Università
Ca'Foscari
Venezia

Corso di laurea
Magistrale

in Scienze Filosofiche

Tesi di Laurea

Considerazioni sulla IA a partire dal pensiero di Tommaso d'Aquino

Relatore
Ch. Prof. Paolo Pagani

Correlatrice
Ch.ma Prof.ssa Teresa Scantamburlo

Laureando
Marco Peris
881676

Anno accademico
2023/2024

Sommario

Introduzione 4

1 Intelligenza artificiale: alcune coordinate essenziali 6

1.1 Breve storia dell'IA: nascita e sviluppo 6

1.2 Il significato di "Intelligenza Artificiale" 17

1.3 Le forme dell'intelligenza artificiale 23

1.3.1 Weak AI e Strong AI 26

2 AI, Machine learning e Deep learning 28

2.1 Il labirinto di C. Shannon 28

2.2 Apprendimento in IA 31

2.3 L'approccio simbolico 33

2.4 L'approccio sub-simbolico 34

2.4.1 Com'è fatto un neurone artificiale 35

2.5 Tipi di apprendimento 38

2.5.1 Apprendimento supervisionato 38

2.5.2 Apprendimento non-supervisionato 39

2.5.3 Apprendimento semi-supervisionato 39

2.5.4 Apprendimento umano e artificiale 40

2.6 *Deep learning* e reti neurali artificiali 41

2.6.1 Il neurone sigmoidale 42

2.6.2 Come apprende una rete neurale 44

2.6.3 La questione degli hidden layers 51

2.7 Un nuovo tipo di IA: *GenAI* 53

2.7.1 Cos'è e come funziona ChatGPT 54

3 Gli agenti razionali 62

3.1 Cosa si intende per "agente" 62

3.2 Il concetto di razionalità 63

3.3 I “programmi agente” 66
3.4 Prendere una decisione 69
3.4.1 La teoria dell’utilità attesa 70
3.4.2 Utilità attesa e agenti razionali 73
4 Etica dell’IA 78
4.1 Agenti morali artificiali 82
4.2 Principali questioni relative alla *machine ethics* 86
4.3 Approcci all’etica computazionale 89
4.3.1 L’approccio utilitarista 90
4.3.2 L’approccio deontologico e l’approccio misto 92
4.3.3 L’etica delle virtù 93
4.3.4 Approcci top-down, bottom-up e misto 94
4.4 Orientamenti etici per un’IA affidabile 95
4.5 Stato attuale dell’etica dell’IA 98
5 La teoria dell’azione in Tommaso d’Aquino 99
5.1 Struttura generale della *Summa* 100
5.1.1 Organizzazione dell’articolo 101
5.2 La questione del fine ultimo 101
5.2.1 Il fine ultimo e gli agenti morali artificiali 104
5.3 Il ruolo della volontà 105
5.4 La scansione dell’agire umano in dodici atti 106
5.4.1 Questione 12: l’intenzione 109
5.4.2 Questione 14: la deliberazione o consilium 112
5.4.3 Questione 15: il consenso 115
5.4.4 Questione 13: l’elezione o scelta 116
5.5 Considerazioni intorno alle conseguenze dell’azione 124
6 Conclusioni 129

7 Bibliografia 132

Introduzione

Il presente lavoro si propone di affrontare il tema dell'intelligenza artificiale secondo due diverse prospettive, ingegneristica e filosofica, attraverso un metodo che possa far dialogare tra loro le due diverse discipline, mettendo in risalto i risultati e le sfide ingegneristiche che stanno dietro agli attuali modelli di IA, ma al contempo innestando una riflessione filosofica su alcuni dei presupposti di tali modelli, con particolare attenzione verso il campo dell'etica dell'IA.

Si è deciso di affrontare questo tema per approfondire lo studio dell'intelligenza artificiale non sono dal punto di vista filosofico, ma anche per andare a studiare nello specifico le modalità di funzionamento degli algoritmi di IA. Come si vedrà nei capitoli finali, dedicati alla ricostruzione del processo decisionale degli "agenti morali artificiali", la distinzione tra intelligenza umana e "intelligenza" sintetica passa anche attraverso la comprensione dei differenti processi che le regolano. L'equivalenza dei risultati (o delle *performance*) non deve trarre in inganno: le applicazioni di IA sono strumenti progettati dall'uomo, basati su una logica statistica per il raggiungimento di un certo obiettivo, attraverso la minimizzazione di una funzione di costo. Il fatto che queste applicazioni riescano a scrivere un testo articolato o possano prendere decisioni in maniera più rapida di un essere umano riguarda l'atto nella sua esteriorità, ma il modo in cui l'uomo e la macchina arrivano al risultato è completamente diverso.

Nel Cap.1 si presenterà una breve storia dell'IA, partendo dal progetto presentato nella conferenza di *Darmouth* ed arrivando fino alle applicazioni più recenti e potenti dell'IA. Si andrà poi a precisare che cosa si intende per "intelligenza artificiale", mostrando come questo termine non sia univocamente definibile, ma possa essere inteso secondo quattro principali interpretazioni. Verranno inoltre affrontati i temi della *weak AI* e *strong AI*, differenziandoli rispetto ai concetti di ANI e AGI.

Nel Cap.2 verrà introdotta la differenza tra approccio simbolico e approccio sub-simbolico: quest'ultimo sta alla base di tutti i moderni algoritmi di *deep learning*. Si andranno successivamente a presentare alcuni algoritmi di apprendimento, che vengono normalmente utilizzati per addestrare i modelli di IA e si proporrà un confronto tra apprendimento umano e artificiale. Si avrà poi un'ampia sezione riguardante il neurone sigmoidale e i suoi principi di funzionamento, presentati anche dal punto di vista matematico. Infine verrà presentato il metodo di funzionamento di *ChatGPT*, concentrandosi in particolare sulle sue capacità semantico-sintattiche, mettendone in evidenza i limiti rispetto a quelle propriamente umane.

Nel Cap.3 si tratterà degli agenti artificiali, partendo dalla definizione di “agente”, ma soprattutto concentrandosi sul significato di “razionale”, che come si vedrà è strettamente connesso con la dimensione pratica dell’agire. Si procederà presentando le differenze che ci sono tra un agente e l’altro a seconda dei differenti *software* da essi implementati, chiamati “programmi agente”. L’ultima parte si concentrerà nello studio della teoria dell’utilità attesa, che rappresenta la base teorica su cui poggia la costruzione del processo decisionale degli agenti artificiali.

Nel Cap.4 verrà introdotto il campo dell’etica dell’IA, andando a presentare i principali approcci utilizzati e le teorie etiche di riferimento, mostrando come l’approccio utilitarista si presti meglio degli altri a una traduzione algoritmica. Per ultimo verrà introdotto l’importante documento della Commissione Europea sugli orientamenti per un’IA affidabile, che rappresenta una delle migliori pubblicazioni al momento per comprendere come la dimensione etica avrà un’importanza enorme nei futuri sviluppi dell’IA.

Nel Cap.5 si presenterà la teoria dell’azione in Tommaso d’Aquino, ponendo particolare attenzione al commento filosofico, ma cercando comunque di interloquire anche con il mondo ingegneristico, per mostrare come l’analisi operata da Tommaso possa essere utilizzata per progettare un sistema etico di controllo delle applicazioni dell’IA. Attraverso la rilettura di Tommaso saremo in grado di porre in rilievo alcune differenze sostanziali tra il modo di agire prettamente umano e quello invece artificiale.

1 Intelligenza artificiale: alcune coordinate essenziali

Lo scopo di questo capitolo è quello di offrire alcune coordinate essenziali per orientarsi nel panorama dell'intelligenza artificiale: si partirà presentando una breve storia dalle origini negli anni '50 del secolo scorso fino ai giorni nostri, delineando gli eventi salienti per lo sviluppo di questa tecnologia; verrà affrontato il problema definitorio attorno al concetto di IA attraverso la presentazione di vari approcci, proponendo la definizione più corretta al momento disponibile nel panorama di studio; si procederà poi ad un approfondimento circa gli algoritmi che consentono all'IA di imparare, concentrandosi in particolare sugli algoritmi relativi al *deep learning*, relativi a reti neurali modellate sul funzionamento dei neuroni presenti nel cervello umano. Per rendere conto della complessità operativa che sta dietro a tali algoritmi, si presenterà un esempio di funzionamento di una rete neurale artificiale dotata di *back-propagation* dell'errore, utilizzando anche la notazione matematica. A conclusione di questa prima Parte verranno introdotte - senza pretesa di esaustività - le nuove tecnologie riguardanti le IA "generative", prendendo come esempio l'ormai famosa ChatGPT e cercando di mostrare come lavorino gli algoritmi presentati nelle parti precedenti per permettere i risultati sorprendenti di cui tutti noi oggi siamo testimoni.

1.1 Breve storia dell'IA: nascita e sviluppo

Il termine "intelligenza artificiale" viene usato per la prima volta il 31 agosto 1955 da John McCarthy insieme a Marvin Lee Minsky, Claude Elwood Shannon e Nathaniel Rochester[1] all'interno del documento intitolato *Proposta per il progetto di ricerca estivo di Darmouth sull'intelligenza artificiale*[2], nel quale si progettava di avviare durante l'estate dell'anno successivo uno studio della durata di due mesi attorno all'IA:

> Lo studio è da condurre sulla base dell'ipotesi che ogni aspetto relativo all'apprendimento o una qualsiasi altra caratteristica dell'intelligenza possano in principio essere descritti così precisamente da poter costruire una macchina capace di simulare tali caratteristiche. Si tenterà di capire come costruire

[1] J. McCarthy (1927-2011) e M.L. Minsky (1927-2016) sono matematici statunitensi, considerati tra i padri dell'informatica e dell'intelligenza artificiale. C.E. Shannon (1916-2001) è matematico ed ingegnere statunitense, considerato il padre della teoria dell'informazione. N. Rochester (1919-2001) è un ingegnere elettronico statunitense che ha sviluppato Assembly: un linguaggio di programmazione a basso livello. All'epoca della "Proposta", McCarthy era assistente professore a Dartmouth, Minsky faceva parte della *Harvard Society of Fellows* come *Junior Fellow*, Shannon lavorava come ricercatore presso i laboratori della *Society Bell Telephone* e Rochester era impiegato presso la IBM.

[2] Testo originale reperibile al link: dartmouth.dvi (stanford.edu), oppure "AI magazine", volume 27, numero 4, 1955.

> macchine in grado di utilizzare il linguaggio, formare astrazioni e concetti, risolvere problemi ora solo riservati all'uomo e migliorare sé stesse. Riteniamo che un progresso significativo in uno o più di questi campi possa essere fatto se un gruppo accuratamente selezionato di scienziati ci lavorano insieme per un'estate[3].

I punti principali su cui la ricerca si sarebbe dovuta soffermare erano sette: la programmazione di computer automatici, la programmazione di un computer per l'uso del linguaggio, lo sviluppo di reti neurali, lo sviluppo di una teoria della computabilità, l'auto miglioramento, la capacità di astrazione, il rapporto tra casualità e creatività.

Nell'estate del 1956 i lavori presentati nella "Proposta" presero il via presso il Dartmouth College, con il nome di *Progetto di ricerca estivo di Dartmouth sull'intelligenza artificiale*. I due mesi previsti non bastarono per raggiungere tutti gli ambiziosi obiettivi del progetto ma gettarono le basi per tutte le successive ricerche del gruppo. Nello stesso anno venne presentato il *Logic Theorist*[4], considerato il primo programma di ragionamento automatico[5]. Realizzato dal matematico A. Newell, dall'economista e psicologo H.A. Simon e dal programmatore C. Shaw, era in grado di dimostrare 38 dei 52 teoremi presenti nei *Principia Mathematica* di Whitehead e Russell. In particolare, la prova del teorema 2.85 fu svolta in modo differente rispetto a quella presente nei *Principia*, ma non venne accettata per la pubblicazione in *The Journal of Symbolic Logic*, probabilmente a causa del fatto che tra gli autori citati c'era lo stesso programma *Logic Theorist*. Dopo questo successo, Newell e Simon si misero a sviluppare il *General Problem Solver* o GPS, progettato fin da principio per imitare i procedimenti umani durante la risoluzione dei problemi. Nel 1959, Herbert Gelernter scrisse il *Geometry Theorem Prover*, in grado di dimostrare teoremi di geometria.

La conferenza di Darmouth fu preceduta da due avvenimenti importanti che risultarono centrali per lo sviluppo futuro dell'IA: la presentazione del primo modello di rete neurale artificiale e la pubblicazione dell'articolo di Turing[6] sulla rivista britannica *Mind* che rispondeva alla domanda *"Can machines think?"*. Entrambi gli apporti furono decisi sia sul piano ingegneristico-informatico che sul piano filosofico.

---

[3] *Ivi.*

[4] Cfr. A. Newell, H.A. Simon, *The logic theory machine, a complex information processing system*, giugno, 1956, reperibile su: logictheorymachine.pdf (cmu.edu)

[5] Simon, a tal proposito, dichiarò: «abbiamo inventato un programma per computer capace di pensare in modo non numerico, e abbiamo quindi risolto l'antichissimo problema mente-corpo».

[6] Cfr. A.M Turing, *Computing Machinery and Intelligence,* "Mind", vol. 59, no. 236, 1950, pp. 433–460. *JSTOR*, http://www.jstor.org/stable/2251299.

Nel 1943 due studiosi statunitensi, il matematico W.H Pitts e il neurofisiologo W.S. McCulloch proposero per la prima volta un modello computazionale[7] ispirato al funzionamento dei neuroni all'interno del cervello umano. Per arrivarci misero insieme la loro conoscenza delle basi di fisiologia cerebrale e del funzionamento dei neuroni, gli studi di logica formale di Russell e Whitehead e la teoria della computazione di Turing. L'unità base di calcolo di tale modello venne chiamato "neurone di McCulloch-Pitts" e rimase l'antesignana di tutte le reti neurali fino ai giorni nostri: si tratta del modello più semplice in grado di imitare il modo di funzionamento dei neuroni.

Come vedremo in modo più approfondito in uno dei seguenti paragrafi, il neurone di McCulloch-Pitts è in grado di ricevere segnali di input e di trasmettere o meno un segnale in uscita, a seconda dei parametri interni regolabili dall'operatore. Con questa configurazione era in grado di simulare qualsiasi funzione logica[8] (ad esclusione dello XOR[9], anche detta OR-esclusivo). Le limitazioni principali di questo modello risiedono nel fatto che i valori di input/output potevano essere solo binari (0 o 1, -1 o 1) e che non includevano un algoritmo di auto-apprendimento, non essendo in grado così di aggiornare in maniera autonoma i parametri interni.

Nel 1950 A. Turing pubblica sulla rivista "*Mind*" un articolo che contiene una proposta per discernere quando una macchina esprime un comportamento intelligente oppure no; questo criterio assunse negli anni successivi il nome di *test di Turing* e rappresenta ancora oggi un riferimento fondamentale intorno agli studi sull'intelligenza artificiale – si prenda ad esempio il test di Turing proposto nella declinazione morale[10], per capire quando una macchina può essere ritenuta in grado di agire come un essere umano.

