse spiegazioni sarebbe da respingere anche quando si limitasse a offrire suggerimenti al decisore umano. Quest'ultimo potrebbe, infatti, essere indotto a seguire passivamente le indicazioni del sistema (non essendo in grado di valutarle), rinunciando all'esame critico di fatti e norme e alla responsabilità della decisione.

4.3.2 Obiettivi e proxy delle predizioni giuridiche

Al fine di cogliere il senso delle predizioni giuridiche è utile distinguere l'obiettivo diretto della predizione automatica (quale decisione futura è maggiormente probabile?), e un possibile obiettivo finale per il quale il sistema viene utilizzato, che può essere diverso dall'obiettivo diretto (quale sarà/sarebbe una decisione "corretta"). Anche quando, nella maggioranza dei casi, l'obiettivo diretto tenda ad allinearsi all'obiettivo finale dell'utilizzo del sistema, l'allineamento può essere imperfetto. Per esempio, se un sistema per

la diagnosi medica viene addestrato sulla base delle diagnosi effettuate da medici umani (senza correggere le diagnosi che poi si rivelino errate), il suo obiettivo diretto sarà fornire indicazioni corrispondenti a quelle che avrebbe adottato il medico medio, obiettivo che si avvicina (almeno così sperano i pazienti), ma non si identifica con quello di effettuare diagnosi corrette. Similmente, se un sistema automatico per valutare l'affidabilità creditizia viene addestrato sulla base delle decisioni passate, il sistema avrà l'obiettivo diretto di prevedere le valutazioni del funzionario medio, che si avvicinano (in misura maggiore o minore), ma non si identificano con l'obiettivo finale di determinare la probabilità con cui il richiedente restituirà il debito.

Un sistema destinato a prevedere l'esito delle sentenze può aver l'obiettivo diretto di prevedere la sentenza del giudice medio, obiettivo che si avvicina a quello dell'utilizzatore avvocato o cittadino, solitamente interessati a prevedere il comportamento del giudice della propria causa, così da impostare di conseguenza il proprio comportamento processuale; quell'obiettivo ha una importanza più limitata per il giudice, il cui compito è adottare una decisione giuridicamente corretta. Come abbiamo osservato sopra, per il giudice può essere più utile avere un sistema che non si limiti a fornire una sola decisione possibile, ma indichi piuttosto, nel caso di divergenze nella giurisprudenza, accanto all'ipotesi maggioritaria, anche quelle che trovino meno seguito.

Nelle applicazioni di giustizia predittiva, si deve individuare con precisione l'obiettivo diretto delle predizioni del sistema, per valutarne la rilevanza rispetto all'obiettivo finale che ci si propone di raggiungere. Per esempio, possiamo distinguere due tipi di sistemi impiegati al fine di determinare se concedere o meno la libertà vigilata a un imputato. Il primo sistema potrebbe essere addestrato sulla base di un insieme di addestramento che associa richieste di libertà vigilata alle corrispondenti decisioni dei giudici. La predizione diretta del sistema riguarderebbe in questo caso la decisione del giudice medio (o anche le diverse soluzioni risultanti da diversi indirizzi presenti nella giurisprudenza). Il secondo sistema potrebbe invece effettuare le proprie predizioni sulla base di esempi che riportino il comportamento successivo di un insieme di imputati che abbiano goduto della libertà vigilata. In questo caso il sistema predirebbe la probabilità che l'imputato reiteri il reato o si sottragga alla pena, fornendo un'indicazione più "oggettiva" di cui il giudice potrebbe tener conto.

La possibilità di una predizione "oggettiva" non esiste quando, come nel caso della predizione dell'esito di controversie giudiziarie, l'obiettivo da prevedere sia proprio la decisione umana e non vi sia un riscontro obiettivo della correttezza della decisione. In questo caso, il sistema si limiterà a "predire" quale nuova decisione potrà allinearsi alle decisioni passate nel loro complesso (alla decisione del giudice "medio"), riproducendo le virtù (accuratezza, imparzialità) così come i vizi (imprecisioni, iniquità) dell'indirizzo prevalente. Per cogliere il rilievo fattuale e quello normativo di un sistema inteso a prevedere le decisioni dei giudici è importante chiarire il rapporto tra decisione media e decisione corretta.

Assumiamo che il sistema sia in grado di prevedere la decisione media dei giudici. Per esempio, se il 50% dei giudici decide cause di un certo tipo accordando un risarcimento di circa 1000 euro, mentre il 25% dei giudici accorda un risarcimento di 500 euro e altri di 1500, il sistema prevedrà un risarcimento atteso di 1000 euro. Qualora, una volta introdotto il sistema, tutti i giudici ne seguissero le indicazioni, accordando sempre un risarcimento di 1000 euro a tutti i nuovi casi, si sarebbe raggiunto un risultato importante. Si sarebbe ridotto il "rumore" o la dispersione nei risultati del sistema: il sistema non tratterebbe in modo diverso casi eguali.[202] Ciò faciliterebbe la previsione da parte di cittadini, che saprebbero quale risposta attendersi dal sistema giuridico. Ma si potrebbe dire che in tal modo sia aumentata la "giustizia" del sistema? Ciò dipende da dove stia la decisione più giusta o giuridicamente corretta. Se essa è compresa tra 500 e 1500, consistendo, per esempio, nel valore di 800, sembra che la risposta debba essere positiva: convergendo su tale cifra si limita l'ingiustizia media, cioè la divergenza media tra il risarcimento attribuito e quello ideale (essa ammonterà a 200 per ogni persona, piuttosto che $200 * 0.5 + 300 * 0.25 + 700 * 0.25 = 100 + 75 + 175 = 350$). Tuttavia, è da osservare che il diritto possiede altri metodi per rimediare alle divergenze (come il meccanismo del precedente), e che il disaccordo può stimolare un dibattito che conduca a convergere verso le soluzioni migliori.

Bisogna peraltro osservare che, come già ricordato, non necessariamente un sistema dovrà essere destinato a fornire solo la soluzione "media". Questo orientamento sembra ragionevole quando dal sistema ci si attenda un'unica soluzione, possibilmente per dare esecuzione ad essa senza un ulteriore intervento umano. Quando, invece, il sistema sia destinato a interagire con l'uomo, fornendogli elementi di conoscenza —a questo riguardo si parla talvolta di computazione cognitiva, *cognitive computing*— può essere utile che il sistema fornisca indicazioni alternative, alla luce degli esempi divergenti contenuti nel suo insieme di addestramento (o delle regole divergenti nella sua base di conoscenza), come si è evidenziato nei paragrafi precedenti.

In conclusione, un sistema "predittivo", o anzi previsionale, potrebbe contribuire a migliorare la pratica del diritto, se sviluppato in modo da fornire indicazioni accurate e accompagnate da spiegazioni adeguate, e applicato con piena consapevolezza delle sue funzionalità e dei suoi limiti, nel rispetto di etica e diritto. Per giudici, avvocati, e parti si tratterebbe di un aiuto ad avere una migliore conoscenza del diritto in azione, come risulta dalle decisioni passate, che ciascuno potrebbe utilizzare a seconda delle proprie funzioni e dei propri interessi. La possibilità di un uso di sistemi di intelligenza artificiale è stata prospettata da autorevoli giuristi anche con riferimento al nostro ordinamento. Si è anzi affermato che un sistema automatico potrebbe avere un ruolo assimilabile a quello dell'Avvocato Generale nei processi davanti alla Corte di Giustizia, proponendo considerazioni autorevoli, ma non vincolanti per il giudicante.[203] A tutt'oggi, peraltro, non sono disponibili sistemi in grado di rispondere adeguatamente a tutti i requisiti appena indicati, ma il rapido sviluppo delle tecnologie, accompagnato dalla crescente attenzio-

ne per gli aspetti etici e giuridici, ci fanno ben sperare per il prossimo futuro. Infine, è appena il caso di ricordare che, benché qui si sia posto l'accento sull'uso dell'IA nella giustizia, non mancano ricerche e applicazioni dedicate ad altre funzioni pubbliche, come la legislazione[204] o la pubblica amministrazione.[205]

Conclusione

Come si è osservato nella prefazione, il presente volume si propone di introdurre il giurista all'intelligenza artificiale, con un'esposizione sintetica e accessibile anche a chi sia privo di conoscenze tecniche, ma non superficiale.

Spero che il lettore, giunto alla fine del volume, possa disporre di alcuni strumenti per meglio cogliere la dinamiche tecnologiche e sociali dell'IA, e forse anche per contribuire ad indirizzarle.

L'intelligenza artificiale ha infatti bisogno del diritto; solo grazie ad un'efficace disciplina giuridica l'IA potrà svilupparsi nel rispetto delle esigenze individuali e dei valori sociali. Si tratta di una condizione meramente necessaria, non sufficiente per tale obiettivo: sono altresì necessari sviluppi tecnologici nella direzione di un'intelligenza artificiale "centrata sull'uomo" (*human centred AI*), che il diritto può tuttavia contribuire a promuovere.

Il diritto, d'altro lato, ha bisogno dell'intelligenza artificiale, le cui tecnologie possono contribuire a migliorare il lavoro del giurista, e l'efficienza-efficacia del sistema giuridico. Tali tecnologie sono necessarie, in particolare, per affrontare tematiche collegate alla tecnologie dell'informazione e alla stessa intelligenza artificiale (es. il controllo su informazioni e interazioni online, la verifica della legalità ed equità degli algoritmi, ecc.) Più in generale l'IA può contribuire a rendere maggiormente informata e consapevole l'attività del giurista, potenziandone le capacità di analisi e decisione.

I due bisogni appena menzionati potranno essere soddisfatti solo se il giurista potrà contribuire in modo efficace e consapevole alla regolazione dell'IA e al suo uso nel diritto. Spero che questo lavoro possa rappresentare un modesto ma utile contributo in questa direzione.

Note

[1]Sofocle, nei versi 333 e 334, usa l'aggettivo "deinós", il cui significato include sia il meraviglioso (o stupefacente) sia il terribile (o tremendo).

[2]Il lettore interessato a una traduzione italiana potrà cercare la più recente interrogando Internet o usando la funzione di ricerca disponibile presso il sito del Sistema Bibliotecario Nazionale (`www.sbn.it`).

[3]R. L. Gregory, «Intelligence», in *The Oxford Companion to the Mind*, a cura di R. L. Gregory, Oxford University Press, 1987, p. 375-379, p. 375. *Testo originale*: "Innumerable tests are available for measuring intelligence, yet no one is quite sure of what intelligence is, or even of just what is that the available tests are measuring."

[4]Vedi S. J. Russell e P. Norvig, *Artificial Intelligence. A Modern Approach*, 4ª ed., Pearson, 2021, Capitolo 1.

[5]Anche se Aristotele preferisce qualificare l'uomo, più modestamente, come "animale bipede senza penne" (Aristotele, *Opere*, vol. VI: *Metafisica*, Laterza, 1996, p. 1037).

[6]Queste parole furono pronunciate da Vico nell'opera giovanile G. Vico, *De antiquissima Italorum sapientia*, Laterza, [1709] 1917, par. 1, 1. Qui le riprendo quale mero spunto, senza avere la pretesa di cogliere con esattezza filologica il pensiero di Vico, che applicava l'idea della corrispondenza tra il vero e il fatto non (ovviamente) all'IA, ma bensì alla matematica e alla storia, conoscibili in quanto creazioni umane.

[7]Sul tema del rapporto tra mente, sistema nervoso e corpo, si veda da ultimo A. R. Damasio, *Feeling & knowing: making minds conscious*, Penguin, 2021.

[8]S. J. Russell e P. Norvig, *Artificial Intelligence. A Modern Approach*, 3ª ed., Prentice Hall, 2010, p. 2.

[9]J. Haugeland (a cura di), *Artificial Intelligence: The Very Idea*, MIT, 1985. *Testo originale*: "The exiting new effort to make computers think [...] machines with minds, in the full and literal sense."

[10]R. E. Bellman, *An Introduction to Artificial Intelligence: Can Computer Think?*, Boyd e Fraser, 1978. *Testo originale*: "The automation of activities that we associate with human thinking, activities such as decision-making, problem-solving, learning."

[11]E. Charniak e D. McDermott, *Introduction to Artificial Intelligence*, Addison-Wesley, 1985. *Testo originale*: "The study of mental faculties though the use of computational models."

[12]P. H. Winston, *Artificial Intelligence*, Addison-Wesley, 1992. *Testo originale*: "The study of the computations that make it possible to perceive, reason, and act."

[13]R. Kurzweil, *The Age of Spiritual Machines*, Orion, 1999. *Testo originale*: "The art of creating machines that perform functions that require intelligence when performed by people."

[14]E. Rich e K. Knight, *Artificial Intelligence*, McGraw-Hill, 1991. *Testo originale*: "The study of how to make computers do things at which, at the moment, people are better."

[15]D. L. Poole et al., *Computational Intelligence: A Logical Approach*, Oxford University Press, 1998. *Testo originale*: "Computational Intelligence is the study of the design of intelligent agents."