Il test è chiamato originariamente da Turing "gioco dell'imitazione": un gioco in cui figurano tre persone, ovvero un uomo (A), una donna (B) ed un interrogatore (C). Tutti e tre i partecipanti

---

[7] Cfr. W.S. McCulloch e W. Pitts, *A logical calculus of the ideas immanent in nervous activity*, "Bulletin of Mathematical Biology", Vol. 52, No 1-2, pp. 99-115, 1990. Reperibile: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiX-MTZw6aIAxWXgf0HHaMmKQUQFnoECBEQAQ&url=https%3A%2F%2Fwww.cs.cmu.edu%2F~.%2Fepxing%2FClass%2F10715%2Freading%2FMcCulloch.and.Pitts.pdf&usg=AOvVaw0yhn00mFm05rk66caMGT8O&opi=89978449

[8] Le tre funzioni elementari in logica booleana sono: NOT, AND e OR. Tali operatori logici implementati nei circuiti elettronici assumono il nome di porte logiche.

[9] Un enunciato composto col connettivo XOR è vero solamente se le proposizioni di partenza hanno valore di verità opposto.

[10] Si veda C. Allen *et alii*., *Prolegomena to any future artificial moral agent*, "Journal of Experimental and Theoretical Artificial Intelligence", vol. 12, 2000, pp. 254-255.

sono in stanze diverse e possono comunicare tra di loro solo attraverso telescriventi. L'interrogatore sa solamente che sta comunicando con X e Y, ma non sa quale dei due sia uomo e quale donna. Attraverso delle domande ad entrambi gli altri due partecipanti, l'interrogatore alla fine del gioco sarà in grado di dire se "X è A e Y è B" o "X è B e Y è A": assegnare cioè il sesso corretto alle due persone nelle altre stanze. Per il giocatore B, scopo del gioco è quello di aiutare l'interrogatore, mentre per A lo scopo è quello di confonderlo. Cosa succederebbe se ad A si sostituisse una macchina? Le probabilità che l'interrogatore sbagli ad indovinare rimangono le stesse di quando A è una persona in carne ed ossa?[11]

Negli sviluppi successivi il test perde l'aspetto relativo al sesso di A e B e si concentra sulle capacità di manipolazione del linguaggio dei computer: C pone delle domande ad A e B, e dalle risposte deve essere in grado di capire chi è la persona in carne ed ossa e chi la macchina. Nello stesso articolo Turing dichiara di aspettarsi che «entro cinquant'anni sarà possibile programmare computer […] capaci di partecipare al gioco dell'imitazione così bene che un interrogatore medio non avrà più del 70% di probabilità di fare la corretta identificazione dopo cinque minuti di domande»[12]. Nel 2014, 60 anni dopo la morte di Turing, presso la *Royal Society* di Londra il chatbot Eugene Goostman[13] fu il primo programma in grado di ingannare il 33% di una commissione di 30 giudici portandoli a credere, durante una sessione di conversazioni scritte (300 in totale) di cinque minuti ciascuna – che comprendeva anche altri 4 chatbot e 5 esseri umani –, che fosse un essere umano in carne ed ossa e non un computer[14]. Eugene era stato programmato per presentarsi come un bambino di 13 anni residente ad Odessa, capace di compiere piccoli errori grammaticali nel rispondere in inglese, per risultare più convincente. La prestazione di Eugene fu contestata su più punti: una delle critiche mosse fu che secondo la formulazione originale del test di Turing, per il quale è necessario che l'esaminatore sbagli con la stessa frequenza sia quando deve distinguere tra computer e uomo sia quando deve distinguere tra due esseri umani, Eugene non era stato in grado di passare il test, ingannando molto meno del 50% dei giudici. Nonostante tutto, il risultato conseguito dal chatbot fu notevole nel panorama scientifico.

---

[11] Cfr. A.M. Turing, *Computing Machinery and Intelligence,* pp. 433-434.

[12] I*vi*, p. 442.

[13] Si veda l'articolo sul *Time*, 2014, Interview: Eugene Goostman Passes the Turing Test | TIME

[14] Un test "al contrario" di Turing è quello CAPTCHA (*Completely Automated Public Turing-test-to-tell Computers and Humans Apart*), utilizzato nei siti per verificare che l'utente sia un essere umano e non un bot. A tal proposito vengono utilizzate scritte deformate, immagini e spunte per la verifica.

Nel testo, Turing propone anche una via di sviluppo dell'IA: anziché programmare manualmente l'intelligenza della macchina, il lavoro andrebbe concentrato nello sviluppare algoritmi di apprendimento e successivamente, tramite questi, insegnare alla macchina. Rimane in ogni caso aperta ancora ai giorni nostri la questione su che cosa significhi che "una macchina può pensare", se sia possibile, e in che senso, che il pensiero sia attribuito anche alle macchine. Relativamente a tale problema, Turing non prende una posizione definitiva, definendolo a tratti *meaningless* – prospettando una sua risoluzione in tempi futuri –, ma svolgendo alla fine dell'articolo un lavoro confutatorio diretto verso le nove principali obiezioni che venivano al tempo sollevate di attribuire pensiero alle macchine; ossia[15]:

1. L'obiezione "teologica", secondo cui il pensiero è funzione solamente dell'anima immortale dell'uomo, data da Dio ad ogni uomo e donna, ma non agli altri animali o alle macchine. Dunque nessun animale o macchina può essere in grado di pensare.
2. L'obiezione della "testa nella sabbia", secondo cui le conseguenze portate da macchine pensanti sarebbero troppo terribili da immaginare. Meglio sperare e credere che non siano mai in grado di farlo.
3. L'obiezione "matematica", individuata nei teoremi di incompletezza di Gödel, secondo cui è impossibile costruire un sistema logico assiomatico completo, perché al suo interno esisteranno sempre proposizioni indecidibili (il cui valore di verità non è dimostrabile all'interno dello stesso sistema).
4. L'argomento della "coscienza", secondo cui le macchine non saranno in grado di pensare finché non riusciranno ad essere coscienti dei propri stati cognitivi ed emozionali.
5. L'argomento delle "varie incapacità", per cui si tratta di trovare sempre una capacità X che la macchina non è in grado di compiere (ad esempio: fare errori, amare, imparare dall'esperienza, essere gentile, assaporare dei frutti, ecc…)
6. L'obiezione di Lady Lovelace, secondo cui l'*Analytical Engine* – un computer meccanico ad uso universale progettato dal matematico inglese Charles Babbage nel 1837 – non può originare alcunché di nuovo, ma solamente ciò che noi siamo in grado di programmare.
7. Argomento della "continuità del sistema nervoso", che si oppone all'architettura dei computer di allora, basati su sistemi discreti.

[15] Cfr. *Computing Machinery and Intelligence,* pp. 443-453.

8. Argomento “dell’informalità del comportamento”, secondo cui è impossibile produrre un insieme di regole morali in grado di descrivere come un uomo dovrebbe agire in ogni situazione. Da ciò deriverebbe il fatto che l’uomo non è una macchina, poiché non obbedisce ad un set di istruzioni morali fornito a priori, ma è in grado di agire autonomamente.
9. L’argomento della “percezione extra-sensoriale”, riferito ai fenomeni di telepatia, chiaroveggenza, precognizione e psico-cinesi.

Molte di queste posizioni sono valide ancora oggi, mentre alcune sono state parzialmente confutate: segno che la domanda sollevata da Turing risulta ai tempi odierni ancora attualissima.

L’argomento 4 e l’argomento 1 pongono la questione se sia possibile ridurre la mente umana al cervello. In questa ipotesi essa rappresenterebbe solamente un epifenomeno di connessioni biologiche, e quindi sarebbe possibile riprodurla su supporti di silicio costruiti in modo simile al cervello. Se invece cervello e mente sono qualitativamente diversi, sia pur connessi l’uno all’altro, quest’ultima non sia in alcun caso riproducibile artificialmente. Da un punto di vista tecnico-scientifico, la questione è ancora molto lontana dall’essere risolta, in quanto lo studio del funzionamento del cervello è molto complesso: siamo in grado di tracciare tutti gli impulsi nervosi che partono dalla retina e di seguirne il percorso, ma non sappiamo dove si formi effettivamente l’immagine che noi vediamo.

L’obiezione 3 e 6 individuano il dominio entro cui le macchine, anche quelle “intelligenti”, possono agire, ossia all’interno dei paletti posti dall’algoritmo progettato dall’uomo. Filosoficamente ciò sta ad indicare che qualunque applicazione di IA si muove all’interno di un orizzonte procedurale con principi ben definiti che in nessun caso essa potrà eludere, mentre l’uomo, l’unico ente aperto su di un orizzonte trascendentale, è in grado di andare oltre i contenuti finiti ed è in grado di scegliere le regole da seguire (ad esclusione dei principi trascendentali propri dell’essere). Un programma addestrato sulle regole della geometria classica non potrà in alcun modo decidere di fare a meno del V postulato e di inoltrarsi nelle geometrie non-euclidee, cosa che in modo evidente è stato possibile all’essere umano, libero di interrogarsi attorno ai presupposti del proprio discorso. Si potrebbe obiettare che, anche se l’uomo vive all’interno di un orizzonte più ampio, è inevitabilmente orientato da alcuni principi su cui non ha scelta, vedi ad esempio il principio di non contraddizione (PDNC). Il fatto principale, già ben notato da Aristotele nella sua *Metafisica*, è che tali principi possiedono un carattere elenctico, non sono evadibili o trascendibili, ma sono trascendentali: «*élenchos* è la forma del trascendentale: la forma che esso assume quando, in qualche modo, viene negato. Dire che il

trascendentale è l'elenctico, è come dire che è l'intrascendibile; dove l'intrascendibilità implica l'impossibilità del collocarsi altrove o a distanza rispetto ad esso»[16].

L'argomento 8 anticipa un problema di particolare interesse per i giorni nostri, ossia come fare in modo che gli "agenti intelligenti" guidati dall'IA possano inglobare al loro interno delle regole etiche che consentano loro di affrontare in modo efficace imprevisti in un mondo complesso (pensiamo ad esempio alle applicazioni di guida automatizzata o agli impieghi in guerra di macchine autonome). I principali approcci utilizzati sono quello deontologico, quello utilitaristico-consequenzialistico e l'approccio misto.

Nel 1957 Frank Rosenblatt presenta il suo percettrone[17] (*perceptron*), un lavoro che parte dal modello di neurone di McCulloch-Pitts, ma che lo supera in due aspetti cruciali: è infatti in grado di correggere in autonomia i valori interni grazie ad un algoritmo di apprendimento e può accettare in input valori compresi nell'intervallo tra [0,1]. Dimostra inoltre il "teorema di convergenza del percettrone" per cui, se due classi sono linearmente separabili, l'algoritmo converge ad una soluzione. Nasce così l'apprendimento automatico, denominato per la prima volta *machine learning* nel 1959 da Arthur Samuel nel suo libro *Some studies in machine learning using the game of checkers*[18]. Nel 1960 Rosenblatt realizzerà il "Mark 1 perceptron", una macchina elettromeccanica capace di riconoscere delle immagini di dimensione 20x20 pixel. Negli stessi anni, McCarthy inventa il linguaggio di programmazione Lisp (List Processor), che verrà ampiamente utilizzato nelle prime applicazioni di IA.

A questi primi anni successivi dalla conferenza di Darmouth, caratterizzati da un grande sviluppo e interesse verso l'intelligenza sintetica, seguì una fase definita "primo inverno dell'IA" (detto *AI winter*), in cui si assistette ad un taglio dei fondi da parte degli Stati e delle Università e ad un parziale rallentamento delle relative ricerche, pur rimanendo vivace il lavoro intellettuale ed ingegneristico.

Tre sono i problemi principali a cui andava incontro lo sviluppo dell'intelligenza artificiale[19]:

---

[16] Cfr. P. Pagani, *La forma dell'élenchos*, "Divus Thomas", vol. 118, n. 2, 2015, pp. 152-191.

[17] F. Rosenblatt, *The perceptron: a probabilistic model for information storage and organization in the brain*, "Psychological Review", vol. 65, No 6, 1958, Rosenblatt1958.pdf (upenn.edu). Si veda anche rosenblatt-1957.pdf (bpb-us-e2.wpmucdn.com), *The perceptron, a perceiving and recognizing automaton*, Cornell aeronautical laboratory, Ithaca, 1957.

[18] A.L. Samuel, *Some studies in machine learning using the game of checkers*, pubblicato originariamente in "IBM Journal", Vol.3, No. 3, luglio 1959. Reperibile su: Samuel.pdf (mit.edu)

[19] Useremo come riferimento il testo S. Russel, P. Norvig, (a cura di) F. Amigoni, *Intelligenza artificiale, un approccio moderno*, Vol.1, Pearson Italia, Milano-Torino 2021, pp. 23-24.

- Il modo in cui i software venivano programmati per svolgere particolari compiti non era frutto di un'attenta osservazione del compito in questione, non teneva in conto le caratteristiche della soluzione né come l'algoritmo potesse raggiungere tali soluzioni in modo affidabile, ma ricalcava piuttosto "un'introspezione" di come gli umani affrontavano il problema.
- Il modo in cui i primi programmi di risoluzione dei problemi si muovevano era una combinazione di passi che giungesse alla soluzione corretta. Nel caso di poche azioni possibili e limitate soluzioni il metodo si rivelò efficace e si era convinti che al crescere della complessità sarebbe bastato aumentare la dotazione hardware, tramite memorie più grandi e processori più veloci. Non fu tuttavia così, e anche i dimostratori automatici di teoremi non furono in grado di svolgere il loro compito qualora i teoremi avessero coinvolto più di qualche dozzina di fatti. Nel 1973 Lighthill, incaricato dal governo britannico di stendere una relazione sullo stato degli studi sull'AI, fra le altre cose fece riferimento all'«esplosione combinatoria[20]» come il più importante scoglio nello sviluppo dell'intelligenza artificiale. Il problema fu risolto solamente con lo sviluppo della teoria della complessità computazionale.
- Emersero inoltre dei limiti fondamentali che riguardavano le strutture base su cui costruire l'AI: nel 1969 ad opera di Marvin Lee Minsky e Seymour Papertm uscì il libro *Perceptrons*[21] in cui veniva criticata la struttura e le reali capacità del percettrone di Rosenblatt: la pubblicizzata abilità di riconoscere pattern era nella realtà dei fatti molto limitata ed i problemi che poteva trattare si riducevano a quelli linearmente separabili.

Tuttavia in alcuni campi i lavori procedettero con buoni risultati: nel 1965 e poi nel 1968, a opera di Edward Albert Feigenbaum, Bruce Buchanan e Joshua Lederberg, il programma DENDRAL[22] – scritto in linguaggio Lisp – diventa il primo "sistema esperto", cioè un programma in grado di risolvere problemi in un particolare campo di applicazione con prestazioni simili ad un esperto umano dello stesso ambito. Il programma presentava un comportamento intelligente poiché esprimeva un alto livello di competenza nell'ambito in cui operava, e in concreto aiutava i chimici a ricostruire la struttura molecolare a partire dai dati forniti da uno spettrometro di massa.

[20] Cfr. Lighthill Report, Cambridge University, luglio 1972, accessibile su: lighthill.dvi (ed.ac.uk)

[21] M. Minsky, S. A. Papert, *Perceptrons: an introduction to computational geometry*, "The MIT Press", 2017, https://doi.org/10.7551/mitpress/11301.001.0001

[22] B.G. Buchanan, E.A. Feigenbaum, J. Lederberg, *Heuristic DENDRAL: A program for generating explanatory hypotheses in organic chemistry*, "Machine Intelligence", 4, 1968. (PDF) Heuristic DENDRAL: A program for generating explanatory hypotheses in organic chemistry (researchgate.net)

DENDRAL era un programma *knowledge-driven*[23], in cui l'importanza non ricadeva tanto nella potenza di calcolo ma nel fatto che in memoria erano presenti grandi quantità di dati e tecniche specifiche per il problema che sarebbe andato a risolvere; questa si può chiamare la "base di conoscenza" della macchina. Vi era poi un "motore inferenziale" che si occupava di elaborare le conoscenze in memoria tramite leggi logiche con i dati di input inseriti. Il programma era molto raffinato per l'epoca ed era diviso in 5 sub-programmi[24]: il *preliminatory inference maker*, il *data adjustor*, lo *structure generator*, il *predictor* e l'*evaluation function*.

Lo stesso Feigenbaum, assieme ad altri studiosi, nel 1972 sviluppò MYCIN: un sistema esperto in grado di aiutare i medici nella diagnosi di malattie infettive. MYCIN differiva per due aspetti particolari da DENDRAL[25]: in primo luogo non conteneva in memoria alcun modello teorico generale da cui dedurre le regole, ma queste ultime venivano acquisite tramite interviste agli esperti; in secondo luogo, MYCIN incorporava un metodo di calcolo dell'incertezza, in grado di rispecchiare l'incertezza associata alla conoscenza medica. La prima applicazione commerciale di un sistema esperto che ebbe fortuna fu R1[26], presso la Digital Equipment Corporation: un programma che si occupava di aiutare la configurazione di nuovi sistemi informatici. Nel 1986 si stimò che il risparmio nella configurazione dei sistemi, grazie all'utilizzo di R1 ammontava a 40 milioni di dollari l'anno.

Nel 1974 il matematico statunistense Paul Werbos presentò nella sua tesi di laurea un modo per superare i limiti del percettrone tramite l'utilizzo di tre strati di neuroni, al posto del singolo strato: nasceva così il *multi-layer perceptron*. Tuttavia le critiche indicate da Minsky e Papert nel loro libro resero necessario, affinché si affermasse definitivamente questo tipo di rete, attendere il 1986 quando D. Rumelhart, G.E. Hinton e R.J. Williams concepirono l'algoritmo di retropropagazione dell'errore (*error backpropagation algorithm*). Il modello si basava si un approccio connessionistico alla conoscenza e rappresentava l'alternativa all'approccio simbolico utilizzato fino a quel momento. Da questo momento in poi lo sviluppo dell'AI ebbe un nuovo impulso: dai sistemi esperti, basati su una logica booleana e su di una programmazione puntuale, si passò a lavori basati più sull'utilizzo della probabilità e dell'apprendimento automatico.

---

[23] Si veda anche E.A. Feigenbaum *et alii.*, *DENDRAL: a case study of the first expert system for scientific hypothesis formation*, "Artificial Intelligence", vol. 61, 1993, pp. 209-261. Reperibile su dendral-history.pdf (mit.edu)

[24] Cfr. *Heuristic DENDRAL*, pp. 210-213.

[25] Cfr. *Intellligenza artificiale: un approccio moderno*, p.25.

[26] Cfr. *ibidem*.

Citiamo il lavoro di J. Pearl del 1988, *Probabilistic reasoning in intelligent systems*[27], che portò a una nuova considerazione della probabilità e della teoria delle decisioni all'interno dell'IA, in particolare per lo studio sulle reti bayesiane ed i modelli di Markov.

A partire dalla fine degli anni '90, lo sviluppo di hardware sempre più potenti e la disponibilità di una grande quantità di dati – i cosiddetti big data – su cui poter addestrare gli algoritmi di apprendimento ha portato l'IA al raggiungimento di importanti traguardi: nel 1997 il supercomputer dell'IBM Deep Blue battè Garry Kasparov, l'allora campione del mondo degli scacchi. Deep Blue era in grado di analizzare 200 milioni di posizioni al secondo prima di compiere la sua mossa. Nel 2001 al Super Bowl di Tampa la polizia utilizzò un programma di riconoscimento facciale per individuare la presenza di criminali.

Un'altra branca del *machine learning*, quella del *deep learning*, i cui esperimenti erano iniziati già nel 1970 ed ebbero successo durante gli anni 90 sotto forma di reti neurali convoluzionali, riscosse molta popolarità a partire dagli anni successivi al 2010, soprattutto nel riconoscimento vocale e delle immagini: nel 2010 IBM Watson vince il gioco a quiz americano *Jeopardy!* contro tre campioni umani: era in grado di comprendere il linguaggio naturale per poi cercare la risposta nell'immensa memoria di cui era dotato. Nel 2011 Apple presenta l'assistente vocale Siri, seguita nel 2014 da Amazon con Alexa. Nel 2012 un sistema di *deep learning* sviluppato all'Università di Toronto dal gruppo di Geoffrey Hinton partecipò ad una competizione di classificazione di immagini a partire da migliaia disponibili (cane, auto, foglia, ecc…) con risultati eccellenti rispetto ai metodi utilizzati precedentemente. La rete neurale profonda di Google DeepMind, AlphaGo, batte i campioni umani nel gioco del Go nel 2015, 2016 e 2017; la sua versione aggiornata AlphaGo Zero è in grado di addestrarsi autonomamente, giocando contro sé stessa, senza più richiedere input da parte di esseri umani, ad eccezione delle regole del gioco.

Nel 2024 l'IA trova applicazione nei più svariati ambiti, che elencheremo di seguito, senza pretesa di completezza[28]:

- L'ambito medico, come abbiamo visto, è uno dei primi campi di applicazione dei sistemi esperti. Ora gli algoritmi di IA sono utilizzati nella diagnosi di malattie, soprattutto quando quest'ultima si basa sul riconoscimento delle immagini. Il lavoro è ancora svolto insieme agli specialisti medici, che attraverso le loro conoscenze integrano il lavoro della

[27] J. Pearl, *Probabilistic reasoning in intelligent systems*, Morgan Kaufmann Publisher Inc., San Francisco 1988.

[28] Cfr. *Intelligenza artificiale, un approccio moderno*, pp. 31-33.

macchina. Si stanno testando robot con AI in grado di rapportarsi autonomamente con i pazienti per ricordare la somministrazione dei farmaci.

- Un altro ambito è quello della interpretazione delle immagini. Una branca centrale della robotica è quella della produzione di robot che siano in grado di muoversi autonomamente nello spazio riconoscendo immagini. Dagli equipaggiamenti di sicurezza nelle auto per evitare collisioni, alla guida autonoma di veicoli (es. Waymo), a robot quadrupedi come Spot, sviluppato dalla Boston Dynamics, senza dimenticare applicazioni di supporto a persone con disabilità (es. ipovedenti).
- Pensiamo all'ambito delle raccomandazioni: dai più semplici filtri anti-spam, in grado di categorizzare autonomamente le mail, alle applicazioni utilizzate dai grandi colossi quali Google, Apple, Amazon, Netflix, Youtube, ecc.. che usando algoritmi di profilazione attraverso l'incrocio di dati sono in grado di suggerirci il prossimo prodotto da acquistare. Se i dati sono in sé neutri, il loro utilizzo non lo è: a tal proposito si può ricordare lo scandalo che coinvolse Facebook e Cambridge Analytica nel 2018, che ebbero accesso ai dati di 87 milioni di persone a loro insaputa. In seguito vennero attuate leggi più stringenti per evitare la profilazione e raccolta di dati indiscriminata: il GDPR[29] del 2016 dell'UE ne è un esempio.
- Quanto al riconoscimento vocale: nel 1990 Dragon Dictate è il primo programma nel mercato che consente all'utente di dettare testi al computer dopo una fase di addestramento tramite frasi preimpostate. Ora il riconoscimento vocale è presente su tutti i telefoni e computer di ultima generazione, attraverso gli assistenti vocali Siri (Apple), Cortana (Microsoft), Alexa (Amazon) e Google Assistant (Google), in grado sia di riconoscere il parlato (*speech to text*) che di fornire risposte (*text to speech*) adeguate al contesto.
- Quanto alla traduzione automatica, nel 2024 il più famoso traduttore disponibile online, Google traduttore, consente ormai la traduzione dall'italiano a più di 100 lingue e viceversa: si tratti di parole, di frasi o di intere pagine web o di documenti. Uno dei competitor, DeepL, utilizza un modello linguistico di ultima generazione (*Large Language Model* o LLM) e algoritmi di addestramento proprietari per fornire prestazioni comparabili.

[29] Cfr. La protezione dei dati nell'UE - Commissione europea (europa.eu)

- Veicoli robotizzati: *Shakey*[30], il primo robot in grado di muoversi autonomamente risale al 1966, era stato ingegnerizzato presso il *Stanford Research Institute*. Nel 2014 la guida autonoma per i veicoli motorizzati su ruote è stata regolamentata tramite l'introduzione di una classificazione da 0, nessuna automazione, a 5, automazione completa. Nel 2018 Google ha iniziato ad offrire in Arizona, tramite *Waymo*[31], un servizio di taxi senza conducente, per poi espanderlo a Los Angeles e San Francisco nel 2023. Il software di queste auto incorpora sia dati provenienti dalla guida reale su strada sia quelli provenienti da milioni di ore di simulazione di guida virtuale. Tra i veicoli dotati di guida autonoma troviamo anche aerei, come per esempio vediamo nelle applicazioni militari dei droni e dei veicoli marini (barche o sottomarini[32]).

1.2 Il significato di "Intelligenza Artificiale"

Prima di tutto, vista la recente esplosione dell'IA, sgombriamo il campo da un equivoco diffuso: "IA", "apprendimento automatico" e "*deep learning*" non sono concetti da confondere tra di loro e non sottendono lo stesso significato. Il *deep learning* è un sottoinsieme del più generale apprendimento automatico, il quale viene ricompreso all'interno dell'insieme IA. Alcuni sistemi possono ricadere all'interno dell'"insieme IA" senza per questo dover includere algoritmi di *deep learning*. Definire che cosa si intenda per "intelligenza artificiale" rappresenta da sempre una difficoltà tra gli stessi esperti del settore, poiché non è scontato cosa si debba intendere con il termine "intelligenza" e quali siano le caratteristiche di un'intelligenza implementata ad una macchina e/o da un *software*.

La questione non riguarda solamente l'aspetto tecnico-ingegneristico ma rappresenta anche una questione filosofica: l'intelligenza si deve intendere come il solo risultato di processi chimico-fisici che avvengono all'interno del cervello mediante entità base, i neuroni, grazie ad una grande capacità di calcolo dei processi in parallelo, oppure come un tratto distintivo dell'essere umano che va oltre la sua configurazione puramente biologica, pur essendo influenzata da quest'ultima? A considerazioni metafisiche differenti corrispondono differenti risposte alla domanda.

Nell'immaginario comune il concetto di IA è spesso collegato all'universo della fantascienza, in cui robot intelligenti prendono il controllo della terra – o di una nave spaziale – e si rivoltano

[30] Cfr. Stanford's robotics legacy | Stanford Report
[31] Cfr. Waymo - Self-Driving Cars - Autonomous Vehicles - Ride-Hail
[32] Si veda per i droni marini, Dynamic robotic tracking of underwater targets using reinforcement learning | Science Robotics

contro gli umani, sfruttando le loro superiori caratteristiche prestazionali. Si pensi ad esempio a HAL 9000, il computer della nave spaziale *Discovery* nel film *2001: Odissea nello spazio*, o alla saga di *Terminator*. Tuttavia, tale visione dell'IA è molto distante dalle sue attuali capacità, per cui se in alcuni ambiti essa è in grado di esprimere risultati degni di nota – tra i quali la produzione di testi o il riconoscimento delle immagini –, in altri non è in grado neppure di compiere attività di cui un bambino di uno/due anni è capace di fare: ad esempio la manipolazione di oggetti o una competente interazione con l'ambiente circostante; in questi casi i robot appaiono goffi, indecisi ed estremamente lenti nelle azioni.

Come abbiamo visto, uno dei primi che ha cercato di uscire dalla difficoltà definitoria è stato Alan Turing, nel suo articolo *Can machine think?*[33] In cui propone la soluzione del "gioco dell'imitazione" per valutare se le prestazioni di una macchina possono essere messe a confronto con le capacità umane. Il problema definitorio viene così trasformato in un problema dialogico – e indirettamente viene tradotto in termini prestazionali.

Questa definizione "mimetica" dell'IA è quella che ha avuto maggior diffusione e che riscontriamo anche nelle definizioni correnti: riportiamo ad esempio quella del Parlamento europeo che definisce l'IA come «l'abilità di una macchina di mostrare capacità umane quali il ragionamento, l'apprendimento, la pianificazione e la creatività», che permette «ai sistemi di capire il proprio ambiente, mettersi in relazione con quello che percepiscono e risolvere problemi e agire verso un obiettivo specifico». Inoltre, «i sistemi di IA sono capaci di adattare i proprio comportamento analizzando gli effetti delle azioni precedenti e lavorando in autonomia»[34].

Più concisa la definizione della IBM, per cui «l'intelligenza artificiale è una tecnologia che consente a computer e macchine di simulare l'intelligenza e la capacità di risoluzione dei problemi degli esseri umani»: «da sola o in combinazione con altre tecnologie (ad esempio, sensori, geolocalizzazione, robotica) l'AI può eseguire compiti che altrimenti richiederebbero l'intelligenza o l'intervento umano»[35].

Riportiamo infine la definizione presentata nel documento della Commissione Europea curato, nel 2018, da una commissione indipendente di esperti di alto livello, per cui

---

[33] Turing, A. M., *Computing Machinery and Intelligence,* "Mind", vol. 59, no. 236, 1950, pp.433–460. *JSTOR*, http://www.jstor.org/stable/2251299.

[34] Cfr. Che cos'è l'intelligenza artificiale? | Tematiche | Parlamento europeo (europa.eu)

[35] Cfr. Cos'è l'Intelligenza Artificiale (AI)?| IBM, da notare che la ricerca sul sito americano di IBM (What Is Artificial Intelligence (AI)? | IBM) presenta una definizione differente: «l'intelligenza artificiale è una tecnologia che consente ai computer e alle macchine di simulare l'apprendimento umano, la comprensione, la capacità di risoluzione dei problemi, il processo decisionale, la creatività e l'autonomia» (traduzione mia)

«l'intelligenza artificiale si riferisce ad un sistema che mostra comportamenti intelligenti analizzando l'ambiente e prendendo decisioni – con un certo grado di autonomia – per raggiungere specifici obiettivi»[36].

Dalle definizioni sopra riportate notiamo, che esse più che dare una definizione teoretica di cosa significhi "intelligenza", ne presentano una pratica molto vicina a quella promossa da Turing: la macchina deve essere in grado di "mostrare" o di "simulare" in concreto atteggiamenti che l'uomo riconduce solitamente a sé stesso, in quanto essere intelligente. L'attenzione si sposta quindi sulla *performance* che la macchina è in grado di portare in quanto agente, sia essa solamente virtuale (bot) o anche dotata di un *hardware* di supporto (robot).

Per fornire un riassunto completo dei possibili approcci all'IA, utilizziamo la suddivisione presente in uno dei testi più completi sull'argomento: *Artificial Intelligence, a modern approach*[37] di S. Russell e P. Norvig (presente anche all'interno dell'enciclopedia di filosofia di Stanford[38]):

	Dimensione umana	**Razionalità ideale**
Basato sul ragionamento	Sistemi che pensano umanamente	Sistemi che pensano razionalmente
Basato sul comportamento	Sistemi che agiscono umanamente	Sistemi che agiscono razionalmente

La tabella presenta i possibili approcci che sono stati usati nello sviluppo dell'IA. In particolare, se seguiamo gli indicatori in neretto della prima riga orizzontale, si può intendere l'intelligenza come fedeltà alla prestazione umana oppure come formale esercizio di razionalità. Se invece si considerano gli indicatori in neretto della prima colonna a sinistra, l'intelligenza può essere indicata dalla correttezza del ragionamento oppure dal tipo di comportamento.