[16]N. Nilsson, *Artificial Intelligence: A New Synthesis*, Morgan Kaufmann, 1998. *Testo originale*: "AI [. . .] is concerned with intelligent behaviour in artifacts."

[17]Su questa connessione, vedi a esempio J. L. Pollock, *Cognitive Carpentry: A Blueprint for How to Build a Person*, MIT, 1995.

[18]R. A. Brooks, *Intelligence without Reason*, AI Memo, MIT Artificial Intelligence Laboratory, 1991:*Testo originale*: "Problem solving behavior, language, expert knowledge and application, and reason, are all pretty simple once the essence of being and reacting are available. That essence is the ability to move around in a dynamic environment, sensing the surroundings to a degree sufficient to achieve the necessary maintenance of life and reproduction. This part of intelligence is where evolution has concentrated its time—it is much harder. I believe that mobility, acute vision and the ability to carry out survival related tasks in a dynamic environment provide a necessary basis for the development of true intelligence".

[19]Anche se il contesto in cui quelle capacità si formarono, la vita di piccoli gruppi di raccoglitori-cacciatori dell'età della pietra, è molto diverso da quello attuale. Ciò fa si che alcune nostre tendenze psicologiche naturali—come la tendenza a non tener conto adeguatamente di rischi aventi bassa probabilità di verificarsi, o a sottovalutare vantaggi o svantaggi destinati a verificarsi in futuro lontano—non siano adeguate a società nelle quali la vita è lunga e relativamente sicura, ed è possibile fare piani a lungo termine. Su queste e altre innate tendenze umane all'irrazionalità si sono soffermati recentemente gli studi di economia comportamentale (*behavioural economics*), che hanno dato luogo a importanti studi in tema di diritto ed economia (*behavioural law and economics*).

[20]H. A. Simon, *Models of Man: Social and Rational*, Wiley, 1957, p. 198: "La capacità della mente umana nel formulare e nel risolvere problemi complessi è molto piccola rispetto alla dimensione dei problemi la cui soluzione è richiesta per un comportamento oggettivamente razionale nel mondo reale o anche per un'approssimazione praticabile a tale razionalità oggettiva". *Testo originale*: "The capacity of the human mind for formulating and solving complex problems is very small compared with the size of the problems whose solution is required for objectively rational behaviour in the real world or even for a practicable approximation to such objective rationality."

[21]H. A. Simon, *Models of Man: Social and Rational*, Wiley, 1957, p. 198. *Testo originale*: "We cannot within practicable computational limits generate all the admissible alternatives and compare their respective merits. Nor can we recognize the best alternative, even if we are fortunate enough to generate it early, until we have seen all of them. We satisfice by looking for alternatives in such a way that we can generally find an acceptable one after only moderate search."

[22]G. Gigerenzer e W. Gaissmaier, «Heuristic Decision Making», in *The Annual Review of Psychology*, vol. LXII (2011), p. 451-82.

[23]L. Floridi e F. Cabitza, *Intelligenza artificiale*, Bompiani, 2021: "[O]ggi l'AI non è il matrimonio tra ingegneria (artefatti) e biologia (intelligenza almeno animale, se non umana), ma il divorzio dell'agire (agency) dalla necessità di essere intelligenti per aver successo. Credo che sia questo l'unico modo corretto di interpretare l'AI = agere sine intelligere. Di qui segue che il modo migliore di concettualizzare l'AI è nei termini di una risorsa crescente di agency (come dicevo, "agenzia" non si dice in italiano) interattiva, autonoma e spesso in grado di autoapprendere, capace di affrontare un numero crescente di problemi e compiti che altrimenti richiederebbero l'intelligenza e l'intervento umano (e possibilmente una quantità illimitata di tempo) per essere eseguiti con successo."

[24]E. Esposito, *Artificial Communication*, MIT, 2022.

[25]Vedi per esempio C. Castelfranchi e F. Paglieri, «The Role of Beliefs in Goal Dynamics: Prolegomena to a Constructive Theory of Intention», in *Synthese*, vol. CLV (2007), p. 237-63. Sulle architetture fondate sull'idea che sistemi digitali possano agire attuando intenzioni basate su desideri e credenze (il modello Belief-Desire-Intention – BDI) vedi: per i concetti fondamentali, M. Bratman, *Intentions, Plans and Practical Reasoning*, Harvard University Press, 1987; per una rassegna drecent di modelli informatici, L. de Silva et al., «BDI Agent Architectures: A Survey», in *Proceedings of IJCAI-2020*, 2020, p. 4914-4921.

[26]L'idea che nozioni psicosociali possano essere applicate anche alle macchine risale alle origini della cibernetica. Vedi A. Rosenblueth et al., «Purpose and Teleology», in *Philosophy of Science*, vol. X (1943,), p. 18-24. L'uso di concetti quali credenza, intenzione, comprensione, fiducia per descrivere i sistemi di IA si ritrova nella maggior parte lavori tecnologici dedicate all'IA. Vedi S. J. Russell e P. Norvig, *Artificial Intelligence. A Modern Approach*, 4ª ed., Pearson, 2021. Per un tentativo di costruire alcuni concetti giuridici in essi si applichino anche a sistemi intelligenti, vedi: G. Sartor, «Cognitive Automata and the Law: Electronic Contracting and the Intentionality of Software Agents», in *Artificial Intelligence and Law*, vol. XVII (2009), p. 253-90; F. Lagioia e G. Sartor, «AI systems under criminal law: A legal analysis and a regulatory perspective», in *Philosophy of Technology*, vol. XXXIII (2019), p. 433-465.

[27]J. McCarthy, *What Is Artificial Intelligence*, rapp. tecn., Stanford University, 2007. *Testo originale*: "[AI] is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable".

[28]AI-HLEG, High-Level Expert Group on Artificial Intelligence, *A definition of AI: Main capabilities and scientific disciplines*, European Commission, 2019. *Testo originale*: "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions."

[29]AI-HLEG, High-Level Expert Group on Artificial Intelligence, *A definition of AI: Main capabilities and scientific disciplines*, European Commission, 2019. *Testo originale*: "As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization),

and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)."

[30]AI-HLEG, High-Level Expert Group on Artificial Intelligence, *A definition of AI: Main capabilities and scientific disciplines*, European Commission, 2019.*Testo originale*: "A robot is a physical machine that has to cope with the dynamics, the uncertainties and the complexity of the physical world. Perception, reasoning, action, learning, as well as interaction capabilities with other systems are usually integrated in the control architecture of the robotic system. In addition to AI, other disciplines play a role in robot design and operation, such as mechanical engineering and control theory. Examples of robots include robotic manipulators, autonomous vehicles (e.g. cars, drones, flying taxis), humanoid robots, robotic vacuum cleaners, etc."

[31]G. A. Bekey, «Current Tends in Robotics: Technology and Ethics», in *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT, 2012, p. 18-34, p. 18. *Testo originale*: "A machine, situated in the world, that senses, thinks, and acts. Thus, a robot must have sensors, processing ability that emulates some aspects of cognition, and actuators. Sensors are needed to obtain information from the environment. Reactive behaviours (like the stretch reflex in humans) do not require any deep cognitive ability, but on-board intelligence is necessary if the robot is to perform significant tasks autonomously, and actuation is needed to enable the robot to exert forces upon the environment. Generally, these forces will result in motion of the entire robot or one of its elements (such as an arm, a leg, or a wheel)."

[32]Russell e Norvig, *Artificial Intelligence. A Modern Approach* cit., Capitolo 26.

[33]`www.research.philips.com`. *Testo originale*: "This is our vision of 'Ambient Intelligence': people living easily in digital environments in which the electronics are sensitive to people's needs, personalized to their requirements, anticipatory of their behavior and responsive to their presence."

[34]Di "disincantamento del mondo" (*Entzauberung der Welt*) parlava Max Weber [1864-1920] con riferimento all'attitudine mentale propria della società scientifico-tecnologica, in una celebre scritto dedicato alla "scienza come professione" (M. Weber, *La scienza come professione*, Bompiani, [1919] 2008.

[35]Ciò non esclude che il funzionamento di un oggetto intelligente sia spiegabile anche secondo le leggi fisiche che ne governano il funzionamento, ma di fatto —data la complessità tecnologica dell'oggetto stesso, la variabilità del suo comportamento, e l'imperscrutabilità del suo meccanismo interno— solo un'interpretazione teleologica del suo comportamento ci consentirà di comprenderlo e solo l'assunzione di una prospettiva comunicativa ci consentirà di interagire con esso in modo appropriato. L'idea che sia necessario adottare tale prospettiva (che egli chiama prospettiva intenzionale, *intentional stance*) nell'interagire con sistemi complessi, capaci di azione finalistica, è sviluppata da Daniel C. Dennett, vedi D. C. Dennett, *Consciousness Explained*, Little Brown, 1991; D. C. Dennett, *Kinds of Minds: Towards an Understanding of Consciousness*, Basic, 1997.

[36]G. Sartor e A. Loreggia, *Study: The impact of algorithms for online content filtering or moderation ("upload filters")*, European Parliament, 2020.

[37]L. Floridi, *Etica dell'intelligenza artificiale*, Cortina, 2021, p. 218.

[38]N. Bostrom, *Superintelligence*, Oxford University Press, 2014.

[39]Questa possibilità era già descritta da S. Butler, «Darwin Among The Machines», in *The Press* (June 13th 1863). Lo stesso Butler, nel romanzo Erewhon immagina una guerra tra fautori e oppositori delle macchine intelligenti, conclusasi con la vittoria di questi ultimi.S. Butler,

Erewhon Or Over the Range, Trubner, 1872. Il tema è ripreso nella saga di Dune,F. Herbert, *Dune*, Chilton, 1965, che si svolge in un mondo senza robot, esito di una guerra vittoriosa (chiamata Butlerian Jihad) contro le macchine intelligenti.

[40]Questo tema è sviluppato da N. Bostrom, *Superintelligence*, Oxford University Press, 2014 e M. Tegmark, *Life 3.0. Being Human in the Age of Artificial Intelligence*, Knopf, 2017.

[41]R. Kurzweil, *The Singularity is Near*, Viking, 2005.

[42]Vedi J. R. Searle, «Minds, Brains and Programs», in *The Behavioural and Brain Science* (1980), p. 417-57. Altri autori (futurologi o scrittori di fantascienza) parlano di IA forte in un significato che si avvicina a quello di "intelligenza generale artificiale", per far riferimento all'obiettivo di realizzare sistemi artificiali le cui competenze cognitive sono generali, e potenzialmente raggiungono o superano le capacità umane.

[43]J. R. Searle, «Minds, Brains and Programs», in *The Behavioural and Brain Science* (1980), p. 417-57, p. 417. *Testo originale*: "The appropriately programmed computer really is a mind in the sense that computers given the right programs can be literally said to understand and have other cognitive states."

[44]Tratta da J. Copeland, *Artificial Intelligence*, Blackwell, 1993.

[45]A. M. Turing, «Computer Machinery and Intelligence», in *Mind*, vol. LIX (1950), p. 433-60. *Testo originale*: "I shall replace the question ['Can machines think?'] by another, which is closely related to it and is expressed in relatively unambiguous words. The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (*A*), a woman (*B*), and an interrogator (*C*) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels *X* and *Y*, and at the end of the game he says either '*X* is *A* and *Y* is *B*' or '*X* is *B* and *Y* is *A*' [...] In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (*B*) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as 'I am the woman, don't listen to him!' to her answers, but it will avail nothing as the man can make similar remarks. We now ask the question, 'What will happen when a machine takes the part of *A* in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?'."

[46]Si tiene una competizione annuale, denominata "Loebner price", che prevede un premio molto elevato (finora mai assegnato) per il sistema che superi il test di Turing, e un premio più basso (assegnato ogni anno), per il sistema che riesca a nascondere la propria identità più a lungo.

[47]Searle, «Minds, Brains and Programs» cit.

[48]Sui limiti dell'AI, vedi Russell e Norvig, *Artificial Intelligence. A Modern Approach* cit., Sezione 27.1. Per una discussione filosofica dell'argomento, vedi Dennett, *Consciousness Explained* cit. La tesi che gli elaboratori elettronici oggi disponibili sono fondamentalmente diversi dal cervello umano, e quindi incapaci di dar vita al fenomeno della coscienza, è stata ripresa dal neurologo Gerald Edelman, premio Nobel per la medicina, vedi G. M. Edelman e G. Tononi, *A Universe of Consciousness*, Basic, 2000.

[49]L'idea di "pensiero cieco" o "pensiero simbolico" può essere ricondotta a Leibniz, per usa questi termini per caratterizzare il processo mediante il quale "ragioniamo con le parole, con il loro oggetto virtualmente assente dalla nostra mente".G. W. Leibniz, *Meditations on Knowledge, Truth, and Ideas*, a cura di J. Bennett, 2004.

[50]Si veda, recentemente D. Hofstadter, «The Shallowness of Google Translate», in *The Atlantic* (gen. 2018).

[51]Vedi L. Floridi e M. Chiriatti, «GPT-3: Its Nature, Scope, Limits, and Consequences», in *Minds and Machines* (2020).

[52]E. M. Bender e A. Koller, «Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data», in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, p. 5185-5198.

[53]Y. Bisk et al., «Experience Grounds Language», in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online nov. 2020, p. 8718-8735.

[54]Il termine *robot* deriva dalla parola ceca *robota*, che significa "lavorare", e potrebbe quindi tradursi con "lavoratore", ma la traduzione inglese conservò il termine originale, che fu poi adottato in altre lingue.

[55]A. C. Clarke, *2001: A Space Odyssey*, (Prima edizione 1968), Penguin, 2000, pubblicata per la prima volta nel 1969.

[56]Bostrom, *Superintelligence* cit.

[57]S. J. Russell, *Human Compatible*, Viking, 2019.

[58]I. Asimov, *I, Robot*, (Prima edizione 1950), Collins, 1968, p. 8. *Testo originale*: "(1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. (2) A robot must obey orders given to it by human beings, except where such orders would conflict with the First Law. (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law". Le tre leggi sulla robotica furono formulate da Asimov per la prima volta nel 1942 (nel racconto *Runaround*, in italiano "circolo vizioso"), e si ritrovano in numerose opere dello stesso autore.

[59]I. Asimov, *Robots and Empire*, Collins, 1985, Cap. 18. *Testo originale*: "a robot may not injure humanity, or, by inaction, allow humanity come to harm".

[60]G. W. F. Hegel, *Fenomenologia dello spirito*, Nuova Italia, 1933, Sezione IV, A.

[61]Questo tema è affrontato in I. McEwan, *Machines Like Me*, Vintage, 2019.

[62]La linea di ricerca "neurale" rimarrà marginale per alcuni decenni ma, come si vedrà nella Sezione 1.4.4, sarà ripresa, tornando al centro dell'IA, negli anni '80.

[63]Come afferma la proposta della conferenza, elaborata da John McCarthy e Marvin Minsky (che in seguito daranno contributi fondamentali all'IA), assieme a Nathan Rochester e Shannon (vedi J. McCarthy et al., *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 1956, `http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html`). *Testo originale*: "Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it."

[64]A. Newell e H. A. Simon, «Computer Science as Empirical Inquiry: Symbols and Search», in *Communications of the ACM*, vol. XIX (1976), p. 113-26. *Testo originale*: "A physical symbol system consists of a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure). Thus a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical

way (such as one token being next to another). At any instant of time the system will contain a collection of these symbol structures. Besides these structures, the system also contains a collection of processes that operate on expressions to produce other expressions: processes of creation, modification, reproduction, and destruction. A physical symbol system is a machine that produces through time an evolving collection of symbol structures."

[65] Ivi. *Testo originale*: "A physical symbol system has the necessary and sufficient means for intelligent action."

[66]Ivi. *Testo originale*: "A physical-symbol system is an instance of a universal machine. [...] Thus the symbol-system hypothesis implies that intelligence will be realized by a universal computer."

[67]Un odierno PC può eseguire programmi di IA con rapidità superiore a quella delle costose *Lisp machines* degli anni '70.

[68]J. McCarthy e P. Hayes, «Some Philosophical Problems of Artificial Intelligence», in *Readings in Artificial Intelligence*, a cura di B. Webber e N. Nilsson, Morgan Kaufmann, 1987, p. 431-50, p. 27. *Testo originale*: "An entity is intelligent if it has an adequate model of the world (including the intellectual world of mathematics, understanding of its own goal and other mental processes), if it is clever enough to answer a wide variety of questions on the basis of that model, if it can get additional information from the external world when required, and can perform such tasks in the external world as its goal demand and its physical abilities permit."

[69]McCarthy e Hayes usano il termine "euristica" nel senso generico di "attinente alla ricerca di soluzioni", piuttosto che nel senso specifico di "euristica" quale metodo congetturale caratterizzante la razionalità limitata.

[70]J. McCarthy e P. Hayes, «Some Philosophical Problems of Artificial Intelligence», in *Readings in Artificial Intelligence*, a cura di B. Webber e N. Nilsson, Morgan Kaufmann, 1987, p. 431-50, p. 27. *Testo originale*: "The epistemological part is the representation of the world in such a form that the solution of the problems follows from the facts expressed in the representation. The heuristic part is the mechanism that on the basis of the information solves the problem and decides what to do."

[71]Tra i sistemi più frequentemente citati nella letteratura, ricordiamo Dendral, utilizzato per l'identificazione di strutture molecolari sulla base dei dati forniti da apparecchiature sperimentali, e Mycin, utilizzato nella diagnosi medica.

[72]Per una critica dei sistemi esperti, classici riferimenti sono J. Weizenbaum, *Computer Power and Human Reason: From Judgement to Calculation*, Freeman, 1977 e H. Dreyfus e S. Dreyfus, *Mind over Machine*, Blackwell, 1986. Per una teoria computabile della coerenza, vedi P. Thagard, *Conceptual Revolutions*, Princeton University Press, 1992 e, per un'applicazione in ambito giuridico, T. J. M. Bench-Capon e G. Sartor, «A Quantitative Approach to Theory Coherence», in *Proceedings of the Fourteenth Annual Conference on Legal Knowledge and Information Systems (JURIX)*, IOS, 2001, p. 53-62 e T. J. M. Bench-Capon e G. Sartor, «A Model of Legal Reasoning with Cases Incorporating Theories and Values», in *Artificial Intelligence*, vol. CL (2003), p. 97-142.

[73]I giochi furono uno dei primi campi di indagine e sperimentazione l'IA, e ricerche e sviluppi in materia sono continuate fino ai nostri giorni. Nel 1997 il computer Big Blue, prodotto da IBM, riuscì a sconfiggere il campione del mondo di scacchi Gary Kasparov. Recentemente Deep Mind ha realizzato un sistema in grado di sconfiggere il campione di mondo del go, un antico gioco cinese, estremamente complesso.

[74]Gli scienziati sono molto divisi sulle prospettive dello sviluppo dell'IA, in particolare sulla possibilità che si raggiunga la superintelligenza, vedi Bostrom, *Superintelligence* cit. Per una prospettiva critica circa la possibilità di realizzare l'IA generale e la superintelligenza, si veda Floridi, *Etica dell'intelligenza artificiale* cit.

[75]Lo studioso statunitense Layman Allen, propose di usare i connettivi della logica formale per esprimere i rapporti tra proposizioni nei testi normativi e nei contratti, e di specificarne in modo univoco (tramite indentazione) l'ambito di applicazione. Per distinguere quei connettivi dalle espressioni corrispondenti nel linguaggio naturale, suggerì di usare le maiuscole (come si è fatto qui negli esempi). Vedi L. E. Allen, «Towards a Normalized Language to Clarify the Structure of Legal Discourse», in *Deontic Logic, Computational Linguistics, and Legal Informations Systems*, a cura di A. A. Martino, North Holland, 1982, p. 349-407.

[76]C. J. Hogger e R. A. Kowalski, «Logic Programming», in *Encyclopedia of Artificial Intelligence*, a cura di S. C. Shapiro e D. N. Y. Eckroth, Wiley, 1987, p. 544-58, p. 544.

[77]Per una discussione del ragionamento defeasible, con particolare riferimento al diritto, vedi G. Sartor, «Defeasibility in Law», in *Handbook of Legal Reasoning and Argumentation*, a cura di G. Bongiovanni et al., Springer, 2018, p. 315-64.

[78]Questa soluzione è stata adottata in particolare nel Logical English, un sistema che consente al giurista di rappresentare contenuti giuridici in un formalismo vicino al linguaggio naturale, e provvede a tradurre automaticamente i contenuti così espressi in linguaggi di programmazione logica, come Prolog o ASP (*Answer Set Programming*). Vedi R. Kowalski e A. Datoo, «Logical English meets legal English for swaps and derivatives», in *Artificial Intelligence and Law*, vol. XXX (2022), p. 163-197.

[79]Uso la locuzione "correttezza" quale formulazione neutrale rispetto al problema della possibilità di assegnare valori di verità alle norme e alle conseguenze derivabili da esse, problema che non può essere affrontato in questa sede.

[80]Sull'idea di conoscenza tacita, vedi M. Polanyi, *The tacit dimension*, University of Chicago Press, [1966] 2009.

[81]Turing, «Computer Machinery and Intelligence» cit.

[82]F. Schauer, *Profiles, Probabilities and Stereotypes*, Belknap, 2003.

[83]Per un'introduzione aggiornata ai concetti fondamentali delle reti neurali, vedi Russell e Norvig, *Artificial Intelligence. A Modern Approach* cit., Sezione 18.7. Un classico riferimento è D. E. Rumelhart e J. L. McClelland (a cura di), *Parallel Distributed Processes: Explorations in the Microstructure of Cognition*, MIT, 1986.

[84]Per una discussione sui limiti dell'apprendimento automatico,G. Marcus e E. Davis, *Rebooting AI: building artificial intelligence we can trust*, Pantheon Books, 2019.

[85]Il capitolo è stato scritto da Francesca Lagioia e Giovanni Sartor. Si basa, in gran, parte, sui seguenti contributi: F. Lagioia e G. Sartor, «Profilazione e decisione algoritmica: dal mercato alla sfera pubblica», in *Federalismi*, vol. XI (2020), p. 85-110; G. Sartor e F. Lagioia, *Study: The impact of the General Data Protection Regulation on artificial intelligence*, European Parliament, 2020; G. Sartor et al., *Study: Regulating targeted and behavioural advertising in digital services How to ensure users' informed consent*, European parliament, 2021.

[86]H. R. Varian, «Computer Mediated Transactions», in *American Economic Review:* vol. C (2010), p. 1-10.

[87]R. Kurzweil, *How to Create a Mind*, Viking, 2012. *Testo originale*: "Through [information] technologies we can address the grand challenges of humanity, such as maintaining a healthy

environment, providing the resources for a growing population (including energy, food, and water), overcoming disease, vastly extending human longevity, and eliminating poverty. It is only by extending ourselves with intelligent technology that we can deal with the scale of complexity needed."

[88]J. C. R. Licklider, «Man-Computer Symbiosis», in *IRE Transactions on Human Factors in Electronics*, vol. HFE-1, n. March (1960), p. 4-11. *Testo originale*: "making decisions and controlling complex situations without inflexible dependence on predetermined programs."

[89]E. Brynjolfsson e A. McAfee, *Race Against the Machine*, Digital Frontier Press, 2011.

[90]R. M. Ryan e E. L. Deci, *Self-Determination Theory Basic Psychological Needs in Motivation, Development, and Wellness*, Guilford, 2017.

[91]A. McAfee e E. Brynjolfsson, *Machine, Platform, Crowd*, Norton, 2019.

[92]Su problemi e prospettive del rapporto tra lavoro e IA, vedi Y. Harayama et al., «Artificial Intelligence and the Future of Work», in *Reflections on Artificial Intelligence for Humanity*, a cura di B. Braunschweig e M. Ghallab, Springer, 2021, p. 53-67.

[93]Sui problemi giuridici della pubblicità mirata, vedi G. Sartor et al., *Study: Regulating targeted and behavioural advertising in digital services How to ensure users' informed consent*, European parliament, 2021.

[94]S. F. Janka e S. Uhsler, «Antitrust 4.0-the rise of Artificial Intelligence and emerging challenges to antitrust law», in *European Competition Law Review*, vol. XXXIX, n. 3 (2018), p. 112-123; E. Calvano et al., «Artificial intelligence, algorithmic pricing and collusion», in *American Economic Review*, vol. CX (2020), p. 3267-97.

[95]E. Pariser, *The Filter Bubble*, Penguin, 2011.

[96]C. R. Sunstein, *Republic: divided democracy in the age of social media*, Princeton University Press, 2020.

[97]C. Burr e N. Cristianini, «Can Machines Read our Minds?», in *Minds and Machines*, vol. XXIX (2019), p. 461-494.

[98]F. Bosco et al., «Profiling Technologies and Fundamental Rights and Values: Regulatory Challenges and Perspectives from European Data Protection Authorities», in *Reforming European Data Protection Law*, a cura di S. Gutwirth et al., Springer, 2015. *Testo originale*: "Profiling is a technique of (partly) automated processing of personal and/or non-personal data, aimed at producing knowledge by inferring correlations from data in the form of profiles that can subsequently be applied as a basis for decision-making. A profile is a set of correlated data that represents a (individual or collective) subject. Constructing profiles is the process of discovering unknown patterns between data in large data sets that can be used to create profiles. Applying profiles is the process of identifying and representing a specific individual or group as fitting a profile and of taking some form of decision based on this identification and representation." See also M. Hildebrandt, «Profiling and AML», in *The Future of Identity in the Information Society. Challenges and Opportunities*, a cura di K. Rannenberg et al., Springer, 2009.

[99]Tratta da F. Lagioia et al., «Fairness through Group Parities? The Case of COMPAS-SAPMOC», in *AI and Society* (2022).

[100]D. Kahneman et al., *Noise: A Flaw in Human Judgement*, Collins, 2021.