La combinazione di questi quattro modi di intendere l'intelligenza porta a differenti modi di sviluppare l'IA: per un approccio "umano" si useranno i risultati di una scienza empirica (es. psicologia), mentre per un approccio "razionalista" si useranno metodi provenienti dalla matematica e dall'ingegneria.

[36] Cfr. A definition of AI (europa.eu)

[37] S.J. Russell e P. Norvig, *Artificial Intelligence, A Modern Approach*, Third edition, Pearson Education Inc., New Jersey 2010, pp. 2-5.

[38] Cfr. Artificial Intelligence (Stanford Encyclopedia of Philosophy)

L'intersezione della casella in basso a sinistra – "l'agire umanamente" – è l'approccio proposto da Turing, di cui abbiamo già ampiamente discusso. Come indicano gli autori del testo in questione, gli studi rivolti a questo approccio non hanno avuto come fine principale la progettazione di macchine in grado di passare il test di Turing, costruite *ad hoc* per ingannare i giudici sottoposti al test; piuttosto, essi si sono impegnati nello studio dei principi base dell'intelligenza, per poi andare ad implementarli in opportuni modelli, i quali venivano poi sottoposti al test per verificare la validità degli studi stessi (e andare poi a correggerli e svilupparli ulteriormente).

Il secondo approccio è quello individuato nella casella in alto a sinistra, cioè il "pensare umanamente": su questo l'ambito interdisciplinare sono competenti le scienze cognitive. Per essere in grado di progettare e costruire programmi in grado di pensare come un essere umano è necessario anzitutto studiare e conoscere il funzionamento dei processi di pensiero propri dell'uomo. Abbiamo visto come già nelle prime applicazioni dell'IA, ad esempio nel *General Problem Solver*, l'intento non fosse solamente quello di risolvere correttamene i problemi ma, primariamente, di indagare i processi di pensiero con i quali gli umani giungevano alla soluzione, per poi riprodurli nel programma.

Durante i primi sviluppi dell'IA non fu raro che si confondessero gli approcci: ad esempio, dato un algoritmo in grado di eseguire in maniera efficace un'attività, si era portati a concludere che quello fosse un buon modello anche per l'esecuzione umana, o viceversa. Una chiara separazione dei due approcci ha consentito di migliorare lo sviluppo dell'IA e delle scienze cognitive, preservando il carattere di interdisciplinarità – ad esempio nello studio tra visione artificiale e neurofisiologia della visione. Al riguardo, gli attuali modelli basati su reti neurali artificiali che si occupano di riconoscere oggetti e simboli, sfruttano gli studi di neurofisiologia della visione, i quali indicano che l'informazione proveniente dalla vista non viene processata in un unico passaggio, ma ci sono diversi strati di neuroni che si occupano di affinare gli impulsi provenienti dalla retina, prima determinando i contorni degli oggetti e poi aggiungendo tutti gli altri particolari. Questi studi sulla visione sono stati tradotti in un opportuno modello computazionale che si è rivelato estremamente efficace rispetto ai metodi utilizzati prima.

Il "pensare razionalmente" è quell'approccio di tradizione "logicista"[39] che si occupa di indagare le leggi logiche del pensiero volte a individuare e formalizzare i processi di

[39] È così chiamata dagli studiosi dell'IA che a tale tradizione appartengono, ma non necessariamente per questo i filosofi che citeremo appartengono *stricto sensu* alla corrente "logicista".

ragionamento corretti. Comunemente, come anche gli autori del libro affermano, tale approccio è fatto risalire agli studi di Aristotele, che per primo fornì una sistematizzazione coerente della logica all'interno dell'*Organon*. Di particolare interesse per l'IA è la parte di logica del ragionamento, individuata nella figura del sillogismo, per cui a partire da una premessa maggiore e da una minore, legate tra loro da un termine medio, si arriva una conclusione valida. Perché essa sia anche vera, c'è bisogno che le entrambe le premesse siano vere.

È doveroso notare che questo approccio – quindi l'attenzione per la logica e il rapporto tra pensiero e logica – è stato foriero di sviluppi nei più svariati campi del sapere, a partire dalla filosofia, per passare alla matematica ed arrivare fino all'ingegneria. Questa tradizione comprende, oltre al già citato Aristotele, Leibniz, che intuisce la possibilità di creare una logica simbolica e un'algebra della logica. Quest'ultima sarà realizzata pienamente da Boole, cui si deve remotamente il linguaggio base di programmazione dei computer, sia dei primi che di quelli recenti.

Nello stesso alveo si inseriscono anche matematici del calibro di Frege, Cantor, Hilbert[40] e Gödel[41], che pone da parte sua un limite alle capacità dei computer di eguagliare la mente umana, riflettendo sui risultati dei suoi teoremi di incompletezza. Si può notare che Turing stesso ricade in questo approccio, visto che il tema del gioco dell'imitazione parte dalla possibilità di creare un calcolatore universale in grado di svolgere qualsiasi calcolo logico. Una visione di questo tipo considera la mente umana solamente dal punto di vista logico-razionale, ma tralascia aspetti altrettanto importanti, quali ad esempio le emozioni, le passioni o il senso comune, che appartengono alla nostra intelligenza in quanto umana.

L'ultimo approccio – "l'agire razionalmente" – è quello a cui aderiscono gli autori del libro. In questa prospettiva lo studio dell'IA è finalizzato alla costruzione di agenti in grado di interfacciarsi con le persone e con il mondo, che agiscono in modo razionale cioè, secondo i nostri autori, «in modo da ottenere il miglior risultato o, in condizioni di incertezza, il miglior risultato atteso»[42]. Il vantaggio di questo approccio è molteplice: in primo luogo può implementare all'interno dei suoi programmi una pluralità di linguaggi – logico, probabilistico, fuzzy[43], ecc… – e, in secondo luogo,

---

[40] Pensiamo ai 23 problemi irrisolti che Hilbert propose in occasione del Secondo congresso internazionale di matematica, tenutosi a Parigi nell'agosto del 1900. Chiamati *problemi di Hilbert*, alcuni vennero dimostrati indecidibili, cioè non risolvibili computazionalmente tramite un algoritmo.

[41] Per una prospettiva più completa si può consultare: M.Davis, traduzione di G.Rigamonti, *Il calcolatore universale*, Adelphi, Milano 2003.

[42] Cfr. *Intelligenza artificiale, un approccio moderno*, p.6.

[43] Come linguaggio logico intendiamo quello della logica di Boole, in cui un elemento può assumere solamente i valori 0 e 1. La logica fuzzy, che può essere considerata un'estensione di quella booleana,

è più “verificabile” nei suoi obiettivi: è più semplice mostrare che una agente artificiale ha agito razionalmente, assumendo come valida la definizione data poco sopra, rispetto a dimostrare di esser riusciti a imitare i processi umani di pensiero. La definizione di “agente” data è molto ampia, difatti qualsiasi programma per computer si interfaccia con delle persone e, ad esempio, anche una caldaia agisce mantenendo la temperatura costante all’interno dei caloriferi di casa.

Tuttavia, ciò non basta per ritenerli “agenti razionali”: quello che manca è la capacità attiva di adattamento rispetto ai cambiamenti esterni – che siano questi input umani o cambiamenti fisici dell’ambiente circostante. Inoltre, all’interno della categoria degli agenti razionali sarà presente una stratificazione a seconda del compito per cui sono progettati: un robot aspirapolvere è da ritenersi certamente meno raffinato di un’automobile a guida autonoma, pur condividendo alcune tecnologie di base; così come il suggeritore del telefono rispetto ad un programma come ChatGPT. Si può parlare anche di “agenti a razionalità limitata” qualora le informazioni per attuare la migliore azione non siano complete e/o non vi sia il tempo per svolgere tutti i calcoli necessari alla sua determinazione, per cui è necessario applicare un processo di approssimazione per arrivare al risultato atteso.

Dopo quanto abbiamo visto, possiamo fare alcune osservazioni conclusive: una definizione rigorosa e univoca di “intelligenza artificiale” non è ancora disponibile, ma presenta sfumature differenti a seconda degli approcci ma anche della cultura che si trova a definirla: come afferma P. Benanti[44], la nostra eredità filosofica greca ci porta a non fermarci al “come sembra”, ma ad andare più a fondo per comprendere “che cos’è”[45]; invece, in altre culture come quella giapponese o americana, questa necessità può non essere sentita e quindi l’urgenza di definire la differenza tra umano e artificiale in alcuni ambiti non si pone.

Abbiamo visto nella presentazione dei differenti approcci all’IA, in cui gli studiosi si pongono obiettivi differenti a seconda della concezione di partenza, che è sicuramente in parte determinata culturalmente. Russell e Norvig affermano che il paradigma (o modello standard) prevalente nello studio dell’IA è stato lo «studio e la costruzione di agenti che fanno la cosa

---

prevede invece che le variabili possano assumere tutti i valori compresi tra 0 e 1. L’uso dei metodi propri della teoria della probabilità si rende necessario quando la conoscenza del mondo non è certa – ad esempio nel caso di un’automobile a guida autonoma – e dev’essere attuata la migliore azione in presenza di incertezza.

[44] Cfr. Le frontiere dell’umano: i dilemmi dell’algoretica. Intervista a Paolo Benanti - Pandora Rivista

[45] Il test di Turing può richiamare alla mente il mito della caverna di Platone: il fatto che una macchina imiti le capacità o i comportamenti umani abbastanza bene da apparire umana non può bastare per risolvere la questione ontologica che sta alla base della questione. La capacità di imitazione è su un piano ben diverso rispetto ad un’ipotetica possibilità di riduzione dell’umano all’artificiale.

giusta»[46]: dapprima basato sulla logica e su piani definiti, per poi evolversi grazie agli studi intorno all'apprendimento automatico e alla teoria della probabilità applicata all'agire anche in situazione di incertezza. Possiamo però affermare che gli approcci visti in precedenza non hanno mai lavorato a compartimenti stagni, ma i diversi studi hanno avuto ricadute l'uno sugli altri; il metodo dell'agente razionale è diventato il modello di riferimento poiché lo studio sull'AI è nato sia come questione teoretica – indagare la possibilità di imitare i processi di pensiero umani – ma anche come una questione tecnica, ossia la replicabilità di tali processi attraverso una macchina.

Per concludere questa sezione riportiamo la definizione di IA elaborata da un gruppo indipendente di esperti di alto livello per la Commissione Europea nel 2018: essa è in grado di ricomprendere tutti gli approcci sopra presentati e di offrire una panoramica esaustiva di che cosa si intenda al momento per "intelligenza artificiale". Non è da escludere che con l'avanzare delle tecnologie tecniche la definizione possa essere nuovamente aggiornata:

> I sistemi di intelligenza artificiale (IA) sono sistemi software (ed eventualmente hardware) progettati dall'uomo che, dato un obiettivo complesso, agiscono nella dimensione fisica o digitale percependo il proprio ambiente attraverso l'acquisizione di dati, interpretando i dati strutturati o non strutturati raccolti, ragionando sulle conoscenze, o elaborando le informazioni derivate da questi dati e decidendo le migliori azioni da intraprendere per raggiungere l'obiettivo dato. I sistemi di IA possono usare regole simboliche o apprendere un modello numerico, e possono anche adattare il loro comportamento analizzando come l'ambiente è influenzato dalle loro azioni precedenti. Come disciplina scientifica, l'IA include diversi approcci e diverse tecniche, come l'apprendimento automatico (di cui l'apprendimento profondo e l'apprendimento per rinforzo sono esempi specifici), il ragionamento meccanico (che include la pianificazione, la programmazione, la rappresentazione delle conoscenze e il ragionamento, la ricerca e l'ottimizzazione), e la robotica (che include il controllo, la percezione, i sensori e gli attuatori e l'integrazione di tutte le altre tecniche nei sistemi ciberfisici).[47]

1.3 Le forme dell'intelligenza artificiale

Oltre ad una definizione di carattere generale, all'interno dell'ambito dell'IA è possibile individuare una gerarchia legata alle diverse capacità che i sistemi artificiali possono avere ed esprimere. Tutte le applicazioni di IA disponibili al momento sono capaci di svolgere bene uno specifico compito in uno specifico dominio di conoscenza: i software utilizzati in ambito medico aiutano nella prevenzione, nella diagnosi ed anche nel trattamento delle malattie, ma non

[46] Cfr. *Intelligenza artificiale, un approccio moderno*, p.6, dove giusto non è da intendersi in senso solamente morale: per IBM Watson il "giusto" era rappresentato da rispondere correttamente alle domande. Il giusto è relativo all'obiettivo che viene dato alla macchina e all'impiego operativo della stessa.

[47] Cfr. Ethics guidelines for trustworthy AI | Shaping Europe's digital future (europa.eu), alla voce *Definition of AI*.

possiedono alcuna capacità legata all'interpretazione del linguaggio, come ad esempio gli assistenti vocali o ChatGPT. Tutte queste applicazioni condividono, sia pure in modi differenti, i metodi e le applicazioni propri del *machine learning* e del *deep learning*, ma nessuna di queste è in grado di imparare le abilità dell'altra.

Nella letteratura tali sistemi rientrano all'interno della definizione di ANI: *Artificial Narrow Intelligence*, chiamata anche *weak AI* – IA debole. Anche se a volte i due significati possono venire usati in modo equivalente, questo non è tuttavia completamente corretto: il primo significato è inteso in modo tecnico, per cui il campo di applicazione dell'IA è sin da principio concepito come "ristretto" (*narrow*); il secondo invece incorpora anche una riflessione metafisico-ontologica, in cui la debolezza è rapportata alle capacità dell'intelligenza umana – che è quindi un'intelligenza "forte". Tutti i sistemi di IA attualmente esistenti rientrano all'interno della definizione di *weak AI*.

Si possono poi suddividere in due diverse sottosezioni: le *reactive machine AI* e le *limited memory AI*: le prime reagiscono solamente agli stati presenti del mondo, non hanno memoria e non possono pianificare eventi per il futuro – i programmi in grado di giocare a scacchi appartengono a questa categoria. Le seconde possono invece immagazzinare le informazioni passate ed usarle per agire o per monitorare l'evoluzione dei dati – il correttore automatico del cellulare ne è un esempio.

Il secondo tipo di IA che è possibile individuare è l'AGI: *Artificial General Intelligence*; anch'essa eredita le difficoltà incontrate nella definizione di IA, per cui, a seconda degli approcci cambia il modo di intendere la generalità di tale intelligenza artificiale: per alcuni studiosi si tratterà di creare una "intelligenza sintetica", per altri una "intelligenza computazionale", un'"intelligenza naturale", una "architettura cognitiva" o una "architettura cognitiva ispirata biologicamente"[48]. Ognuna di queste realizza differenti raccolte di concetti, approcci e prospettive, ma tutte condividono l'ipotesi di lavoro per cui

> la creazione e lo studio di una intelligenza sintetica con un ambito sufficientemente ampio (ad esempio a livello umano) e una forte capacità di generalizzazione, è qualitativamente diversa dalla creazione e dallo studio di intelligenze sintetiche con un ambito significativamente più ristretto e una capacità di generalizzazione più debole.[49]

[48] Cfr. B. Goertzel, *Artificial General Intelligence: Concept, State of the Art, and Future Prospects*, "Journal of Artificial Intelligence", vol. 5, n. 1, 2014, pp. 1-46.