[101]J. Kleinberg et al., «Human Decisions and Machine Predictions», in *The Quarterly Journal of Economics*, vol. CXXXIII (2017), p. 237-93.

[102]C. O'Neil, *Weapons of math destruction: how big data increases inequality and threatens democracy*, Crown Business, 2016, p. 16. *Testo originale*: "An algorithm processes a slew of

statistics and comes up with a probability that a certain person might be a bad hire, a risky borrower, a terrorist, or a miserable teacher. That probability is distilled into a score, which can turn someone's life upside down. And yet when the person fights back, 'suggestive' countervailing evidence simply won't cut it. The case must be ironclad. The human victims of WMDs [...] are held to a far higher standard of evidence than the algorithms themselves."

[103]J. Kleinberg et al., «Discrimination in the Age of Algorithm», in *Journal of Legal Analysis*, vol. X (2018), p. 113-174, p. 113.

[104]Varian, «Computer Mediated Transactions» cit.

[105]Sui profili giuridici dei registri distribuite e delle blockchain, vedi P. De Filippi e A. Wright, *Blockchain and the law: The rule of code*, Harvard University Press, 2018, Sul rapporto tra linguaggi per gli smart contract e linguaggi per l'intelligenza artificiale: G. Governatori et al., «On legal contracts, imperative and declarative smart contracts, and blockchain systems», in *Artificial Intelligence and Law*, vol. XXVI (2018), p. 377-409. Per un recente tentativo di sviluppare un linguaggio per gli smart contract: S. Crafa et al., «Stipula: a domain specific language for legal contracts», in *Prolala - Programming Languages and the Law*, 2022.

[106]A. Pentland, *Social Physics: How Social Networks Can Make Us Smarter*, Penguin, 2015, p. 28.

[107]S. Zuboff, *The Age of Surveillance Capitalism*, Hachette, 2019.

[108]J. M. Balkin, «The Constitution in the National Surveillance State», in *Minnesota Law Review*, vol. XCIII (2008), p. 1-25.

[109]K. Polanyi, *The Great Transfomation*, First ed. 1944, Beacon Press, [1944] 2001.

[110]S. Zuboff, *The Age of Surveillance Capitalism*, Hachette, 2019, p. 514. *Testo originale*: "Surveillance capitalism annexes human experience to the market dynamic so that it is reborn as behavior: the fourth "fictional commodity'. Polanyi's first three fictional commodities — land, labor, and money— were subjected to law. Although these laws have been imperfect, the institutions of labor law, environmental law, and banking law are regulatory frameworks intended to defend society (and nature, life, and exchange) from the worst excesses of raw capitalism's destructive power. Surveillance capitalism's expropriation of human experience has faced no such impediments".

[111]N. Cristianini e T. Scantamburlo, «On Social Machines for Algorithmic Regulation», in *AI and Society* (2019).

[112]Balkin, «The Constitution in the National Surveillance State» cit., p. 3.

[113]Sui problemi e le opportunità dello Stato dell'informazione, chiamato anche "Stato algoritmico: H.-W. Micklitz et al., *Constitutional Challenges in the Algorithmic Society*, Cambridge University Press, 2021.

[114]Pariser, *The Filter Bubble* cit.

[115]Sunstein, *Republic: divided democracy in the age of social media* cit.

[116]G. Ruffo e M. Tambuscio, «Capire la diffusione della disinformazione e come contrastarla», in *Federalismi*, vol. XI (2020), p. 73-84; M. Cavino, «Il triceratopo di Spielberg. Fake news, diritto e politica», in *Federalismi*, vol. XI (2020), p. 32-42; F. Pizzetti, «Fake news e allarme sociale: responsabilità, non censura», in *Medialaws* (2017), p. 48-59.

[117]C. R. Sunstein, *On rumors: how falsehoods spread, why we believe them, what can be done*, Farrar, Straus e Giroux, 2009, p. 24-25. *Testo originale*: "Whenever a threat looms or a terrible event has occurred, rumors are inevitable. [...] In the aftermath of a crisis, numerous speculations will be offered. To some people, those speculations will seem plausible, perhaps

because they provide a suitable outlet for outrage and blame. Terrible events produce outrage, and when people are outraged, they are all the more likely to accept rumors that justify their emotional states, and also to attribute those events to intentional action."

[118]Sulla disinformazione, vedi: G. Pitruzzella e O. Pollicino, *Disinformation and hate speech. A European Constitutional Perspective*, Bocconi University Press, 2020.

[119]A. Hern, «Cambridge Analytica: how did it turn clicks into votes», in *Guardian* (6 maggio 2018). Sulla relazione tra populismo, democrazia e Internet, vedi M. Barberis, *Come internet sta uccidendo la democrazia*, Chiarelettere, 2020.

[120]Hern, «Cambridge Analytica: how did it turn clicks into votes» cit.

[121]Sui profili giuridici del credito sociale, vedi: E. Consiglio e G. Sartor, «Il sistema di credito sociale cinese: una "nuova" regolazione sociotecnica mediante sorveglianza, valutazione e sanzione», in *TIGOR* (2021), p. 138-61; D. Mac Sithigh e M. Siems, «The Chinese Social Credit System: A Model for Other Countries?», in *Modern Law Review*, vol. LXXXII (2019), p. 1034-1071.

[122]Il concetto di "dividuo" fu originariamente proposto da G. Deleuze, «Postscriptum sur les sociétés de contrôle par Gilles Deleuze», in *L'Autre Jurnal* (1990).

[123]Abusando della prospettiva elaborata da M. Walzer, *Spheres of Justice*, Basic Books, 1983.

[124]L. Floridi e M. Taddeo, «What is data ethics?», in *Philosophical Transactions of the Royal Society A* (2016), p. 374: *Testo originale*: " "branch of ethics that studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values)."

[125]Per un'introduzione a vari aspetti dell'etica dell'IA; vedi F. Fossa et al., *Automi e persone*, Carocci, 2021.

[126]Per un'ampia introduzione e discussione, vedi L. Floridi, *Etica dell'intelligenza artificiale*, Cortina, 2021.

[127]L. Floridi et al., «AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations», in *Minds and Machines*, vol. XXVIII (2018), p. 689-707.

[128]AI-HLEG, High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission, 2019; per la presentazione di diversi approcchi alla *trustworthy AI*, vedi F. Heintz et al. (a cura di), *Trustworthy AI - Integrating Learning, Optimization and Reasoning - First International Workshop, TAILOR 2020, Virtual Event, September 4-5, 2020, Revised Selected Papers*, Springer, 2021.

[129]Floridi et al., «AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations» cit.

[130]F. Pasquale, *The black box society: the secret algorithms that control money and information*, Harvard University Press, 2015.

[131]Per una rassegna dei metodi per spiegare il funzionamento di sistemi che si presentano come scatole nere, si veda R. Guidotti et al., «A Survey Of Methods For Explaining Black Box Models», in *ACM Computing Surveys*, vol. LI, n. Article 93 (2018), p. 1-42.

[132]J. M. Balkin, «The Three Laws of Robotics in the Age of Big Data», in *Ohio State Journal Law Journal*, vol. LXXVIII (2017), p. 1217-241.

[133]Pariser, *The Filter Bubble* cit.

[134]Sul rapporto tra diritti fondamentali e tecnologie dell'informazione, vedi G. Sartor, «Human Rights and Information Technologies», in *The Oxford Handbook on the Law and Regulation of Technology*, a cura di R. Brownsword et al., Oxford University Press, 2017, p. 424-450; specificamente su diritti e IA, vedi G. Sartor, «Artificial intelligence and human rights: Between law and ethics», in *Maastricht Journal of European and Comparative Law*, vol. XXVII, n. 27 (2020), p. 705-719.

[135]Dati i limiti della nostra analisi non potranno essere presi in considerazione numerosi temi e approfondimenti. Vedi su responsabilità, diritti, e altri aspetti della disciplina dell'IA: U. Ruffolo (a cura di), *Intelligenza Artificiale e Responsabilità*, Giuffrè, 2018; U. Ruffolo (a cura di), *Intelligenza artificiale.Il diritto, i diritti, l'etica*, Giuffrè, 2020; U. Ruffolo, *XXVI lezioni di diritto dell'intelligenza artificiale*, Giappichelli, 2021; A. D'Aloia, *Intelligenza artificiale e diritto. Come regolare un mondo nuovo*, Franco Angeli, 2021. Sul rapporto tra algoritmi e regolazione; A. Reichman e G. Sartor, «Algorithms and Regulation», in *Constitutional Challenges in the Algorithmic Society*, Cambridge University Press, 2021, p. 131-81.

[136]Sul rapporto tra protezione dei dati e IA: G. Sartor e F. Lagioia, *Study: The impact of the General Data Protection Regulation on artificial intelligence*, European Parliament, 2020; F. Pizzetti (a cura di), *Intelligenza artificiale, protezione dei dati personali e regolazione*, Giappichelli, 2018.

[137]Per un commento all'Art. 22, vedi, F. Lagioia et al., «Art. 22. Processo decisionale automatizzato relativo alle persone fisiche, compresa la profilazione», in *Codice della privacy e Data Protection*, Giuffrè, 2021, p. 378-390.

[138]Proposta di Regolamento del Parlamento Europeo e del Consiglio che stabilisce regole armonizzate sull'IA (Legge sull'IA), COM(2021) 206 final 2021/0106(COD).

[139]Un'iniziativa analoga proviene dalla California, che nel 2019 ha adottato il *Bot Disclosure and Accountability Act*, che vieta l'uso di *bot* (agenti software) allo scopo di ingannare e fuorviare gli utenti di piattaforme e siti internet. In particolare, la sezione 17941(a) rende illegale l'uso di *bot* per comunicare o interagire online con individui con l'intento di indurli in errore circa la natura artificiale del *bot*.

[140]Proposta di Direttiva del Parlamento Europeo e del Consiglio relativa all'adeguamento delle norme in materia di responsabilità civile extracontrattuale all'intelligenza artificiale (Direttiva sulla responsabilità da intelligenza artificiale), COM(2021) 206 final 2021/0106(COD).

[141]European Commission, Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Action Plan against Disinformation JOIN(2018), consultabile al sito: `https://ec.europa.eu/digital-single-market/en/news/action-plan-against-disinformation`.

[142]Pitruzzella e Pollicino, *Disinformation and hate speech. A European Constitutional Perspective* cit.

[143]F. Galli et al., «Potenzialità e limiti della moderazione algoritmica», in *Ecoscienza*, vol. XIII, n. 3 (2022), p. 41-43.

[144]Sul concetto di autonomia, applicato ai robot, vedi G. Sartor e A. Omicini, «The Autonomy of Technological Systems and Responsibilities for their Use», in *Autonomous Weapons Systems: Law, Ethics, Policy*, a cura di N. Bhuta et al., Cambridge University Press, 2016, p. 39-74.

[145]Sulla disciplina dei veicoli autonomi, vedi E. Al Mureden e G. Calabresi, *Driverless cars. Intelligenza artificiale e futuro della mobilità*, Il Mulino, 2021. Sul controverso tema delle scelte "etiche" affidate a veicoli autonomi, vedi G. Contissa et al., «The Ethical Knob», in *Artificial intelligence and Law*, vol. XXV (2017), p. 365-78. Sui robots medici e l'IA in medicina: F. Lagioia, *L'intelligenza artificiale in sanità: un'analisi giuridica*, Giappichelli, 2020.

[146]Vedi U. Pagallo, *The laws of robots*, Springer, 2013 and M. R. Calo, «Robotics and the Lessons of Cyberlaw», in *California Law Review*, vol. CIII (2015), p. 101-48.

[147]Per una teoria generale della responsabilità per danni causati da sistemi intelligenti, vedi A. Beckers e G. Teubner, *Three Liability Regimes for Artificial Intelligence: Algorithmic Actants, Hybrids, Crowds*, Hart, 2019.

[148]G. Hallevy, *When Robots Kill: artificial intelligence under criminal law*, Northeastern University Press, 2013; Lagioia e Sartor, «AI systems under criminal law: A legal analysis and a regulatory perspective» cit.

[149]Pagallo, *The laws of robots* cit.

[150]S. Chopra e L. F. White, *A Legal Theory for Autonomous Artificial Agents*, University of Michigan Press, 2011.

[151]Per uno sviluppo di questa tesi, rinvio a G. Sartor, «Cognitive Automata and the Law: Electronic Contracting and the Intentionality of Software Agents», in *Artificial Intelligence and Law*, vol. XVII (2009), p. 253-90.

[152]M. R. Calo, «Robots and Privacy», in *Robot Ethics: The Ethical and Social Implications of Robotics*, Cambridge University Press, 2012, p. 187-201.

[153]Come il *peculium* di cui disponevano gli schiavi nel diritto romano, per le proprie spese personali.

[154]C. Novelli et al., «A conceptual framework for legal personality and its application to AI», in *Jurisprudence* (2021).

[155]Il c.d. *jus in bello*, in opposizione allo *jus ad bellum,* che determina le condizioni nelle quali l'uso della forza armata è lecito (ad esempio, per respingere un'aggressione).

[156]Sul concetto di autonomia e la sua applicazione alle armi, vedi G. Sartor e A. Omicini, «The Autonomy of Technological Systems and Responsibilities for their Use», in *Autonomous Weapons Systems: Law, Ethics, Policy*, a cura di N. Bhuta et al., Cambridge University Press, 2016, p. 39-74.

[157]Rinvio il lettore interessato alla fantascienza al racconto di Philip K. Dick *Second Variety* (da cui fu tratto il film *Screamers*) nel quale la realizzazione di armi-robot intelligenti, la cui costruzione è affidata a fabbriche completamente automatizzate, conduce alla distruzione dell'umanità (P. K. Dick, «Second Variety», in *Minority Report*, Gollancz, 2002, p. 61-101).

[158]R. Arkin, *Governing Lethal Behavior in Autonomous Robots*, CRC Press, 2009.

[159]J. K. Galbraith, *American Capitalism: The Concept of Countervailing Power*, Houghton Mifflin, 1956.

[160]Sulle tecnologie dell'intelligenza artificiale per il diritto, vedi K. D. Ashley, *Artificial Intelligence and Legal Analytics*, Cambridge University Press, 2017. Per una discussione con riferimento al nostro paese: A. Santosuosso, *Intelligenza artificiale e diritto. Perché le tecnologie di IA sono una grande opportunità per il diritto*, Mondadori, 2020; Bassoli, *Algoritmica giuridica. Intelligenza artificiale e diritto*, Amon, 2021. Sull'uso dell'IA nella determinazione degli assegni di mantenimento: E. Al Mureden e R. Rovatti (a cura di), *Gli assegni di mantenimento tra disciplina legale e intelligenza artificiale*, Giappichelli, 2020. Per un'analisi delle

prime applicazioni di intelligenza artificiale e diritto, vedi: G. Sartor, *Le applicazioni giuridiche dell'intelligenza artificiale: la rappresentazione della conoscenza*, Giuffrè, 1990.

[161]D. Waterman e M. Peterson, «Rule-Based Models of Legal Expertise», in *Proceeding National Conference on Artificial Intelligence - AAAI*, 1980, p. 272-5.

[162]Il modello e la metodologia del progetto sviluppato dall'Imperial College sono illustrati in M. J. Sergot et al., «The British Nationality Act as a Logic Program», in *Communications of the ACM*, vol. XXIX (1986), p. 370-86.

[163]M. Palmirani et al., «LegalRuleML: XML-Based Rules and Norms», in *Rule-Based Modeling and Computing on the Semantic Web*, a cura di F. Olken et al., Springer Berlin Heidelberg, Berlin, Heidelberg 2011, p. 298-312; G. Governatori et al., «Logic and the Law: Philosophical Foundations, Deontics, and Defeasible Reasoning», in *Handbook of Deontic Logic and Normative Systems, Volume 2*, College Publications, 2022, p. 657-764.

[164]Kowalski e Datoo, «Logical English meets legal English for swaps and derivatives» cit.

[165]Per la prospettiva ottimista degli anni '80 sui sistemi esperti giuridici, vedi R. Susskind, *Expert Systems in Law: A Jurisprudential Inquiry*, Oxford University Press, 1988.

[166]Come affermano P. Johnson e G. Masri, «Making Better Determinations, Discussion Paper No.7», in *Future Challenges for E-government*, Australian Government, Department of Finance e Deregulation, 2004. Gli stessi autori individuano tre criteri principali per valutare la funzionalità dei processi determinativi: il costo del processo determinativo (costo da determinarsi tenendo conto sia gli oneri di cui deve farsi carico il richiedente, sia quelli di cui si fa carico la PA), la sua qualità (nei termini dell'accuratezza giuridica, consistenza, equità, tempestività, trasparenza), la sua controllabilità (cioè l'adeguatezza della motivazione fornita al cittadino, al fine di verificare la correttezza e l'imparzialità del processo determinativo). Secondo gli stessi autori, i risultati ottenuti provano che i sistemi basati sulla conoscenza consentono di ottenere risultati notevoli sotto i seguenti profili: riduzione del personale, aumenti di produttività, riduzione dei costi di gestione, aumento della coerenza delle decisioni, riduzione dei ricorsi contro gli atti amministrativi, aumento della qualità del lavoro (grazie all'aiuto del sistema, anche il personale esecutivo può affrontare con autonomia casi complessi, che prima potevano essere risolti solo dai dirigenti). In particolare, l'uso di tali sistemi ha consentito di porre rimedio a frequenti disguidi ed errori, evitando per esempio che i destinatari ricevessero prestazioni superiori o inferiori a quelle dovute. Inoltre, grazie alle indicazioni fornite dai sistemi basati sulla conoscenza si è esteso l'accesso alle prestazioni pubbliche, che molti degli aventi diritto non richiedevano mancando di informazioni adeguate e precise.

[167]C. Montesquieu, *L'esprit des lois*, 1748, Livre 11, Chap. 6. disponibile presso il sito `http://www.uqac.uquebec.ca/zone30/Classiques_des_sciences_sociales`. *Testo originale*: "Les juges de la nation ne sont [...] que la bouche qui prononce les paroles de la loi; des êtres inanimés qui n'en peuvent modérer ni la force ni la rigueur."

[168]Sull'interpretazione di regole e precedenti, in una prospettiva comparatistica, vedi D. N. MacCormick e R. S. Summers (a cura di), *Interpreting Statutes: A Comparative Study*, Darthmouth, 1991, e D. N. MacCormick e R. S. Summers (a cura di), *Interpreting Precedents: A Comparative Study*, Darthmouth, 1997.

[169]Tra le opere che considerano queste dimensioni del diritto, vedi, tra altri, R. Alexy, *Theorie der Grundrechte*, Suhrkamp, 1985, e R. M. Dworkin, *Law's Empire*, Kermode, 1986.

[170]Per una critica dei tentativi di formalizzare il diritto mediante regole, vedi P. Leith, «Clear Rules and Legal Expert Systems», in *Automated Analysis of Legal Texts*, a cura di A. Martino e

F. Socci, North Holland, 1986, p. 661-79. Vedi anche L. T. McCarty, «Artificial Intelligence and Law: How to Get There from Here», in *Ratio Juris*, vol. III (1990), p. 189-200.

[171]Lo stesso ragionamento basato su regole, d'altro canto, ha conosciuto importanti sviluppi nell'ambito delle ricerche di intelligenza artificiale e diritto. Si è arricchito della possibilità di rappresentare con precisione i concetti normativi, sia i concetti deontici di base (obbligo, permesso, facoltà), sia i concetti correlati all'idea di obbligo verso una persona (obbligazioni, diritti all'adempimento), sia i concetti potestativi (potere di modificare posizioni giuridiche, soggezione a tale potere, capacità negoziale, ecc.). Per la rappresentazione degli obblighi verso altre persone, vedi H. Herrestad e C. Krogh, «Obligations Directed from Bearers to Counterparties», in *Proceedings of the 5th International Conference on Artificial Intelligence and Law (ICAIL)*, ACM, 1995, p. 210-8. Sui concetti potestativi, vedi A. J. Jones e M. J. Sergot, «A Formal Characterisation of Institutionalised Power», in *Journal of the IGPL*, vol. IV (1996), p. 429-45. Per un modello che include concetti deontici, obblighi verso altre persone, e concetti potestativi, vedi G. Sartor, «Fundamental Legal Concepts: A Formal and Teleological Characterisation», in *Artificial Intelligence and Law*, vol. XXI (2006), p. 101-42. Inoltre, sono stati sviluppati modelli intesi a cogliere la dinamica dei sistemi di norme (cioè le modifiche di ordinamenti normativi in seguito ad abrogazioni e annullamenti), e gli impatti di tale dinamica sulle norme applicabili e sulle posizioni giuridiche conseguenti. Tra gli studi al riguardo, vedi C. E. Alchourrón e D. Makinson, «Hierarchies of Regulations and Their Logic», in *New Studies on Deontic Logic*, a cura di R. Hilpinen, Reidel, 1981, p. 123-48, R. Hernandez Marin e G. Sartor, «Time and Norms: A Formalisation in the Event-calculus», in *Proceedings of the Seventh International Conference on Artificial Intelligence and Law (ICAIL)*, ACM, 1999, p. 90-100, e più recentemente G. Governatori e A. Rotolo, «Changing legal systems: legal abrogations and annulments in defeasible logic», in *Logic Journal of IGPL*, vol. XVIII (2010), p. 157-94.

[172]Non tutti i procedimenti cognitivi sono riducibili alla logica e neppure al ragionamento, inteso in generale come il procedimento secondo il quale la mente passa da premesse a conclusioni in conformità a schemi generali di inferenza. Diverse discipline (psicologia, neurologia, intelligenza artificiale, ecc.) concordano nel sostenere che il ragionamento ha un ruolo limitato in numerosi processi cognitivi: nella percezione (che ha ruolo basilare nell'accertamento dei fatti), nella memoria, nell'analogia, nella formazione di ipotesi. Nel campo della morale, e anche del diritto, il ragionamento non sostituisce la spontanea empatia che ogni essere umano prova nei confronti dei propri simili né la comprensione intuitiva che otteniamo proiettando noi stessi nelle condizioni altrui. Pertanto, nessuna "logica" (quale tecnica del ragionamento) può pretendere di fornire un modello completo dell'attività del giurista, dell'intero processo mediante il quale vengono affrontati e risolti problemi giuridici. Tuttavia, la logica, se sufficientemente estesa, può ricoprire una porzione significativa di tale processo, così da rappresentare un aspetto importante della metodologia del giurista, e da fornire la base per lo sviluppo di software avanzati per il supporto alla decisione giuridica. Per una considerazione approfondita del ruolo della logica nel ragionamento giuridico, mi permetto di rinviare a G. Sartor, *Legal Reasoning: A Cognitive Approach to the Law*, a cura di E. Pattaro, Springer, 2005, p. 1-845, Parte II e a G. Governatori et al., «Logic and the Law: Philosophical Foundations, Deontics, and Defeasible Reasoning», in *Handbook of Deontic Logic and Normative Systems, Volume 2*, College Publications, 2022, p. 657-764.

[173]Sulla motivazione delle sentenze, vedi: M. Taruffo, *La motivazione della sentenza civile*, Cedam, 1975, e T. Mazzarese, *Forme di razionalità delle decisioni giudiziali*, Giappichelli, 1997.

[174]S. Toulmin, *The Uses of Argument*, (1st ed. 1958.), Cambridge University Press, [1958] 2003, p. 94-127.

[175] Se tutti ingredienti per formare gli argomenti (fatti e regole) fossero espressi in forma logica, argomenti siffatti potrebbe essere costruito automaticamente, così come automaticamente sarebbe possibile verificare i rapporti logici tra argomenti e determinare quali argomenti si possano considerare giustificati in base all'informazione disponibile (vedi H. Prakken e G. Sartor, «Argument-based Extended Logic Programming with Defeasible Priorities», in *Journal of Applied Non-classical Logics*, vol. VII (1997), p. 25-75). Per ragioni di semplicità, e poiché si vogliono analizzare i rapporti tra argomenti (piuttosto che la struttura interna degli stessi), qui ci si limita a rappresentare gli argomenti nel linguaggio naturale, illustrandone graficamente le connessioni.

[176]Pollock, *Cognitive Carpentry: A Blueprint for How to Build a Person* cit., p. 38.

[177]Ivi, p. 39.

[178]A questa tecnica argomentativa alludeva probabilmente John Stuart Mill [1806-73] quando osservava che il ragionamento nelle materie pratiche richiede, accanto alla logica positiva (intesa a fondare una conclusione producendo ragioni a suo sostegno), anche una *logica negativa*. Quest'ultima "individua debolezze in teorie o errori in pratica, senza stabilire verità positive" (J. S. Mill, *On Liberty*, (1st ed. 1859.), Penguin, 1974, p. 106). *Testo originale*: "points out weaknesses in theory or errors in practice, without establishing positive truths".

[179]Per citare ancora John Stuart Mill, "in ogni materia nella quale sia possibile una differenza di opinioni, la verità dipende da un bilanciamento delle ragioni in conflitto" (J. S. Mill, *On Liberty*, (1st ed. 1859.), Penguin, 1974, p. 109). *Testo originale*: "on every subject on which difference of opinion is possible, the truth depends on a balance to be struck between two sets of conflicting reasons".

[180]Negli ultimi anni sono state realizzate numerose logiche delle relazioni tra argomenti. Per una rassegna e valutazione critica, vedi H. Prakken e G. Sartor, «Law and logic: A review from an argumentation perspective», in *Artificial Intelligence*, vol. CCXXVII (2015), p. 214-45.

[181]Sulla logica dell'argomentazione formale, il riferimento fondamentale è P. M. Dung, «On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and *n*–Person Games», in *Artificial Intelligence*, vol. LXXVII (1995), p. 321-57. Per una descrizione dettagliata dei diversi schemi argomentativi, vedi D. N. Walton et al., *Argumentation Schemes*, Cambridge University Press, 2008. Per una rassegna degli studi sull'argomentazione giuridica mediante tecniche formali, vedi H. Prakken e G. Sartor, «Law and logic: A review from an argumentation perspective», in *Artificial Intelligence*, vol. CCXXVII (2015), p. 214-45.