[49] *Ivi*, p. 3.

In questo caso l'obiettivo è quello di sviluppare delle IA non limitate ad un singolo ambito della conoscenza – ad es. interpretazione del linguaggio scritto, riconoscimento delle immagini, interpretazione del linguaggio parlato, ecc.. – ma, partendo da una struttura contenente le informazioni di base, capaci di progredire autonomamente e di usare le conoscenze acquisite per formarsene di nuove a seconda delle richieste dell'ambiente – digitale o fisico che sia. Già nel 1950 Turing aveva proposto un approccio di questo tipo: «invece di provare a produrre programmi per simulare la mente di un adulto, perché non provare invece di produrne uno che simula quello di un bambino?»[50]. Dopo aver fornito la struttura di base al programma attraverso opportuni metodi di "educazione" e di insegnamento, esso sarà portato ad acquisire informazioni e poi ad evolversi autonomamente.

Uno degli obiettivi più alti a cui ambisce lo studio dell'AGI è la creazione di una "intelligenza artificiale generale di livello umano"[51], in grado non solo di svolgere qualsiasi capacità intellettuale normalmente riservata all'uomo[52], ma anche di essere in grado di agire autonomamente nel mondo, di interpretare ed esprimere emozioni, di interagire socialmente, e molto altro. Durante l'AGI Roadmap Workshop del 2009 è stata stilata una lista di 14 competenze[53] che l'AGI dovrebbe essere in grado di mostrare[54]. Al momento questo tipo di intelligenza artificiale esiste solamente a livello teorico come campo di ricerca e studio.

L'AGI può anche essere definita come "*strong AI"* (o IA forte) in contrapposizione alla precedente ANI, proprio per via di quella "generalità" di cui sopra si parlava.

Per completezza riportiamo anche l'ultimo tipo di IA che è possibile incontrare nella letteratura, cioè l'ASI – l'*Artificial Super Intelligence*. Essa sarebbe capace di autocoscienza e, grazie alle sue avanzate funzioni cognitive sarebbe in grado di sorpassare tutte le capacità intellettuali umane finora conosciute. Questo tipo di IA esiste solamente come definizione, ma non è oggetto di studio né di ricerca, mancando il tassello fondamentale precedente rappresentato dell'AGI.

---

[50] Cfr. *Computing Machinery and Intelligence*, op. cit., p. 456.

[51] Cfr. S. Adams *et alii*, *Mapping the landscape of human-level artificial general intelligence*, "AI Magazine", vol. 33, n. 1, 2012, pp. 25-42.

[52] Cfr. S. Groppe, S. Jain, *The way forward with AI-Complete Problems*, New Generation Computing, vol. 42, n. 1-5, 2024.

[53] Percezione del mondo, attuazione, memoria, apprendimento, ragionamento, capacità di pianificare, attenzione, motivazione, emozioni, capacità di modellare sé stessi e gli altri, interazione sociale, capacità di comunicare, capacità quantitative, capacità creative, cfr. *Artificial General Intelligence: Concept, State of the Art, and Future Prospects*, p. 10.

[54] *Ivi*, p.9.

1.3.1 Weak AI e Strong AI

Sebbene i concetti di *weak* e *strong AI* siano entrati in letteratura con i significati che abbiamo delineato nel paragrafo precedente, tali termini trovano posto con un significato differente nell'articolo di J. Searle, *Minds, Brains and Programs*[55]. In particolare per la concezione dell'IA debole – definita anche "prudente" dallo stesso autore –, il valore principale dell'uso del computer nello studio della mente risiede nell'essere uno strumento per testare le teorie e formulare ipotesi circa il funzionamento del cervello e della mente umana. Al contrario, per la concezione dell'IA forte, il computer non è un mero strumento che esegua un software ma, se opportunamente programmato, può essere considerato come avente una mente e quindi in grado di capire e di possedere stati cognitivi[56]. Queste sono le definizioni che dà Searle nel suo articolo per differenziare i due tipi di IA, che di fatto riprendono la questione chiave già posta da Turing precedentemente: "le macchine possono pensare?".

L'articolo si concentra nel dimostrare l'impossibilità dell'esistenza di una IA forte, in grado di avere una coscienza e in particolare di esser dotata di intenzionalità. Nella conclusione[57], l'autore si concentra in particolare nel rilevare il dualismo – di stampo cartesiano – che sottostà alla concezione dell'IA forte, per cui è sarebbe possibile replicare la mente umana tramite un raffinato programma, poiché si ritiene che essa non dipenda dalla "sostanza materiale" in cui risiede, ossia il cervello. Al contrario, l'emergere dell'intelligenza umana è possibile proprio grazie al sinolo[58] tra la mente e il corpo, che come tale è inscindibile; mentre un programma può essere eseguito più o meno velocemente su una grande varietà di macchine.

1.3.1.1 L'argomento della stanza cinese

Per criticare la possibilità che un programma di intelligenza artificiale possa avere intenzionalità e coscienza di quello che sta facendo, Searle propone un esperimento mentale[59]: in una stanza chiusa a chiave è presente una persona che non ha alcuna conoscenza di cinese né scritto né parlato; con sé ha solamente una serie di istruzioni scritte in inglese che gli consentono di mettere in relazione dei simboli cinesi con altri simboli cinesi. Dall'esterno gli viene consegnato un foglio con dei simboli cinesi e, attraverso le istruzioni, l'uomo è in grado di restituire un foglio in uscita con simboli cinesi diversi. Mano a mano che la persona impara a destreggiarsi con le

[55] Cfr. J. Searle, *Minds, Brains and Programs*, "The behavioural and brain sciences", Vol. 3 (3), 1980, pp. 417-457.
[56] *Ivi*, pp. 417-418.
[57] *Ivi*, pp. 423-424.
[58] Sinolo non è un termine usato da Searle, ma è esemplificativo della visione presentata nell'articolo.
[59] Cfr. *Minds, Machine and Programs*, pp. 418-419.

istruzioni e queste ultime diventano sempre più precise, il foglio in uscita sembrerà scritto da qualcuno che il cinese lo conosce bene, anche se ciò non corrisponde a verità: il traduttore nella stanza infatti manipola simboli di cui non conosce il significato. Sostituendo alla persona un computer possiamo affermare che esso non capisce quello che sta facendo, ma opera seguendo un programma su simboli che, per lui, non significano niente intrinsecamente. Il lavoro di traduzione si compie solamente sul piano sintattico ma non su quello semantico.

Un paio di considerazioni su quanto affermato in questo articolo che ormai deve fare i conti con lo sviluppo che ha avuto l'IA negli ultimi decenni: (1) se prendiamo come esempio emblematico ChatGPT – questo vale anche per Google Traduttore – non è più possibile affermare *tout court* che esso non lavori anche sul piano semantico: i programmi sono infatti progettati per ricreare il piano semantico tramite modelli matematici. (2) la critica della coscienza rimane invece ancora valida: anche un programma raffinato come ChatGPT non fa altro che manipolare dei simboli secondo metodi matematici e statistici, ma non per questo ha coscienza di ciò che dice o scrive. (3) Se inteso nella declinazione scolastica, il tema dell'intenzionalità[60] rimane ancora attuale: ai programmi di IA e ai computer che li implementano manca una sfera essenziale nelle loro rappresentazioni, la sfera ontologica. I programmi di visione artificiale riconoscono, ma non vedono; gli assistenti vocali ascoltano, ma non odono; ChatGPT manipola simboli, ma non pensa e non conosce, se come pensiero intendiamo quell'unità con originaria l'essere che troviamo affermata nel poema di Parmenide: «è infatti lo stesso pensare ed essere». L'ontologia dell'IA è invece di tipo "unidimensionale", per cui esistono solamente una serie di dati che non si riferiscono ad alcunché di altro.

[60] Nel campo della teoria della conoscenza possiamo fare riferimento alla differenza tra unità o identità intenzionale e unità o identità fisica; citiamo a questo proposito un passo di S.V. Rovighi, *Elementi di filosofia 1*, Editrice Morcelliana, Brescia, 2022, pp. 110-111: «fra alcune delle realtà che constano c'è identità reale, "fisica" (nel senso scolastico del termine), fra certe altre no. […] se, poniamo, *ho freddo, e desidero* avvicinarmi alla stufa, e sono *lieto* di ascoltare il concerto, […] e sono *preoccupato* […], questi sentimenti mi constano come *miei*. Ossia: costituiscono una identità fisica […] nel senso che constano come *modi di essere me*, sì che posso tradurre la mia esperienza dicendo: io, che *sono* infreddolito, *sono* lieto nella musica, ecc. Ma il colore rosso-mattone della stufa […] non mi è dato come un modo di essere me, non costituisce un'identità fisica. […] Non posso tradurre l'esperienza del rosso-mattone […] dicendo: io sono rosso-mattone. […] Ora, per distinguere l'unità che c'è tra me infreddolito […] dall'unità che c'è tra me infreddolito e il rosso che vedo, chiamo la prima unità o identità *fisica* o *reale*; chiamo la seconda unità o identità *intenzionale*. Nella identità intenzionale consiste la conoscenza. E noi abbiamo conoscenza perché abbiamo coscienza di queste due diverse unità».

2 AI, Machine learning e Deep learning

Allo sviluppo dell'intelligenza artificiale si è sempre accompagnata della questione sulla possibilità di programmare degli algoritmi che consentissero alle macchine di imparare dall'esperienza, che fossero cioè in grado di acquisire nozioni attraverso l'interazione con l'ambiente – inizialmente un ambiente di tipo fisico, poi un ambiente digitale. L'obiettivo era quello di slegare la macchina da una programmazione puntuale passo-passo, del tipo *if-then*, e di fare in modo che l'algoritmo riuscisse ad elaborare una soluzione in autonomia anche davanti a situazioni impreviste. Come vedremo, solamente le applicazioni più avanzate di *deep learning* sono in grado di soddisfare queste richieste di apprendimento e autonomia, mentre molte altre applicazioni utilizzano algoritmi più semplici di *machine learning*.

2.1 Il labirinto di C. Shannon

Spinto anche dal confronto con il collega A. Turing, Claude Shannon, conosciuto come il "padre della teoria dell'informazione", nel marzo del 1951 presentò presso la *J. Macy Jr. Foundation* a New York il progetto di una macchina in grado di risolvere labirinti

> this is a maze-solving machine that is capable of solving a maze by trial-and-error means, of remembering the solution, and also of forgetting it in case the situation changes and the solution is no longer applicable[61].

Il progetto della macchina comprendeva: (1) un labirinto metallico composto da venticinque quadrati 5x5 cm separabili da divisori mobili, (2) il sensore a forma di topolino in grado di muoversi autonomamente per il labirinto ed infine (3) un obiettivo mobile montato su un perno. Nella parte sottostante il labirinto erano stati posizionati settantacinque relè collegati tra di loro tramite un circuito, che all'attivazione si disponevano secondo due configurazioni, rispettivamente in grado di far passare oppure bloccare il flusso di corrente. La disposizione dei relè al primo avvio della macchina è casuale e viene definita *nominal position*.

Si procedeva poi a porre l'obiettivo in un punto a piacere all'interno del labirinto e il sensore in una parte differente; all'avvio della macchina il sensore, dal quadrato di partenza applicava la strategia esplorativa: a partire dalla parete nord del quadrato in cui si trovava, tramite un movimento antiorario testava tutte le altre pareti, individuando quale di queste fosse quella aperta; le pareti "aperte" o "chiuse" venivano memorizzate dal sistema cambiando lo stato dei

[61] Cfr. *Cybernetics. Transactions of the Eighth Conference March 15-16*, 1951, Macy Foundation, New York, reperibile su AITopics | Presentation of a Maze-Solving Machine

relè sottostanti. Individuata la parete aperta, il sensore proseguiva nel quadrato successivo e ripeteva la stessa procedura. Reiterando questo algoritmo, il sensore era in grado di arrivare fino all'obiettivo posto all'interno del labirinto; una volta raggiuntolo si fermava.

Se la macchina veniva lasciata alimentata e il sensore preso e spostato in una posizione differente da quella della partenza precedente, veniva attivata una seconda strategia: la *goal strategy* – per cui il sensore era in grado di leggere l'attivazione precedente dei relè e di arrivare rapidamente all'obiettivo, senza controllare nuovamente ogni direzione. Se invece il sensore veniva spostato e il percorso del labirinto modificato, disponendo le pareti in modo differente, il programma era in grado di tornare alla strategia esplorativa per evitare di entrare in un *loop*[62] senza via d'uscita. Attraverso l'esperienza e la memorizzazione, la macchina aveva appreso la strada verso l'obiettivo ed era in grado di modificare il proprio comportamento al mutare dell'ambiente esterno[63].

Dal confronto con altri colleghi vennero in luce i limiti del progetto, che permisero di tracciare le indicazioni di ricerca e sviluppo per i sistemi futuri. Tra queste osservazioni, possiamo individuarne tre di particolare interesse:

- il collega fisico e filosofo von Foerster chiese a Shannon se, dati due differenti percorsi per arrivare all'obiettivo, la macchina fosse in grado di scegliere quello migliore. La riposta fu che la scelta del percorso non rappresentava necessariamente la più breve, ma *probabilmente* la più breve fra le due. Per calcolare quale percorso minimizzasse il tempo di percorrenza, sarebbe stato necessario progettare una macchina molto più complessa e con molta più potenza di calcolo[64]. L'osservazione posta dal collega sollevava una questione molto importante, ossia se la macchina, oltre a trovare l'obiettivo, fosse anche in grado di determinare il percorso più veloce tra quelli disponibili, oppure scegliesse semplicemente il primo che riusciva a scoprire. Sul piano dell'apprendimento questo comporterebbe l'implementazione di un algoritmo comparativo: la macchina non solo apprende, ma è anche capace di scegliere quale sia l'azione più adatta da compiere. La risposta di Shannon mette in risalto una questione che si farà sempre più centrale con il procedere dello sviluppo tecnologico: l'aumento di conoscenze e abilità della macchina richiede un aumento della potenza

[62] È possibile visionare il filmato del funzionamento qui: Where did digital communication begin?

[63] Successivamente la macchina venne denominata Theseus, in onore del protagonista dell'omonimo mito greco.

[64] Cfr. *Cybernetics*, op. cit., p. 178.

computazionale disponibile. Essendo questa potenza limitata, bisognerà trovare un modo per coniugare apprendimento e capacità computazionale: ad esempio, i primi *software* che vinsero delle competizioni di scacchi non andavano a calcolare tutti i possibili risultati della mossa di un pezzo, ma avevano una conoscenza approssimata fino alle tre/quattro mosse successive.

L'evoluzione del risolutore di labirinti ai giorni nostri è da individuare nelle applicazioni come Google Maps, in grado di suggerirci la strada più veloce tra due punti su una mappa: applicazione resa possibile sia dallo sviluppo di algoritmi sempre più efficaci sia dall'aumento di potenza computazionale a disposizione.