[182]Su questo tema, vedi T. F. Gordon e N. Karacapilidis, «The Zeno Argumentation Framework», in *Proceedings of the Sixt International Conference on Artificial Intelligence and Law (ICAIL)*, ACM, 1997, p. 10-9; T. F. Gordon e D. N. Walton, «The Carneades argumentation framework - using presumptions and exceptions to model critical questions», in *Computational Models of Argument. Proceedings of COMMA-06*, a cura di P. Dunne e T. Bench-Capon, IOS Press, 2006, p. 195-207.

[183]K. D. Ashley, «Reasoning with cases and hypotheticals in HYPO», in *International Journal of Man-Machine Studies*, vol. XXXIV (1991), p. 753-96; T. J. M. Bench-Capon, «HYPO'S legacy: introduction to the virtual special issue», in *Artificial Intelligence and Law*, vol. XXV (2017), p. 205-250.

[184]V. Aleven e K. D. Ashley, «Evaluating a learning environment for case-based argumentation skills», in *Proceedings of the sixth international conference on artificial intelligence and law (ICAIL-97)*, ACM, 1997, p. 170-179.

[185]Ivi.

[186]H. Prakken e G. Sartor, «Modelling Reasoning with Precedents in a Formal Dialogue Game», in *Artificial Intelligence and Law*, vol. VI (1998), p. 231-87.

[187]J. F. Horty, «Rules and reasons in the theory of precedent», in *Legal theory*, vol. X (2011), p. 1-33; J. F. Horty, «Reasoning with Dimensions and Magnitudes», in *International Conference on Artificial Intelligence and Law, ICAIL2017*, ACM, 2017.

[188]Horty, «Reasoning with Dimensions and Magnitudes» cit.; T. Bench-Capon e K. Atkinson, *The Roles of Dimensions and Values in Legal CBR*, 2018.

[189]Non mancarono alcuni tentativi di applicare reti neurali a problemi giuridici, come i seguenti: L. Philipps, «Distribution of damages in car accidents through the use of neural networks», in *Cardozo L. Rev.*, vol. XIII (1991), p. 987-1000; F. Romeo e F. Barbarossa, «Simulation of Verdicts in Civil Liability», in *World Congress on Neural Networks, WCNN, San Diego California USA, June 5-9 1994*, Hillsdale New Jersey USA, 1994, vol. I, p. 432-436, J. Zeleznikow e A. Stranieri, «The Split-Up System: Integrating Neural Networks and Rule Based Reasoning in the Legal Domain», in *Proceedings of the Fifth International Conference on Artificial Intelligence and Law (ICAIL)*, ACM, 1995, p. 185-94. Per una discussione dell'uso delle reti neurali nel diritto: F. Romeo, *Lezioni di logica e di informatica giuridica*, Giappichelli, 2012.

[190]A questo riguardo rinviamo a G. Contissa et al., *CLAUDETTE meets GDPR Automating the Evaluation of Privacy Policies using Artificial Intelligence. Study Report, Funded by The European Consumer Organisation*, BEUC, 2018.

[191]Nel nostro caso, l'algoritmo di apprendimento potrebbe operare assegnando ciascuna parola un peso (un valore numerico) tale da far sì che il valore delle clausole legittime sia il più lontano possibile lontano dal valore delle clausole vessatorie. Per esempio, si assuma che presenza di ogni parola sia rappresentata dal numero 1, e la sua assenza da -1. Il valore di una frase (da utilizzare per calcolare la sua distanza da altre frasi) potrebbe determinarsi sommando i valori che si ottengono moltiplicando il numero 1 o -1 assegnato a ciascuna parola (a seconda che essa sia presente o assente) per il peso di quella parola.

[192]Più esattamente gli esperimenti sul corpus giuridico annotato sono stati condotti mediante la procedura *leave-one-out* (lasciane-uno-fuori). Il *dataset* costituito dai documenti marcati dai giuristi esperti è stato utilizzato nel modo seguente: a turno, ciascun documento del dataset è stato escluso e impiegato per verificate la risposta del classificatore, addestrato con riferimento all'insieme dei documenti rimanenti. La prestazione complessiva del classificatore è stata valutata con riferimento all'insieme delle risposte così ottenute.

[193]F. Ruggeri et al., «Detecting and explaining unfairness in consumer contracts through memory networks», in *Artificial Intelligence and Law*, vol. XXX, n. 1 (2022), p. 59-92.

[194]Varie tecniche per l'elaborazione del linguaggio naturale possono essere utilizzate a questi fini, in particolare, i vocaboli possono essere rappresentati oltre che mediante la loro appartenenza o meno alla frase anche mediante l'indicazione della frequenza con la quale appaiono assieme ad altri vocaboli. Per denotare in questo caso si parla di "inserimento di parola" (*word embedding*). Per una rassegna e riferimenti bibliografici, vedi K. D. Ashley, «A Brief History of the Changing Roles of Case Prediction in AI and Law», in *Law in Context*, vol. XXXVI (2019), p. 93-112. Per una comparazione di diversi modelli per la predizione nel diritto, vedi

L. K. Branting et al., «Scalable and explainable legal prediction», in *Artificial Intelligence and Law*, vol. XXIX (2021), p. 213-238.

[195]I. C. Sabo et al., «Clustering of Brazilian legal judgments about failures in air transport service: an evaluation of different approaches», in *Artificial Intelligence* (2021), p. 1-37.

[196]Vedi, anche per riferimenti bibliografici: A. Santosuosso e G. Sartor, *La giustizia predittiva: una visione realistica*, Giurisprudenza italiana, 2022; C. Castelli e D. Piana, «Giustizia predittiva. La qualità della giustizia in due tempi», in *Questione Giustizia*, n. 4 (2018), p. 153-65.

[197]Per una rassegna sulla costruzione automatica di sommari di sentenze, vedi J. Deepali et al., «Summarization of legal documents: Where are we now and the way forward», in *Computer Science Review* (2021), p. 1-14. Sul ruolo delle massime, vedi E. Vincenti, «Massimazione e conoscenza della giurisprudenza nell'era digitale», in *Questione Giustizia*, n. 4 (2018), p. 147-152.

[198]Il sistema Compas, usato in talune giurisdizioni degli Stati Uniti è stato oggetto di un intenso dibattito (J. Larson et al., *How We Analyzeded the COMPAS Recidivism Algorithm*, Pro Publica, 2016). Per una discussione e riferimenti bibliografici, vedi F. Lagioia et al., «Fairness through Group Parities? The Case of COMPAS-SAPMOC», in *AI and Society* (2022).

[199]N. Aletras et al., «Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective», in *Peer J Computer Science*, vol. 2: e93 (2016).

[200]D. M. Katz et al., «A General Approach for Predicting the Behavior of the Supreme Court of the United States», in *PLoS ONE*, vol. XII (2017).

[201]M. Surdeanu et al., «Risk analysis for intellectual property litigation», in *ICAIL-2011*, ACM, 2011, p. 116-20.

[202]Per una discussione del rumore, si veda D. Kahneman et al., *Noise: A Flaw in Human Judgement*, Collins, 2021.

[203]Così U. Ruffolo, «La machina sapiens come "avvocato generale" ed il primato del giudice umano: una interazione virtuosa», in *Lezioni di diritto dell'intelligenza artificiale*, Giappichelli, 2021. Vedi anche: A. Carleo (a cura di), *Decisione robotica*, Il Mulino, 2019; B. Caravita, «Uno staff di laureati per costruire una macchina di intelligenza artificiale che collabori alle funzioni di giustizia», in *Federalismi* (2021), p. 1-8.

[204]M. Palmirani et al., *Legal Drafting in the Era of Artificial Intelligence and Digitisation*, European Commission, 2022; G. Sartor, *The way forward for better regulation in the EU – better focus, synergies, data and technology*, European Parliament, 2022.

[205]D. F. Engstrom et al., *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*, Administrative Conference of the United States, 2020; H. Margetts, «Rethinking AI for Good Governance», in *Dedalus*, vol. CLI (2022), p. 360-71.

Bibliografia

Al Mureden, E. e Calabresi, G., *Driverless cars. Intelligenza artificiale e futuro della mobilità*, Il Mulino, 2021.

Al Mureden, E. e Rovatti, R. (a cura di), *Gli assegni di mantenimento tra disciplina legale e intelligenza artificiale*, Giappichelli, 2020.

Alchourrón, C. E. e Makinson, D., «Hierarchies of Regulations and Their Logic», in *New Studies on Deontic Logic*, a cura di R. Hilpinen, Reidel, 1981, p. 123-48.

Aletras, N., Tsarapatsanis, D., Preoţiuc Pietro, D. e Lampos, V., «Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective», in *Peer J Computer Science*, vol. 2: e93 (2016).

Aleven, V. e Ashley, K. D., «Evaluating a learning environment for case-based argumentation skills», in *Proceedings of the sixth international conference on artificial intelligence and law (ICAIL-97)*, ACM, 1997, p. 170-179.

Alexy, R., *Theorie der Grundrechte*, Suhrkamp, 1985.

Allen, L. E., «Towards a Normalized Language to Clarify the Structure of Legal Discourse», in *Deontic Logic, Computational Linguistics, and Legal Informations Systems*, a cura di A. A. Martino, North Holland, 1982, p. 349-407.

Aristotele, *Opere*, vol. VI: *Metafisica*, Laterza, 1996.

Arkin, R., *Governing Lethal Behavior in Autonomous Robots*, CRC Press, 2009.

Ashley, K. D., «A Brief History of the Changing Roles of Case Prediction in AI and Law», in *Law in Context*, vol. XXXVI (2019), p. 93-112.

— *Artificial Intelligence and Legal Analytics*, Cambridge University Press, 2017.

— «Reasoning with cases and hypotheticals in HYPO», in *International Journal of Man-Machine Studies*, vol. XXXIV (1991), p. 753-96.

Asimov, I., *I, Robot*, (Prima edizione 1950), Collins, 1968.

— *Robots and Empire*, Collins, 1985.

Balkin, J. M., «The Constitution in the National Surveillance State», in *Minnesota Law Review*, vol. XCIII (2008), p. 1-25.

— «The Three Laws of Robotics in the Age of Big Data», in *Ohio State Journal Law Journal*, vol. LXXVIII (2017), p. 1217-241.

Barberis, M., *Come internet sta uccidendo la democrazia*, Chiarelettere, 2020.

Bassoli, *Algoritmica giuridica. Intelligenza artificiale e diritto*, Amon, 2021.

Beckers, A. e Teubner, G., *Three Liability Regimes for Artificial Intelligence: Algorithmic Actants, Hybrids, Crowds*, Hart, 2019.

Bekey, G. A., «Current Tends in Robotics: Technology and Ethics», in *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT, 2012, p. 18-34.

Bellman, R. E., *An Introduction to Artificial Intelligence: Can Computer Think?*, Boyd e Fraser, 1978.

Bench-Capon, T. J. M., «HYPO'S legacy: introduction to the virtual special issue», in *Artificial Intelligence and Law*, vol. XXV (2017), p. 205-250.

Bench-Capon, T. e Atkinson, K., *The Roles of Dimensions and Values in Legal CBR*, 2018.

Bench-Capon, T. J. M. e Sartor, G., «A Model of Legal Reasoning with Cases Incorporating Theories and Values», in *Artificial Intelligence*, vol. CL (2003), p. 97-142.

— «A Quantitative Approach to Theory Coherence», in *Proceedings of the Fourteenth Annual Conference on Legal Knowledge and Information Systems (JURIX)*, IOS, 2001, p. 53-62.

Bender, E. M. e Koller, A., «Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data», in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, p. 5185-5198.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N. e Turian, J., «Experience Grounds Language», in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online nov. 2020, p. 8718-8735.

Bosco, F., Creemers, N., Ferraris, V., Guagnin, D. e Koops, B.-J., «Profiling Technologies and Fundamental Rights and Values: Regulatory Challenges and Perspectives from European Data Protection Authorities», in *Reforming European Data Protection Law*, a cura di S. Gutwirth, R. Leenes e P. de Hert, Springer, 2015.

Bostrom, N., *Superintelligence*, Oxford University Press, 2014.

Branting, L. K., Pfeifer, C., Brown, B., Ferro, L., Aberdeen, J., Weiss, B., Pfaff, M. e Liao, B., «Scalable and explainable legal prediction», in *Artificial Intelligence and Law*, vol. XXIX (2021), p. 213-238.

Bratman, M., *Intentions, Plans and Practical Reasoning*, Harvard University Press, 1987.

Brooks, R. A., *Intelligence without Reason*, AI Memo, MIT Artificial Intelligence Laboratory, 1991.

Brynjolfsson, E. e McAfee, A., *Race Against the Machine*, Digital Frontier Press, 2011.