- Una seconda questione, che approfondiva la precedente, venne dal matematico L.J. Savage: dato un percorso in grado di portare all'obiettivo, la macchina era in grado di *comprendere* se fosse il caso di cercare un nuovo percorso, magari più efficiente del precedente? Mentre la domanda precedente era relativa ad un problema di ottimizzazione del percorso, questa riguardava una "meta-capacità" della macchina, cioè quella di esser in grado di comprendere se fosse necessario trovare una strada differente da quella già nota. Con un parallelo circa l'apprendimento umano, Savage afferma: «se sei già in grado di pelare una patata, perché dovresti prenderti la briga di trovare un altro modo per sbucciarla? Forse la stai già pelando correttamente. Come fai a saperlo?[65]»

  Come vedremo quando parleremo degli agenti razionali, il problema qui posto è relativo alla conoscenza che la macchina ha dell'ambiente circostante: nel caso del topolino di Shannon il "mondo" a sua disposizione era individuato solamente dalle pareti aperte e chiuse. Inoltre, non avendo memoria dei passaggi precedenti mentre si inoltra nel labirinto, non ha la possibilità di valutare un percorso alternativo. Se la macchina avesse avuto a disposizione una telecamera posta sopra al labirinto e addetta a tracciare e calcolare tutti i possibili percorsi per arrivare all'obiettivo, allora avrebbe potuto comprendere quale sarebbe stato il più semplice.

  In modo molto semplificato, questo è quello che succede quando utilizziamo Google Maps o qualunque altra applicazione di navigazione che sia in grado di calcolare un percorso dal punto A al punto B e poi di aggiornarlo e cambiarlo a seconda degli imprevisti rilevati sul percorso iniziale (incidenti, traffico, ecc..). In questo caso, l'ambiente che ha a disposizione il *software* è l'intera cartografia all'interno della

[65] Cfr. *Cybernetics*, p. 179.

quale dove si muove l'auto: esso è quindi in grado di calcolare un percorso alternativo.

- La terza questione si concentra sulla rimozione dell'obiettivo dal labirinto: in tal caso il sensore entra in un *loop* potenzialmente infinito, in cui al completamento di un percorso ne viene cominciato un altro, cercando in tutti i quadrati, senza potersi fermare; solamente con la reintroduzione dell'obiettivo il processo di ricerca termina. Possiamo notare che il comportamento che la macchina esprime è molto lontano da quello che potrebbe essere il comportamento di un essere umano: in un'analoga situazione, dopo aver esplorato ogni angolo del labirinto, un essere umano potrebbe decidere di desistere dall'impresa e di sospendere la ricerca dell'obiettivo. Al contrario la macchina agisce seguendo la procedura prescritta dall'algoritmo ed entra in un *loop* di ricerca di durata virtualmente illimitata, a meno che qualcuno non reintroduca l'obiettivo o la macchina venga spenta dall'esterno. Come si vedrà più avanti, questo è una differenza fondamentale – metafisica – tra i *software* e l'essere umano: per il secondo il tipo di orizzonte è trascendentale, in grado cioè di ricomprendere e superare qualunque dato finito – nel nostro esempio di desistere dalla ricerca –, per il primo è puramente procedurale e quindi finito, delimitato dalle istruzioni contenute nell'algoritmo, ossia dal compito di ricercare nel labirinto qualcosa che, in questo caso, non potrà mai trovare.

Sebbene fosse solo uno dei primi tentativi di programmare una macchina in grado di imparare, la sperimentazione di Shannon diede il via a molti altri studi circa l'apprendimento delle macchine, dando via anche a delle competizioni internazionali di "risolutori di labirinti".

2.2 Apprendimento in IA

Il campo del *machine learning* rappresenta un sottoinsieme di quello dell'IA e si occupa in particolar modo della progettazione di *software* capace di imparare dall'esperienza e il quale non sia necessaria una programmazione puntuale di ogni passaggio, come invece richiesto dai programmi standard. Questo campo è intrinsecamente transdisciplinare includendo l'applicazione di nozioni provenienti dalla statistica, dalla teoria dell'informazione, dalla biologia, dalle scienze cognitive, dalla filosofia, dalla complessità computazionale e dalla teoria del controllo[66].

[66] Cfr. T.M. Mitchell, *Machine Learning*, McGraw-Hill, New York 1997, p. XV.

Prima di andare a vedere quali siano i diversi approcci alla conoscenza e i diversi algoritmi da impiegare, è necessario capire cosa si intende con il concetto di "apprendimento". In proposito, riporteremo due possibili definizioni. La prima è quella riportata da E. Pessa[67], che rappresenta la definizione generale fornita dagli psicologi comportamentisti:

> l'apprendimento è una modifica del comportamento, ovvero delle relazioni tra stimoli e risposte, che si instaura in seguito ad opportune stimolazioni esterne e non è riconducibile a processi di maturazione biologica, a malattie o all'assunzione di sostanze chimiche[68].

Come nota correttamente l'autore, questa definizione porta con sé tre questioni principali da affrontare, per poterla poi adattare al mondo dei *software* e delle macchine; in particolare bisogna definire: (1) cosa si intenda per "stimoli e risposte", (2) cosa si intenda con "relazioni tra stimoli e risposte"; ed infine (3) che tipo di relazione si instauri tra quest'ultimi. La prima questione è un problema relativo al livello di descrizione, per cui bisogna definire quali siano gli stimoli e quali le risposte all'interno di un software o di un robot; la seconda fa riferimento alla rappresentazione della conoscenza, per cui bisogna trovare un modo per rappresentare e mettere in relazione il binomio stimolo-risposta, ad esempio tramite un approccio simbolico oppure sub-simbolico; mentre il terzo punto chiama in causa la scelta dell'algoritmo di apprendimento, che definisce il tipo di relazione a cui la macchina è chiamata a rispondere.

Una differente definizione di apprendimento, più aderente al contesto dell'IA, è contenuta nel libro *Machine Learning*[69] di Tom M. Mitchell

> si dice che un programma per calcolatore apprende dall'esperienza E rispetto ad una certa di classe di compiti T ed a una misura di performance P, se la sua performance nei compiti T, misurate da P, migliorano con l'esperienza E.

Questa definizione è quella entrata ormai in uso comune all'interno dell'ambito del *machine learning*, in quanto è abbastanza ampia per far ricadere al suo interno le più differenti applicazioni di "programmi intelligenti": da quelli che si occupano di filtrare la spam a quelli implementati all'interno delle auto a guida semi-autonoma. È inoltre più semplice della precedente, perché è una definizione facilmente traducibile in termini operativi e verificabile attraverso applicazioni concrete. Per definire ottimamente un problema di apprendimento in

---

[67] Cfr. E. Burattini, R. Cordeschi (a cura di), *Intelligenza artificiale*, Carrocci editori, Roma 2001, pp. 97-98.

[68] Cfr. E. Burattini, R. Cordeschi (a cura di), *Intelligenza artificiale*, Carrocci editori, Roma 2001, pp. 97-98.

[69] T.M. Mitchell, *Machine Learning*, p. 2.

questo caso ci basta identificare tre variabili: la classe di compiti T che il programma è chiamato a compiere, una misurazione P delle performance e, per ultimo, la fonte dell'esperienza E.

Se prendiamo come esempio la macchina di Shannon presentata precedentemente, la classe T sarò rappresentata dal "ricercare l'obiettivo", la classe P potrà esser rappresentata dalla "rapidità" della risoluzione, mentre la classe E sarà il labirinto dove è inserito il sensore. La seconda volta che il sensore si muoverà nel labirinto, la sua *performance* sarà migliorata perché esso si dirigerà molto più rapidamente all'obiettivo ed avrà quindi appreso la strada corretta. Se l'obiettivo resta fisso in quel punto, il processo di apprendimento della macchina si ferma: perché essa ripeterà sempre lo stesso percorso e non ci sarà alcun miglioramento.

Vi sono principalmente due approcci all'apprendimento che vengono usati all'interno del *machine learning* e che si differenziano in modo netto tra loro: l'approccio simbolico e l'approccio sub-simbolico[70].

2.3 L'approccio simbolico

L'IA simbolica, conosciuta anche come *Good Old Fashioned Artificial Intelligence* (GOFAI), fu la prima ad essere sviluppata nella storia dell'IA e del *machine learning*. Essa usa la manipolazione di simboli tramite regole logiche ed il ragionamento deduttivo, per produrre conoscenza. Come esempio possiamo considerare due "sistemi esperti": DENDRAL o GPS (*General Problem Solver*). Entrambi erano dotati di una base di nozioni e di un motore inferenziale, in cui trovavano posto i connettivi e le leggi logiche. Considerato che tali sistemi, per rappresentare la conoscenza, lavorano sulla manipolazione dei simboli ma anche in formule espresse in linguaggio corrente, tali *software* sono facilmente interpretabili anche dall'uomo: uno dei principali vantaggi di tale approccio è così la trasparenza del comportamento della macchina, che può essere predetto e corretto entrando nel suo codice operativo. Il motore inferenziale era progettato e scritto nel dettaglio da programmatori ed era quindi possibile agire sul programma per modificarlo, ampliarne le conoscenze, correggerne gli errori e verificarne il corretto funzionamento.

In generale, il codice di questi software è facilmente leggibile e manipolabile, ed è possibile risalire alla causa di anomalie nel loro comportamento. La capacità di manipolare i simboli è

[70] Esistono anche degli approcci ibridi chiamati *in-between methods*, che cercano di combinare assieme l'approccio simbolico e quello sub-simbolico. Si veda E. Ilkou, M. Koutraki, *Symbolic vs Sub-symbolic AI Methods: friends or enemies?*, Conference on Information and Knowledge Management, 2020.

largamente utilizzata in tutte quelle applicazioni che hanno a che fare con il linguaggio umano: dal riconoscimento vocale alle traduzioni.

I sistemi di IA simbolici sono in grado di gestire problemi ben definiti e statici, tramite un processo di ragionamento rigoroso, ma non restituiscono risultati soddisfacenti qualora abbiano a che fare con grandi moli di dati o debbano produrre valutazioni dinamiche in tempo reale a partire da dati incompleti. Se ad esempio volessimo riconoscere il nostro gatto all'interno di un album di foto, un sistema simbolico sarebbe in grado di riconoscere solamente la foto su cui l'abbiamo addestrata, poiché tutte le altre sarebbero differenti da questa: i pixel della foto di prova non corrisponderebbero con quelli delle altre foto. Il salto nella visione artificiale è stato fatto con l'introduzione di quelle reti neurali che si basano su un approccio differente: quello sub-simbolico.

2.4 L'approccio sub-simbolico

Mentre l'IA simbolica affonda le sue radici nella matematica e nella logica, quella sub-simbolica nasce all'interno del contesto delle neuroscienze, in particolare nello studio delle "percezioni veloci", quali il riconoscimento di volti o del linguaggio parlato. La conoscenza secondo questo approccio non sta nei simboli – tanto che i programmi di questo tipo non contengono un linguaggio *human-friendly*, ma solamente un'enorme quantità di formule matematiche – quanto piuttosto nei dati che vengono processati da strutture complesse, composte da unità semplici – micro-unità cognitive – chiamate neuroni artificiali. La conoscenza per questi sistemi è quindi il risultato di queste interconnessioni e di operazioni matematiche sulle stesse a partire dai dati in input e dal tipo di risultato che si vuole ottenere. Una differenza importante rispetto ai sistemi simbolici visti precedentemente è che quelli sub-simbolici, prima di venire addestrati sui dati, non contengono alcuna conoscenza, ma solamente dei valori e dei parametri inizializzati in modo standard. Inoltre, soffrono del problema di non poter essere trasparenti: dato un certo input è difficile, o quasi impossibile, capire come il sistema sia arrivato ad un certo output: lo vedremo quando andremo ad analizzare nel dettaglio il loro funzionamento. Il principale risultato ottenuto tramite l'approccio sub-simbolico è quello delle reti neurali, che hanno contribuito allo sviluppo esplosivo dell'IA negli ultimi decenni.

2.4.1 Com'è fatto un neurone artificiale

Alla base della rete neurale ci sono le unità di base, i neuroni artificiali, che rappresentano una modellizzazione – e per molti versi una semplificazione – del funzionamento dei neuroni biologici. In fig.1[71] possiamo vedere come è composto un neurone biologico: alle estremità di sinistra vi sono i dendriti – fibre che portano i segnali in arrivo al corpo centrale del neurone – mentre a destra vi è l'assone che, attraverso il collegamento con un altro neurone tramite la sinapsi, farà da canale per trasmettere l'impulso che partirà dal soma. Se l'eccitazione proveniente dai segnali in entrata eccede un certo valore di potenziale, chiamato potenziale di soglia[72], il neurone emetterà un segnale in uscita ed entrerà in uno stato refrattario, per cui non sarà in grado di emettere lo stesso segnale per un certo periodo di tempo. Quando un canale di input è molto utilizzato, diventa potenziato rispetto agli altri ed è così in grado di attivare con più facilità il neurone. Ciascun neurone possiede inoltre un numero di sinapsi tra 5000 e 100000, ed è quindi collegato con un enorme numero di altri neuroni, formando una rete intricatissima[73].

Il neurone artificiale è costruito in modo simile[74], per cui nella parte sinistra trovano posto i

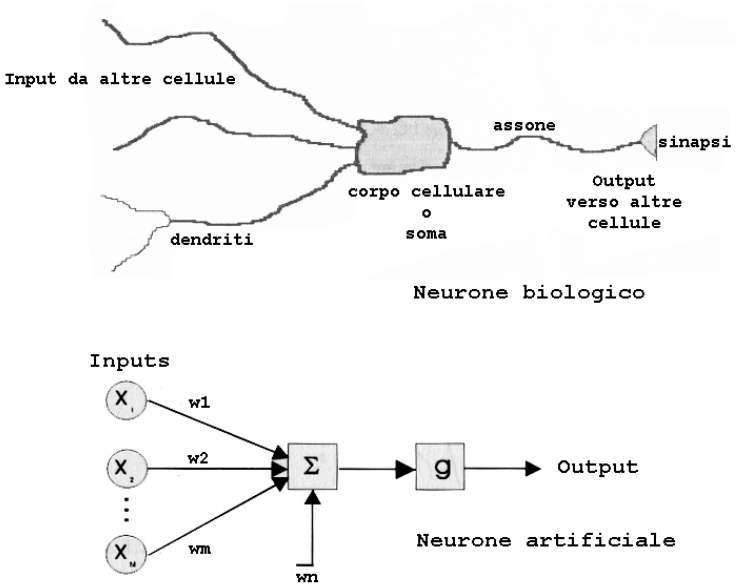


Figura 1

vari input $(x_1 \dots x_n)$ a cui vengono associati dei pesi $(w_1 \dots w_n)$, che rappresentano la "forza"

[71] Il neurone artificiale rappresentato in fig.1 è chiamato anche neurone di McCulloch e Pitts, i primi che hanno costruito un modello artificiale di neurone basato su quello naturale.

[72] Nel processo eccitatorio, da una differenza di potenziale base di -70 mV, si ha un innalzamento a -50 mV.

[73] Il numero di neuroni stimato nel cervello umano è pari a 10^{11}, con una quantità di sinapsi pari a 10^{15}.

[74] Cfr. B. Codenotti, M. Leoncini, *La rivoluzione silenziosa*, Codice edizioni, Torino 2020, pp. 132 e ss.

della connessione di quell'input rispetto al neurone. Essi possono assumere un qualsiasi valore reale: positivo (connessioni eccitatorie), negativo (connessioni inibitorie) o nullo (assenza di connessioni)[75]; in un processo di apprendimento sono proprio questi pesi che vengono modificati per consentire alla rete di imparare. Nelle costruzioni più semplici di neuroni artificiali, tali input vengono sommati, moltiplicati per i rispettivi pesi $\sum_{k=1}^{n} x_k w_k$, e tale sommatoria deve essere maggiore di un certo valore di soglia S. Sia questo valore che quelli dei pesi possono essere decisi dal programmatore; in caso positivo il neurone si attiverà e l'output sarà pari ad 1, in caso contrario l'output sarà pari a 0. Una funzione $f(x)$ di questo tipo è chiamata funzione a gradino (*step function*), con equazione:

$$f(x) = \begin{cases} 0 \; se \; x < S \\ 1 \; se \; x \geq S \end{cases}$$

Dove *x* è pari a $\sum_{k=1}^{n} x_k w_k$ [76].