Burr, C. e Cristianini, N., «Can Machines Read our Minds?», in *Minds and Machines*, vol. XXIX (2019), p. 461-494.

Butler, S., «Darwin Among The Machines», in *The Press* (June 13th 1863).

— *Erewhon Or Over the Range*, Trubner, 1872.

Calo, M. R., «Robotics and the Lessons of Cyberlaw», in *California Law Review*, vol. CIII (2015), p. 101-48.

— «Robots and Privacy», in *Robot Ethics: The Ethical and Social Implications of Robotics*, Cambridge University Press, 2012, p. 187-201.

Calvano, E., Calzolari, G., Denicolò, V. e Pastorello, S., «Artificial intelligence, algorithmic pricing and collusion», in *American Economic Review*, vol. CX (2020), p. 3267-97.

Capek, K., *R.U.R. (Rossum's Universal Robots)*, Penguin, 2004.

Caravita, B., «Uno staff di laureati per costruire una macchina di intelligenza artificiale che collabori alle funzioni di giustizia», in *Federalismi* (2021), p. 1-8.

Carleo, A. (a cura di), *Decisione robotica*, Il Mulino, 2019.

Castelfranchi, C. e Paglieri, F., «The Role of Beliefs in Goal Dynamics: Prolegomena to a Constructive Theory of Intention», in *Synthese*, vol. CLV (2007), p. 237-63.

Castelli, C. e Piana, D., «Giustizia predittiva. La qualità della giustizia in due tempi», in *Questione Giustizia*, n. 4 (2018), p. 153-65.

Cavino, M., «Il triceratopo di Spielberg. Fake news, diritto e politica», in *Federalismi*, vol. XI (2020), p. 32-42.

Charniak, E. e McDermott, D., *Introduction to Artificial Intelligence*, Addison-Wesley, 1985.

Chopra, S. e White, L. F., *A Legal Theory for Autonomous Artificial Agents*, University of Michigan Press, 2011.

Clarke, A. C., *2001: A Space Odyssey*, (Prima edizione 1968), Penguin, 2000.

Consiglio, E. e Sartor, G., «Il sistema di credito sociale cinese: una "nuova" regolazione sociotecnica mediante sorveglianza, valutazione e sanzione», in *TIGOR* (2021), p. 138-61.

Contissa, G., Docter, K., Lagioia, F., Lippi, M., Micklitz, H.-W., Palka, P., Sartor, G. e Torroni, P., *CLAUDETTE meets GDPR Automating the Evaluation of Privacy Policies using Artificial Intelligence. Study Report, Funded by The European Consumer Organisation*, BEUC, 2018.

Contissa, G., Lagioia, F. e Sartor, G., «The Ethical Knob», in *Artificial intelligence and Law*, vol. XXV (2017), p. 365-78.

Copeland, J., *Artificial Intelligence*, Blackwell, 1993.

Crafa, S., Laneve, C. e Sartor, G., «Stipula: a domain specific language for legal contracts», in *Prolala - Programming Languages and the Law*, 2022.

Cristianini, N. e Scantamburlo, T., «On Social Machines for Algorithmic Regulation», in *AI and Society* (2019).

D'Aloia, A., *Intelligenza artificiale e diritto. Come regolare un mondo nuovo*, Franco Angeli, 2021.

Damasio, A. R., *Feeling & knowing: making minds conscious*, Penguin, 2021.

De Filippi, P. e Wright, A., *Blockchain and the law: The rule of code*, Harvard University Press, 2018.

Deepali, J., Malaya, D. B. e Biswas, A., «Summarization of legal documents: Where are we now and the way forward», in *Computer Science Review* (2021), p. 1-14.

Deleuze, G., «Postscriptum sur les sociétés de contrôle par Gilles Deleuze», in *L'Autre Jurnal* (1990).

Dennett, D. C., *Consciousness Explained*, Little Brown, 1991.

— *Kinds of Minds: Towards an Understanding of Consciousness*, Basic, 1997.

De Silva, L., Meneguzzi, F. e Logan, B., «BDI Agent Architectures: A Survey», in *Proceedings of IJCAI-2020*, 2020, p. 4914-4921.

Dick, P. K., «Second Variety», in *Minority Report*, Gollancz, 2002, p. 61-101.

Dreyfus, H. e Dreyfus, S., *Mind over Machine*, Blackwell, 1986.

Dung, P. M., «On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and *n*–Person Games», in *Artificial Intelligence*, vol. LXXVII (1995), p. 321-57.

Dworkin, R. M., *Law's Empire*, Kermode, 1986.

Edelman, G. M. e Tononi, G., *A Universe of Consciousness*, Basic, 2000.

Engstrom, D. F., Ho, D. E., Sharkey, C. M. e Cuéllar, M.-F., *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*, Administrative Conference of the United States, 2020.

Esposito, E., *Artificial Communication*, MIT, 2022.

Floridi, L., *Etica dell'intelligenza artificiale*, Cortina, 2021.

Floridi, L. e Cabitza, F., *Intelligenza artificiale*, Bompiani, 2021.

Floridi, L. e Chiriatti, M., «GPT-3: Its Nature, Scope, Limits, and Consequences», in *Minds and Machines* (2020).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. e Vayena, E., «AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations», in *Minds and Machines*, vol. XXVIII (2018), p. 689-707.

Floridi, L. e Taddeo, M., «What is data ethics?», in *Philosophical Transactions of the Royal Society A* (2016), p. 374.

Fossa, F., Schiaffonati, V. e Tamburrini, G., *Automi e persone*, Carocci, 2021.

Galbraith, J. K., *American Capitalism: The Concept of Countervailing Power*, Houghton Mifflin, 1956.

Galli, F., Loreggia, A. e Sartor, G., «Potenzialità e limiti della moderazione algoritmica», in *Ecoscienza*, vol. XIII, n. 3 (2022), p. 41-43.

Gigerenzer, G. e Gaissmaier, W., «Heuristic Decision Making», in *The Annual Review of Psychology*, vol. LXII (2011), p. 451-82.

Gordon, T. F. e Walton, D. N., «The Carneades argumentation framework - using presumptions and exceptions to model critical questions», in *Computational Models of Argument. Proceedings of COMMA-06*, a cura di P. Dunne e T. Bench-Capon, IOS Press, 2006, p. 195-207.

Gordon, T. F. e Karacapilidis, N., «The Zeno Argumentation Framework», in *Proceedings of the Sixt International Conference on Artificial Intelligence and Law (ICAIL)*, ACM, 1997, p. 10-9.

Governatori, G., Idelberger, F., Milosevic, Z., Riveret, R., Sartor, G. e Xu, X., «On legal contracts, imperative and declarative smart contracts, and blockchain systems», in *Artificial Intelligence and Law*, vol. XXVI (2018), p. 377-409.

Governatori, G. e Rotolo, A., «Changing legal systems: legal abrogations and annulments in defeasible logic», in *Logic Journal of IGPL*, vol. XVIII (2010), p. 157-94.

Governatori, G., Rotolo, A. e Sartor, G., «Logic and the Law: Philosophical Foundations, Deontics, and Defeasible Reasoning», in *Handbook of Deontic Logic and Normative Systems, Volume 2*, College Publications, 2022, p. 657-764.

Gregory, R. L., «Intelligence», in *The Oxford Companion to the Mind*, a cura di R. L. Gregory, Oxford University Press, 1987, p. 375-379.

Guidotti, R., Monreale, A., Turini, F., Pedreschi, D. e Giannotti, F., «A Survey Of Methods For Explaining Black Box Models», in *ACM Computing Surveys*, vol. LI, n. Article 93 (2018), p. 1-42.

Hallevy, G., *When Robots Kill: artificial intelligence under criminal law*, Northeastern University Press, 2013.

Harayama, Y., Milano, M., Baldwin, R., Antonin, C., Berg, J., Karvar, A. e Wyckoff, A., «Artificial Intelligence and the Future of Work», in *Reflections on Artificial Intelligence for Humanity*, a cura di B. Braunschweig e M. Ghallab, Springer, 2021, p. 53-67.

Haugeland, J. (a cura di), *Artificial Intelligence: The Very Idea*, MIT, 1985.

Hegel, G. W. F., *Fenomenologia dello spirito*, Nuova Italia, 1933.

Heintz, F., Milano, M. e O'Sullivan, B. (a cura di), *Trustworthy AI - Integrating Learning, Optimization and Reasoning - First International Workshop, TAILOR 2020, Virtual Event, September 4-5, 2020, Revised Selected Papers*, Springer, 2021.

Herbert, F., *Dune*, Chilton, 1965.

Hern, A., «Cambridge Analytica: how did it turn clicks into votes», in *Guardian* (6 maggio 2018).

Hernandez Marin, R. e Sartor, G., «Time and Norms: A Formalisation in the Event-calculus», in *Proceedings of the Seventh International Conference on Artificial Intelligence and Law (ICAIL)*, ACM, 1999, p. 90-100.

Herrestad, H. e Krogh, C., «Obligations Directed from Bearers to Counterparties», in *Proceedings of the 5th International Conference on Artificial Intelligence and Law (ICAIL)*, ACM, 1995, p. 210-8.

Hildebrandt, M., «Profiling and AML», in *The Future of Identity in the Information Society. Challenges and Opportunities*, a cura di K. Rannenberg, D. Royer e A. Deuker, Springer, 2009.

AI-HLEG, High-Level Expert Group on Artificial Intelligence, *A definition of AI: Main capabilities and scientific disciplines*, European Commission, 2019.

— *Ethics Guidelines for Trustworthy AI*, European Commission, 2019.

Hofstadter, D., «The Shallowness of Google Translate», in *The Atlantic* (gen. 2018).

Hogger, C. J. e Kowalski, R. A., «Logic Programming», in *Encyclopedia of Artificial Intelligence*, a cura di S. C. Shapiro e D. N. Y. Eckroth, Wiley, 1987, p. 544-58.

Horty, J. F., «Reasoning with Dimensions and Magnitudes», in *International Conference on Artificial Intelligence and Law, ICAIL2017*, ACM, 2017.

— «Rules and reasons in the theory of precedent», in *Legal theory*, vol. X (2011), p. 1-33.

Janka, S. F. e Uhsler, S., «Antitrust 4.0-the rise of Artificial Intelligence and emerging challenges to antitrust law», in *European Competition Law Review*, vol. XXXIX, n. 3 (2018), p. 112-123.

Johnson, P. e Masri, G., «Making Better Determinations, Discussion Paper No.7», in *Future Challenges for E-government*, Australian Government, Department of Finance e Deregulation, 2004.

Jones, A. J. e Sergot, M. J., «A Formal Characterisation of Institutionalised Power», in *Journal of the IGPL*, vol. IV (1996), p. 429-45.

Kahneman, D., Sibony, O. e Sunstein, C. R., *Noise: A Flaw in Human Judgement*, Collins, 2021.

Katz, D. M., Bommarito, M. J. e Blackman, J., «A General Approach for Predicting the Behavior of the Supreme Court of the United States», in *PLoS ONE*, vol. XII (2017).

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. e Mullainathan, S., «Human Decisions and Machine Predictions», in *The Quarterly Journal of Economics*, vol. CXXXIII (2017), p. 237-93.

Kleinberg, J., Ludwig, J., Mullainathan, S. e Sunstein, C. R., «Discrimination in the Age of Algorithm», in *Journal of Legal Analysis*, vol. X (2018), p. 113-174.

Kowalski, R. e Datoo, A., «Logical English meets legal English for swaps and derivatives», in *Artificial Intelligence and Law*, vol. XXX (2022), p. 163-197.

Kurzweil, R., *How to Create a Mind*, Viking, 2012.

— *The Age of Spiritual Machines*, Orion, 1999.

— *The Singularity is Near*, Viking, 2005.

Lagioia, F., *L'intelligenza artificiale in sanità: un'analisi giuridica*, Giappichelli, 2020.

Lagioia, F., Rovatti, R. e Sartor, G., «Fairness through Group Parities? The Case of COMPAS-SAPMOC», in *AI and Society* (2022).

Lagioia, F. e Sartor, G., «AI systems under criminal law: A legal analysis and a regulatory perspective», in *Philosophy of Technology*, vol. XXXIII (2019), p. 433-465.

— «Profilazione e decisione algoritmica: dal mercato alla sfera pubblica», in *Federalismi*, vol. XI (2020), p. 85-110.

Lagioia, F., Sartor, G. e Simoncini, A., «Art. 22. Processo decisionale automatizzato relativo alle persone fisiche, compresa la profilazione», in *Codice della privacy e Data Protection*, Giuffrè, 2021, p. 378-390.

Larson, J., Mattu, S., Kirchner, L. e Angwin, J., *How We Analyzeded the COMPAS Recidivism Algorithm*, Pro Publica, 2016.

Leibniz, G. W., *Meditations on Knowledge, Truth, and Ideas*, a cura di J. Bennett, 2004.

Leith, P., «Clear Rules and Legal Expert Systems», in *Automated Analysis of Legal Texts*, a cura di A. Martino e F. Socci, North Holland, 1986, p. 661-79.