Come vedremo la funzione gradino non è adatta alle applicazioni più avanzate e complesse, poiché opera un taglio troppo drastico nel determinare il valore di attivazione. A differenza dei neuroni naturali, i neuroni artificiali non presentano un periodo refrattario alle stimolazioni multiple.

Utilizzando questi neuroni artificiali e disponendoli opportunamente, si vengono a creare le cosiddette reti neurali, che si distinguono a seconda del numero di neuroni e della profondità della rete: variabili in funzione del compito che ciascuna rete è chiamata a svolgere. I principali vantaggi dell'utilizzo delle reti neurali stanno nella loro scalabilità, nella loro modulabilità e nella loro capacità di compiere moltissimi calcoli in parallelo e non solo in serie.

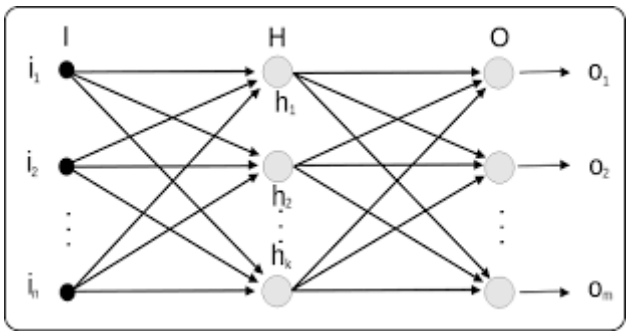


Figura 2

[75] Cfr. E. Burattini, R. Cordeschi (a cura di), *Intelligenza artificiale*, op. cit., p. 99.

[76] È possibile trovare in letteratura una diversa definizione, per cui x è pari a $\sum_{k=1}^{n} x_k w_k - S$, che è per altro equivalente a quella data da noi.

In fig.2 possiamo vedere un esempio di rete neurale semplice, composta da uno strato di neuroni H, che fanno parte dello strato nascosto (*hidden layer*) della rete, e dagli strati di input I e di output O. È da notare che né lo strato di input né lo strato di output sono a rigore da ritenersi composti da neuroni artificiali, perché non compiono alcun calcolo sui dati. Se ad esempio in entrata abbiamo un'immagine composta da 36 pixel, il compito dello strato I sarà quello di ripartire le informazioni di ciascun pixel ad ognuno dei neuroni che stanno al centro, che poi provvederanno a fornire il risultato allo strato di output.

Nella figura le connessioni tra gli strati sono solo di tipo *feed-forward*, ossia aventi un unico verso dall'input verso l'output; alcuni tipi di reti posso avere anche delle connessioni *feedback* tra gli strati – come ad esempio le BAM, *Bidirectional Associative Memory*[77]–, mentre altri tipi di rete possono presentare connessioni tra i neuroni dello stesso strato. Le reti neurali con più strati di neuroni vengono denominate reti neuronali profonde e solamente in quel caso si può parlare di *deep learning*. Attraverso un processo di propagazione all'indietro dell'errore (*error back propagation*), la rete neurale è in grado di aggiornare il valore dei pesi tra i vari strati ed è quindi in grado di apprendere, riducendo l'errore tra il valore di output previsto e quello calcolato.

[77] Cfr. E. Burattini, R. Cordeschi (a cura di), *Intelligenza artificiale*, p. 110. Le reti che fanno uso di queste connessioni vengono chiamate RNN – *Recurrent Neural Network* – per cui l'output è influenzato anche dai dati di input. Una loro applicazione si trova nelle applicazioni per il riconoscimento vocale che usano una tecnologia chiamata *long short-term memory* (LSTM), che grazie alla funzione di memoria sono in grado di comprendere il linguaggio parlato, il contesto di una richiesta e di effettuare traduzioni in tempo reale efficaci ed affidabili.

2.5 Tipi di apprendimento

Esistono principalmente tre differenti modi attraverso i quali i *software* di IA possono essere addestrati per imparare: vi è l'apprendimento supervisionato, il semi-supervisionato e quello non supervisionato; e ciascuno di questi può utilizzare diversi tipi di algoritmi di apprendimento.

2.5.1 Apprendimento supervisionato

Si tratta del più diffuso metodo di *machine learning* utilizzato: quello che da i risultati migliori. Importante all'interno del processo di apprendimento è l'intervento umano, specialmente nella fase di preparazione dei dati da fornire al *software*. Il *dataset* su cui lavora il modello – definito *dataset* "etichettato" (*labeled dataset*) – è infatti composto da due diversi tipi di dati: vi è l'insieme di dati di input $(x_1 \dots x_n)$ che verranno forniti alla macchina e ai quali è associato un insieme di etichette $(\hat{x}_1 \dots \hat{x}_n)$ che rappresenta l'output corretto per ogni rispettivo dato di input. Nel processo di apprendimento il *software* agirà sui dati di input, ad esempio su x_1, e produrrà un output y_1 che andrà poi a confrontare con il valore corretto $\hat{x}_1$. La distanza (definita in modo matematico) tra questi due valori consentirà al modello di aggiornare i parametri e di migliorare la prestazione con i dati successivi. Se ad esempio un *software* deve essere addestrato a riconoscere quali immagini rappresentino un cane e quali un gatto, ad ogni immagine in entrata verrà associata l'etichetta "cane" o "gatto" corretta. All'inizio del processo di apprendimento il *software* parte come una tabula rasa e non sarà in grado di riconoscere correttamente le immagini che gli vengono sottoposte; grazie alle etichette sarà però in grado di stimare l'errore compiuto e, aggiornando i parametri interni, di migliorare le sue capacità di riconoscimento per le immagini successive.

In questo processo è quindi importantissima la qualità dei dati presenti all'interno del *dataset*: essi devono essere completi e consistenti – riferendoci al nostro esempio, devono essere immagini di buona qualità e che rappresentino effettivamente "cani" e "gatti" – e normalizzati, evitando *outlier* che influenzerebbero il modello in maniera negativa. Normalmente vi sono due differenti *dataset*: il primo è il *training set*, che contiene i dati utilizzati per la fase di addestramento del *software*; il secondo è il *test set* che contiene i dati su cui verificare le prestazioni dopo l'addestramento; questi ultimi dati non devono mai esser stati analizzati dal modello, per poter fornire così un risultato valido delle *performance* reali. Dopo aver superato queste due fasi, il programma può essere utilizzato su nuovi dati.

Ci sono due modelli principali di applicazione dell'apprendimento supervisionato: la classificazione e la regressione – lineare o logistica. La prima si occupa di riconoscere e raggruppare i dati in ingresso in categorie predefinite, come accade ad esempio nei programmi di gestione *spam* o di fidelizzazione dei clienti; la regressione lineare si applica invece a problemi di previsione continua, dove si tratta di modellizzare la relazione tra una variabile indipendente e una (o più) variabili dipendenti. La previsione delle vendite di un'azienda sulla base di pubblicità, costi e promozioni può costituirne un esempio.

Per la natura del loro apprendimento, queste applicazioni non sono in grado di trovare pattern nascosti all'interno dei dati e correlazioni tra gli stessi: questi *software* cercheranno solamente ciò che hanno imparato a discriminare.

2.5.2 Apprendimento non-supervisionato

A differenza del precedente, questo metodo non utilizza su dati etichettati, bensì utilizza un *software* che cerca di ricostruire strutture o *pattern* che sono intrinseci all'interno dei dati. I suoi principali campi di applicazione sono: identificazione dei pattern, riduzione della dimensionalità, *clustering*. Il primo è utilizzato ad esempio dalle aziende, per far emergere all'interno di una grande quantità di dati relativa ai clienti le correlazioni, associazioni o tendenze tra i vari dati. Il *clustering* si occupa invece di dividere i dati simili in categorie non specificate a priori. L'ultimo – la riduzione dimensionale – semplifica i dati mantenendo le caratteristiche più rilevanti ed eliminando variabili superflue.

Essendo questi metodi applicati alle reti neurali, una loro criticità risiede nella mancanza di trasparenza e nella difficoltà di interpretare i risultati che il modello produce: non c'è modo di sapere la ragione secondo cui i dati sono stati divisi o raggruppati in un certo modo, e per questo i risultati possono alle volte essere affetti da *bias* impliciti nei dati di partenza[78].

2.5.3 Apprendimento semi-supervisionato

Questa tecnica si pone a metà tra le precedenti; infatti, in questo approccio un modello viene addestrato utilizzando una combinazione di dati etichettati (come nell'apprendimento supervisionato) e di dati non etichettati (come nell'apprendimento non supervisionato), dove

[78] Come esempio più conosciuto citiamo il sistema COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) utilizzato in America che, come fatto notare da alcuni ricercatori nel 2016, stimava come a maggior rischio di recidiva specialmente le persone ispaniche o di colore.

questi ultimi rappresentano la maggior quantità di dati disponibili; è perciò necessario che i dati non etichettati siano di qualità e pertinenti a ciò che andrà a fare il modello, altrimenti buona parte del processo di apprendimento andrebbe vanificato. Se ad esempio vogliamo addestrare un modello a distinguere l'immagine di una macchina da quella di una motocicletta, è necessario che i dati non etichettati per la maggior parte presentino effettivamente immagini con questi veicoli, pena l'inutilità dell'addestramento.

Questo approccio è utilizzato in particolare nella diagnostica medica per immagini, in quanto consente di etichettare un numero relativamente limitato di dati e poi fornire al modello una grande quantità di dati di qualità e ristretti all'ambito della diagnostica per immagini.

Uno dei vantaggi di questo metodo è sicuramente un risparmio di risorse e di tempo – quello che invece sarebbe richiesto per l'etichettatura di tutti i dati, come accade nell'apprendimento supervisionato –; esso richiede in compenso molta precisione nella etichettatura dei pochi dati di partenza.

2.5.4 Apprendimento umano e artificiale

Nonostante i neuroni artificiali si ispirino e prendano come modello il funzionamento dei neuroni biologici, bisogna fare attenzione a non operare un processo di riduzione dell'umano all'artificiale. Infatti, tali modelli nel loro sviluppo attuale non sono ancora comparabili alle capacità dei neuroni presenti nel nostro cervello[79]; sul piano del funzionamento, in quanto i segnali emessi dai neuroni naturali – chiamati *spike* – sono il risultato di alterazioni di tipo chimico-fisico; inoltre, l'attivazione dei neuroni è asincrona, cioè non soggetta ad una sincronizzazione reciproca globale. Ciò vuol dire che all'arrivo di un impulso nervoso, come può essere quello proveniente dall'occhio destro, non si attivano tutte le 10^{15} sinapsi che collegano tutti i neuroni tra di loro, ma solamente le connessioni deputate all'elaborazione del segnale visivo e, in caso, della memoria, per associare ciò che è visto a ciò che è già conosciuto.

Ad una particolare azione, come il riconoscimento di una persona, vengono attivati solamente quei neuroni che si occupano di associare il viso riconosciuto ad un'azione di saluto, ma non l'intera connessione neuronale, per cui, ad esempio, non si attivano i neuroni incaricati di portare al cervello le informazioni relative al tatto. Ciò consente al cervello di essere estremamente efficace dal punto di vista energetico, al contrario di quanto si verifica nelle attuali applicazioni

[79] D. Ielmini, *Intelligenza artificiale: l'approccio neuromorfico*, "Mondo Digitale", Milano 2018, pp. 4 e seguenti.

di IA che richiedono ingenti risorse dal punto vista logistico, energetico, idrico. Alcuni ricercatori tentano attualmente di riprodurre le proprietà fondamentali di calcolo presenti nel cervello attraverso lo studio di circuiti neuromorfici.

Sul piano dell'apprendimento è chiaro che il cervello umano è qualitativamente e quantitativamente superiore rispetto a quello artificiale: mentre per un bambino di quattro anni è sufficiente vedere una volta sola un elefante allo zoo per poi riuscire a riconoscerlo una seconda volta, sia esso vivo o rappresentato in un disegno, per i modelli supervisionati sono invece necessarie migliaia o milioni di immagini di elefanti in tutte le possibili angolazioni ed sposizioni, per poi riuscire a riconoscere l'immagine di un elefante con un'efficacia che si approssimi al 100%.

2.6 *Deep learning* e reti neurali artificiali

Nel paragrafo 2.4.1 abbiamo visto come è costruito e come funziona un neurone artificiale semplice: quello che fu presentato per la prima volta[80] da F. Rosenblatt nel 1958 per il riconoscimento e la classificazione di forme. I risultati non furono però all'altezza delle aspettative ed il progetto si arenò per mancanza di fondi, ponendo in pausa il progetto delle reti artificiali e dell'approccio sub-simbolico. Si è dovuto aspettare il 1986, con la pubblicazione di *Learning representations by back-propagating errors* di D.E. Rumelhart, G. Hinton e R.J. Williams, affinché lo studio delle reti neurali riprendesse e grazie, all'algoritmo di retro-propagazione dell'errore, raggiungesse risultati degni di nota. L'esplosione di questi metodi ci fu nella prima metà degli anni 2000[81], grazie allo sviluppo di processori e di schede grafiche sempre più potenti, ma anche alla disponibilità di una componente essenziale per far funzionare questi modelli: l'accesso a grandi quantità di dati – meglio conosciuti con il nome di *Big Data* – per l'addestramento.

Solamente per le reti neurali multistrato (*deep neural networks*, aventi cioè più di uno strato nascosto), il tipo di apprendimento viene chiamato *deep learning*, dove la "profondità" (*deep*) si

[80] Un modello precedente ed ancora più semplice era quello presentato da McCulloch e Pitts, da cui Rosenblatt prende spunto per la sua versione.

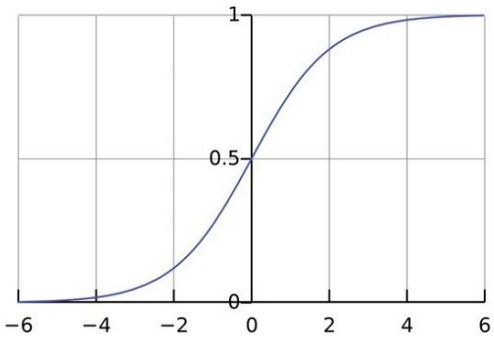
[81] Nel 2007 viene lanciato Google Translate; poco dopo in California compaiono le prime auto a guida autonoma, sempre di proprietà di Google. Nel 2011 il computer Watson della IBM vince la competizione televisiva *Jeopardy*. Cfr. M. Mitchell, *Artificial intelligence, a guide for thinking human*, Penguin Books Ltd, Londra 2020, pp. 37-40.

riferisce solamente al numero degli strati della rete e non ad una qualità particolare dell'apprendimento.

I neuroni che compongono queste reti differiscono solo leggermente da quelli presentati all'inizio del capitolo, specialmente nel tipo di funzione di attivazione e nell'introduzione del valore detto *bias*, che sostituisce il valore di soglia *S*. Vedremo tra poco quale funzione ricopra questo nuovo parametro.