Licklider, J. C. R., «Man-Computer Symbiosis», in *IRE Transactions on Human Factors in Electronics*, vol. HFE-1, n. March (1960), p. 4-11.

Mac Sithigh, D. e Siems, M., «The Chinese Social Credit System: A Model for Other Countries?», in *Modern Law Review*, vol. LXXXII (2019), p. 1034-1071.

MacCormick, D. N. e Summers, R. S. (a cura di), *Interpreting Precedents: A Comparative Study*, Darthmouth, 1997.

— (a cura di), *Interpreting Statutes: A Comparative Study*, Darthmouth, 1991.

Marcus, G. e Davis, E., *Rebooting AI: building artificial intelligence we can trust*, Pantheon Books, 2019.

Margetts, H., «Rethinking AI for Good Governance», in *Dedalus*, vol. CLI (2022), p. 360-71.

Mazzarese, T., *Forme di razionalità delle decisioni giudiziali*, Giappichelli, 1997.

McAfee, A. e Brynjolfsson, E., *Machine, Platform, Crowd*, Norton, 2019.

McCarthy, J., *What Is Artificial Intelligence*, rapp. tecn., Stanford University, 2007.

McCarthy, J. e Hayes, P., «Some Philosophical Problems of Artificial Intelligence», in *Readings in Artificial Intelligence*, a cura di B. Webber e N. Nilsson, Morgan Kaufmann, 1987, p. 431-50.

McCarthy, J., Minsky, M., Rochester, N. e Shannon, C. E., *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 1956, `http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html`.

McCarty, L. T., «Artificial Intelligence and Law: How to Get There from Here», in *Ratio Juris*, vol. III (1990), p. 189-200.

McEwan, I., *Machines Like Me*, Vintage, 2019.

Micklitz, H.-W., Pollicino, O., Reichman, A., Simoncini, A., Sartor, G. e De Gregorio, G., *Constitutional Challenges in the Algorithmic Society*, Cambridge University Press, 2021.

Mill, J. S., *On Liberty*, (1st ed. 1859.), Penguin, 1974.

Montesquieu, C., *L'esprit des lois*, 1748.

Newell, A. e Simon, H. A., «Computer Science as Empirical Inquiry: Symbols and Search», in *Communications of the ACM*, vol. XIX (1976), p. 113-26.

Nilsson, N., *Artificial Intelligence: A New Synthesis*, Morgan Kaufmann, 1998.

Novelli, C., Bongiovanni, G. e Sartor, G., «A conceptual framework for legal personality and its application to AI», in *Jurisprudence* (2021).

O'Neil, C., *Weapons of math destruction: how big data increases inequality and threatens democracy*, Crown Business, 2016.

Pagallo, U., *The laws of robots*, Springer, 2013.

Palmirani, M., Governatori, G., Rotolo, A., Tabet, S., Boley, H. e Paschke, A., «LegalRuleML: XML-Based Rules and Norms», in *Rule-Based Modeling and Computing on the Semantic Web*, a cura di F. Olken, M. Palmirani e D. Sottara, Springer Berlin Heidelberg, Berlin, Heidelberg 2011, p. 298-312.

Palmirani, M., Vitali, F., Van Puymbroeck, W. e Nubla Durango, F., *Legal Drafting in the Era of Artificial Intelligence and Digitisation*, European Commission, 2022.

Pariser, E., *The Filter Bubble*, Penguin, 2011.

Pasquale, F., *The black box society: the secret algorithms that control money and information*, Harvard University Press, 2015.

Pentland, A., *Social Physics: How Social Networks Can Make Us Smarter*, Penguin, 2015.

Philipps, L., «Distribution of damages in car accidents through the use of neural networks», in *Cardozo L. Rev.*, vol. XIII (1991), p. 987-1000.

Pitruzzella, G. e Pollicino, O., *Disinformation and hate speech. A European Constitutional Perspective*, Bocconi University Press, 2020.

Pizzetti, F., «Fake news e allarme sociale: responsabilità, non censura», in *Medialaws* (2017), p. 48-59.

Pizzetti, F. (a cura di), *Intelligenza artificiale, protezione dei dati personali e regolazione*, Giappichelli, 2018.

Polanyi, K., *The Great Transfomation*, First ed. 1944, Beacon Press, [1944] 2001.

Polanyi, M., *The tacit dimension*, University of Chicago Press, [1966] 2009.

Pollock, J. L., *Cognitive Carpentry: A Blueprint for How to Build a Person*, MIT, 1995.

Poole, D. L., Mackworth, A. K. e Goebel, R., *Computational Intelligence: A Logical Approach*, Oxford University Press, 1998.

Prakken, H. e Sartor, G., «Argument-based Extended Logic Programming with Defeasible Priorities», in *Journal of Applied Non-classical Logics*, vol. VII (1997), p. 25-75.

— «Law and logic: A review from an argumentation perspective», in *Artificial Intelligence*, vol. CCXXVII (2015), p. 214-45.

— «Modelling Reasoning with Precedents in a Formal Dialogue Game», in *Artificial Intelligence and Law*, vol. VI (1998), p. 231-87.

Reichman, A. e Sartor, G., «Algorithms and Regulation», in *Constitutional Challenges in the Algorithmic Society*, Cambridge University Press, 2021, p. 131-81.

Rich, E. e Knight, K., *Artificial Intelligence*, McGraw-Hill, 1991.

Romeo, F., *Lezioni di logica e di informatica giuridica*, Giappichelli, 2012.

Romeo, F. e Barbarossa, F., «Simulation of Verdicts in Civil Liability», in *World Congress on Neural Networks, WCNN, San Diego California USA, June 5-9 1994*, Hillsdale New Jersey USA, 1994, vol. I, p. 432-436.

Rosenblueth, A., Wiener, N. e Bigelow, J., «Purpose and Teleology», in *Philosophy of Science*, vol. X (1943,), p. 18-24.

Ruffo, G. e Tambuscio, M., «Capire la diffusione della disinformazione e come contrastarla», in *Federalismi*, vol. XI (2020), p. 73-84.

Ruffolo, U. (a cura di), *Intelligenza Artificiale e Responsabilità*, Giuffrè, 2018.

— (a cura di), *Intelligenza artificiale.Il diritto, i diritti, l'etica*, Giuffrè, 2020.

Ruffolo, U., «La machina sapiens come "avvocato generale" ed il primato del giudice umano: una interazione virtuosa», in *Lezioni di diritto dell'intelligenza artificiale*, Giappichelli, 2021.
— *XXVI lezioni di diritto dell'intelligenza artificiale*, Giappichelli, 2021.
Ruggeri, F., Lagioia, F., Lippi, M. e Torroni, P., «Detecting and explaining unfairness in consumer contracts through memory networks», in *Artificial Intelligence and Law*, vol. XXX, n. 1 (2022), p. 59-92.
Rumelhart, D. E. e McClelland, J. L. (a cura di), *Parallel Distributed Processes: Explorations in the Microstructure of Cognition*, MIT, 1986.
Russell, S. J., *Human Compatible*, Viking, 2019.
Russell, S. J. e Norvig, P., *Artificial Intelligence. A Modern Approach*, 3ª ed., Prentice Hall, 2010.
— *Artificial Intelligence. A Modern Approach*, 4ª ed., Pearson, 2021.
Ryan, R. M. e Deci, E. L., *Self-Determination Theory Basic Psychological Needs in Motivation, Development, and Wellness*, Guilford, 2017.
Sabo, I. C., Dal Pont, T. R., Wilton, P. E. V., Rover, A. J. e Huebner, J. F., «Clustering of Brazilian legal judgments about failures in air transport service: an evaluation of different approaches», in *Artificial Intelligence* (2021), p. 1-37.
Santosuosso, A., *Intelligenza artificiale e diritto. Perché le tecnologie di IA sono una grande opportunità per il diritto*, Mondadori, 2020.
Santosuosso, A. e Sartor, G., *La giustizia predittiva: una visione realistica*, Giurisprudenza italiana, 2022.
Sartor, G., «Artificial intelligence and human rights: Between law and ethics», in *Maastricht Journal of European and Comparative Law*, vol. XXVII, n. 27 (2020), p. 705-719.
— «Cognitive Automata and the Law: Electronic Contracting and the Intentionality of Software Agents», in *Artificial Intelligence and Law*, vol. XVII (2009), p. 253-90.
— «Defeasibility in Law», in *Handbook of Legal Reasoning and Argumentation*, a cura di G. Bongiovanni, G. Postema, A. Rotolo, G. Sartor, C. Valentini e D. Walton, Springer, 2018, p. 315-64.
— «Fundamental Legal Concepts: A Formal and Teleological Characterisation», in *Artificial Intelligence and Law*, vol. XXI (2006), p. 101-42.
— «Human Rights and Information Technologies», in *The Oxford Handbook on the Law and Regulation of Technology*, a cura di R. Brownsword, E. Scotford e K. Yeung, Oxford University Press, 2017, p. 424-450.
— *Le applicazioni giuridiche dell'intelligenza artificiale: la rappresentazione della conoscenza*, Giuffrè, 1990.
— *Legal Reasoning: A Cognitive Approach to the Law*, a cura di E. Pattaro, Springer, 2005, p. 1-845.
— *The way forward for better regulation in the EU – better focus, synergies, data and technology*, European Parliament, 2022.
Sartor, G. e Lagioia, F., *Study: The impact of the General Data Protection Regulation on artificial intelligence*, European Parliament, 2020.
Sartor, G., Lagioia, F. e Galli, F., *Study: Regulating targeted and behavioural advertising in digital services How to ensure users' informed consent*, European parliament, 2021.
Sartor, G. e Loreggia, A., *Study: The impact of algorithms for online content filtering or moderation ("upload filters")*, European Parliament, 2020.

Sartor, G. e Omicini, A., «The Autonomy of Technological Systems and Responsibilities for their Use», in *Autonomous Weapons Systems: Law, Ethics, Policy*, a cura di N. Bhuta, S. Beck, R. Geiss, C. Kress e H. Y. Liu, Cambridge University Press, 2016, p. 39-74.

Schauer, F., *Profiles, Probabilities and Stereotypes*, Belknap, 2003.

Searle, J. R., «Minds, Brains and Programs», in *The Behavioural and Brain Science* (1980), p. 417-57.

Sergot, M. J., Sadri, F., Kowalski, R. A., Kriwaczek, F., Hammond, P. e Cory, H., «The British Nationality Act as a Logic Program», in *Communications of the ACM*, vol. XXIX (1986), p. 370-86.

Simon, H. A., *Models of Man: Social and Rational*, Wiley, 1957.

Sunstein, C. R., *On rumors: how falsehoods spread, why we believe them, what can be done*, Farrar, Straus e Giroux, 2009.

— *Republic: divided democracy in the age of social media*, Princeton University Press, 2020.

Surdeanu, M., Nallapati, R., Gregory, G., Walker, J. e Manning, C., «Risk analysis for intellectual property litigation», in *ICAIL-2011*, ACM, 2011, p. 116-20.

Susskind, R., *Expert Systems in Law: A Jurisprudential Inquiry*, Oxford University Press, 1988.

Taruffo, M., *La motivazione della sentenza civile*, Cedam, 1975.

Tegmark, M., *Life 3.0. Being Human in the Age of Artificial Intelligence*, Knopf, 2017.

Thagard, P., *Conceptual Revolutions*, Princeton University Press, 1992.

Toulmin, S., *The Uses of Argument*, (1st ed. 1958.), Cambridge University Press, [1958] 2003.

Turing, A. M., «Computer Machinery and Intelligence», in *Mind*, vol. LIX (1950), p. 433-60.

Varian, H. R., «Computer Mediated Transactions», in *American Economic Review:* vol. C (2010), p. 1-10.

Vico, G., *De antiquissima Italorum sapientia*, Laterza, [1709] 1917.

Vincenti, E., «Massimazione e conoscenza della giurisprudenza nell'era digitale», in *Questione Giustizia*, n. 4 (2018), p. 147-152.

Walton, D. N., Reed, C. e Macagno, F., *Argumentation Schemes*, Cambridge University Press, 2008.

Walzer, M., *Spheres of Justice*, Basic Books, 1983.

Waterman, D. e Peterson, M., «Rule-Based Models of Legal Expertise», in *Proceeding National Conference on Artificial Intelligence - AAAI*, 1980, p. 272-5.

Weber, M., *La scienza come professione*, Bompiani, [1919] 2008.

Weizenbaum, J., *Computer Power and Human Reason: From Judgement to Calculation*, Freeman, 1977.

Winston, P. H., *Artificial Intelligence*, Addison-Wesley, 1992.

Zeleznikow, J. e Stranieri, A., «The Split-Up System: Integrating Neural Networks and Rule Based Reasoning in the Legal Domain», in *Proceedings of the Fifth International Conference on Artificial Intelligence and Law (ICAIL)*, ACM, 1995, p. 185-94.

Zuboff, S., *The Age of Surveillance Capitalism*, Hachette, 2019.

Finito di stampare nel mese di ottobre 2022
nella Stampatre s.r.l. di Torino
Via Bologna 220