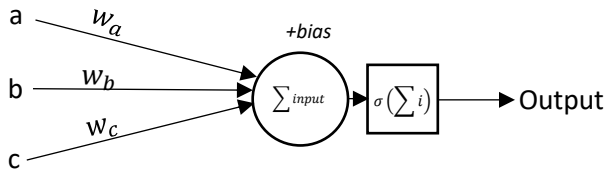
2.6.1 *Il neurone sigmoidale*

Si tratta del modello più utilizzato per la costruzione delle reti e, a differenza del percettrone presentato precedentemente, la sua funzione di attivazione $f(x)$ non è più quella a gradino, ma è sostituita dalla funzione sigmoidale $\sigma(x)$:



Dove $\sigma(x) = \frac{1}{1+e^{-x}}$ in cui compare la funzione e^{-x}, dove la variabile x rappresenta sempre la somma pesata degli input con l'aggiunta di un bias b. Come si può notare dal grafico, la funzione sigmoidale è una versione "addolcita" di quella a gradino e, cosa molto importante, è continua e derivabile su tutto il dominio[82]: condizione necessaria per l'utilizzo del metodo della propagazione all'indietro dell'errore, che richiede le derivate di tale funzione.

Consideriamo come esempio un semplice neurone con tre input (a, b, c) a cui sono associati i relativi pesi (w_a, w_b, w_c):



[82] La funzione a gradino presenta invece una discontinuità nel punto $(0,0)$, dove avviene appunto il passaggio da 0 a 1.

Come di consueto, l'output sarà caratterizzato dalla somma degli input moltiplicati per i rispettivi pesi con l'aggiunta di un nuovo valore b, chiamato *bias*:

$$x = a \cdot w_a + b \cdot w_b + c \cdot w_c + b$$

A questo punto, per avere il valore in output, valutiamo la somma ottenuta inserendola nella formula della sigmoide:

$$\sigma(x) = \sigma(aw_a + bw_b + cw_c + b)$$

Possiamo notare immediatamente che, grazie all'utilizzo di questa funzione, l'output del neurone artificiale non sarà più limitato ai due valori binari 0 e 1, ma potrà assumere tutti i valori compresi tra l'intervallo [0,1]. Inoltre, per valori di x molti grandi o molto piccoli, il comportamento della sigmoide è analogo a quello della funzione gradino: per x molto grande, la funzione e^{-x} tende a 0 e la sigmoide tende al valore 1; per x molto piccolo, la funzione e^{-x} assume valori enormi e dunque il valore della sigmoide si avvicina a 0. Il *bias* è un valore che influisce sulla facilità con cui la funzione σ riesce ad assumere il valore 1, e può assumere sia valori positivi che negativi.

Il *bias*, quindi, consente una maggiore modulabilità dell'output del neurone, poiché consente di determinare con quanta facilità o difficoltà un neurone deve attivarsi a seguito della ricezione di un input; infatti ogni neurone della rete incorpora un proprio valore di *bias*, che sarà diverso dagli altri. Se un neurone ha un valore di *bias* basso si attiverà più facilmente rispetto ad uno con un valore più alto.

Le differenze più importanti rispetto al valore di soglia *S* sono due: (1) nel percettrone classico, superato il valore di soglia *S,* il neurone si attiva con output 1, mentre nel neurone sigmoidale il superamento del valore di *bias* non determina necessariamente un'attivazione completa. (2) Il *bias* è una variabile e, come tale, anch'essa lei viene modificata dal *software* durante la procedura di apprendimento. In questo modo è il *software* in autonomia a determinare quali neuroni debbano attivarsi di più e quali meno. Il valore di soglia, invece, è modificabile solamente dal programmatore.

Il principale vantaggio dell'utilizzo della funzione σ si apprezza per valori interni all'intervallo [0,1], in quanto a piccole modifiche del valore dei pesi e del *bias* varia in modo analogo il valore dell'output. Questo è di importanza cruciale per gli algoritmi di apprendimento, che sfruttano proprio la regolazione fine di tali valori per consentire ai programmi di migliorare la loro *performance*.

2.6.2 *Come apprende una rete neurale*[83]

Come abbiamo avuto già modo di vedere, il funzionamento di una rete neurale e degli algoritmi di apprendimento è radicalmente diverso rispetto ai metodi classici di programmazione e alla logica degli algoritmi standard.

Infatti, l'apprendimento di una rete neurale avviene in modo autonomo tramite la regolazione fine dei pesi e dei *bias*, rispondendo agli stimoli esterni, senza che vi sia l'intervento diretto di un programmatore ad ogni passaggio del codice.

Di seguito mostreremo in che modo tali algoritmi sono progettati e qual è la logica di massima che soggiace a quasi tutte le applicazioni di *deep learning*. Per fare ciò supponiamo che il nostro obiettivo sia quello di addestrare una rete neurale a riconoscere i numeri che vanno da 0 a 9, scritti a mano da una persona qualunque; all'inizio del processo i valori di *bias* e pesi saranno inizializzati a dei valori casuali e quindi la rete non sarà in grado di riconoscere correttamente i numeri, classificandoli erroneamente.

Tramite un procedimento di apprendimento supervisionato, la rete sarà in grado di correggere gradualmente i valori di pesi e *bias*, valutando l'errore commesso rispetto al valore corretto, e sarà quindi in grado di imparare e migliorare la precisione nel riconoscimento dei numeri scritti a mano. Questo è possibile solamente grazie all'utilizzo dei neuroni sigmoidali, in quanto nei neuroni con funzione a scalino – i percettroni – la regolazione fine dei valori non è possibile, data l'oscillazione netta dei valori di output tra 0 e 1. Per una rete neurale di percettroni è quindi molto difficile apprendere[84].

Vediamo ora più in dettaglio come funziona il processo di apprendimento, avvertendo che verranno introdotti concetti matematici essenziali per comprenderne il funzionamento.

Per prima cosa definiamo il tipo di input e come sarà organizzata la rete neurale con cui andremo a lavorare: come dati in entrata avremo delle immagini di numeri scritti a mano, che verranno convertite in immagini digitali della dimensione di 28×28 pixel – per un totale di 784 dati di input –; la rete neurale invece presenta due strati nascosti di neuroni e lo strato di output con 10 possibili uscite, da 0 a 9. Se per esempio l'input x rappresenta il numero 9, l'output $a(x)$

[83] Per questa parte faremo riferimento a B. Codenotti, *La rivoluzione silenziosa*, pp. 132-153. Un'altra fonte da cui prendiamo spunto, che presenta molti materiali interessanti, è raggiungibile dal seguente link: 3Blue1Brown - But what is a Neural Network?

[84] Ed è per questa ragione che le prime applicazioni dei percettroni non furono in grado di soddisfare le aspettative. Il procedimento di apprendimento era lento e limitato al riconoscimento di poche forme.

sarà individuato dal vettore $(0,0,0,0,0,0,0,0,0,1)$, se in input c'è il numero 3, l'output $a(x)$ sarà $(0,0,0,1,0,0,0,0,0,0)$ e così via.

Ai fini della nostra discussione non è essenziale precisare il funzionamento degli strati nascosti, ma supponiamo che la rete ne abbia due con 16 neuroni ciascuno. Tutti i pesi ed i bias della rete vengono inizializzati a valori casuali all'inizio del processo di apprendimento.

Il tipo di apprendimento di cui farà uso questa rete neurale sarà quello di tipo supervisionato, quindi il programma lavorerà su un dataset etichettato, in cui ad ogni vettore di input x_i – con 784 elementi – è associato il vettore y_i – di dimensione 10 – con il risultato corretto.

Indichiamo tale dataset nella forma $[(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)]$, dove n indica i numeri di dati che abbiamo a disposizione. Visto che la rete presenta valori inizializzati in modo casuale, i risultati a_i del vettore di output saranno molto distanti dal valore corretto y_i; la strategia per far "capire" alla rete di quanto sta sbagliando rispetto al dato corretto sta nel quantificare l'errore commesso, attraverso la definizione di una funzione chiamata funzione di costo (*cost function*), che indicheremo con la lettera C.

Esistono diverse funzioni di costo, a seconda del tipo di problema che si andrà a trattare – regressione, classificazione, classificazione multi-classe – e anche dell'architettura della rete. Per il nostro esempio la funzione C sarà definita tramite l'errore quadratico medio.

Consideriamo un dato di training (x_i, y_i) ed il suo corrispettivo output a_i, allora l'errore commesso su un singolo dato corrisponderà alla formula

$$|y_i - a_i| = \sqrt{\sum_{j=1}^{10} (y_{i,j} - a_{i,j})^2}$$

Dove $i = 1, \dots, n$ rappresenta il numero di dati di prova che la rete ha a disposizione. Per capire meglio cosa sta succedendo, procediamo a visualizzare il processo attraverso un'immagine, prendendo per esempio il risultato del riconoscimento del numero 8:

$$
\text{Costo del numero 8}\left\{\begin{array}{lll}
 & a_8 & y_8 \\
0 & (0.70 & -\ 0.00)^2 + \\
1 & (0.17 & -\ 0.00)^2 + \\
\cdot & & \\
8 & (0.66 & -\ 1.00)^2 + \\
9 & (0.04 & -\ 0.00)^2
\end{array}\right.
$$

Come possiamo notare i valori del vettore y_8 relativi al valore 8 sono tutti 0 ad eccezione della posizione corretta, dove vale 1.

I valori del vettore a_8 invece assumono valori compresi tra 0 ed 1 a seconda dei calcoli che i neuroni dei livelli precedenti hanno compiuto[85]. Applicando la formula presentata sopra, otteniamo l'errore che la rete compie per questo singolo dato di addestramento. Se utilizzassimo solo questo dato per correggere la rete, essa certo migliorerebbe la capacità di riconoscimento, ma solamente per il numero 8. Per evitare ciò, andremo a stimare l'errore medio su tutta la popolazione dei dati del nostro set, per cui la funzione di costo C sarà data da

$$C(w,b) = \frac{1}{n}(|y_1 - a_1| + |y_2 + a_2| + \cdots + |y_n - a_n|)$$

Tale funzione dipende dai pesi w e dai bias b, che nella rete da noi considerata raggiungono il ragguardevole numero di 13002: nel primo strato si contano 784×16 pesi + 16 bias, nel secondo strato 16×16 pesi + 16 bias e nel terzo strato 16×10 pesi + 10 bias.

Quello che la funzione C ci restituisce è un numero che descrive la *performance* della rete nel suo complesso e, se si rende necessaria, una correzione dei valori dei pesi e dei bias: tanto più vicino a zero sarà il valore di C, tanto più la rete è precisa ed affidabile nel compito che sta svolgendo.

Per fare ciò è necessario calcolare la derivata[86] della funzione $C(w,b)$, che in analisi matematica viene chiamata *gradiente di C* e viene indicata come ∇C, le cui componenti sono le

[85] Ricordiamo che questo è possibile proprio grazie all'utilizzo dei neuroni sigmoidali: in presenza dei neuroni con funzione a gradino, l'output avrebbe presentato solamente i valori 0 ed 1.

[86] Solitamente per le funzioni di una variabile $y = f(x)$, la derivata si indica con $y' = \frac{d}{dx}f(x)$ e le regole di calcolo sono quelle che si possono apprendere al liceo. Per funzioni in più variabili $z =$

derivate parziali della funzione di partenza rispetto a ciascuna delle variabili. Per rendere chiaro quale sia il processo che viene attuato, presenteremo di seguito delle nozioni base di analisi e spiegheremo brevemente il significato geometrico del gradiente, restringendo il campo momentaneamente a funzioni di una variabile e poi a due variabili.

Il grafico in figura 3 rappresenta un'ipotetica funzione di costo $C(w)$ dipendente solamente da un singolo parametro w. In analisi matematica trovare il valore che minimizza la funzione

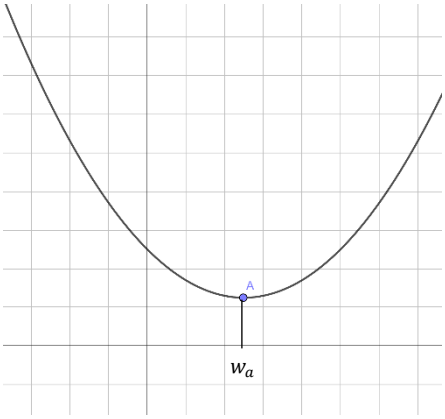


Figura 3

$C(w)$ significa individuare w_a attraverso il calcolo dei punti in cui la derivata prima di tale funzione si annulla. Questo calcolo può essere svolto esplicitamente nei casi delle funzioni più semplici o meglio conosciute, tuttavia non è sempre possibile o agevole farlo, specialmente se si ha a che fare con funzioni molto complesse (nel nostro caso si tratta di una funzione con 13002 parametri).

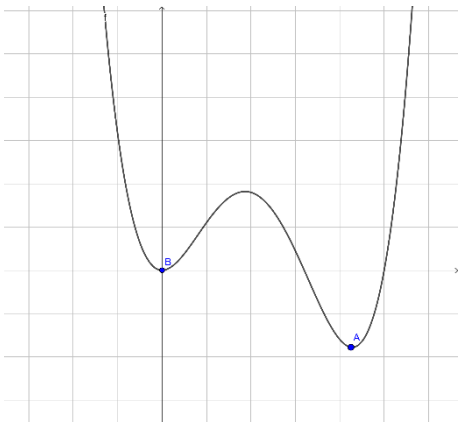


Figura 4

$f(x_1, \dots, x_n)$ la derivata si calcola applicando l'operatore *nabla* ∇ alla funzione *f*, per cui il gradiente della funzione è indicato $z' = \nabla f(x_1, \dots, x_n) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right)$. Per fare un esempio se $f(x) = x^2$, allora $f'(x) = 2x$; se $f(x, y) = \frac{1}{2}x^2 - 5y^3$, allora $\nabla f(x, y) = (x, -15y^2)$.

Un ulteriore problema è dato dal fatto che non tutti i punti di minimo sono equivalenti: anche se si riesce a determinarli, alcuni saranno dei minimi relativi e solo uno sarà il minimo assoluto – se esiste. Come si può vedere in figura 4, il punto *B* è un minimo relativo della funzione in questione, mentre A è un minimo assoluto.

Per riuscire a risolvere il problema ed evitare di calcolare esplicitamente i punti in cui la derivata prima si annulla, è conveniente ragionare sul significato geometrico della derivata prima: essa infatti rappresenta il coefficiente angolare m della generica retta $y = mx + q$, tangente al grafico della funzione in ogni suo punto.

Valutando la funzione in un punto di partenza casuale *x*, si può capire verso che direzione è necessario muoversi per avvicinarsi ad un punto di minimo: se infatti la derivata prima ha un valore negativo, per avvicinarci al punto di minimo dovremmo "spostarci" verso destra; mentre se la derivata assume valore positivo, bisognerà avvicinarsi da sinistra[87].

Figura 5

L'avvicinamento al valore di minimo sarà proporzionale al valore della derivata secondo un parametro η, chiamato tasso di apprendimento, il quale non dovrà essere né troppo alto – pena l'oscillazione attorno al valore di minimo senza mai potervisi avvicinare – né troppo piccolo, altrimenti il processo di avvicinamento richiederebbe troppo tempo e sarebbe troppo dispendioso dal punto di vista computazionale.

Come possiamo vedere nella fig. 4, sebbene il metodo sopra descritto consenta di avvicinarsi al punto di minimo, non è detto che tale punto corrisponda ad un punto di minimo assoluto.

[87] Immagine di 3Blue1Brown

Il procedimento sopra presentato è possibile applicarlo anche per funzioni con più di una variabile, come nell'esempio in fig. 6 del grafico della funzione $f(x,y) = \frac{3}{2}x^2 + \frac{1}{2}y^2$.

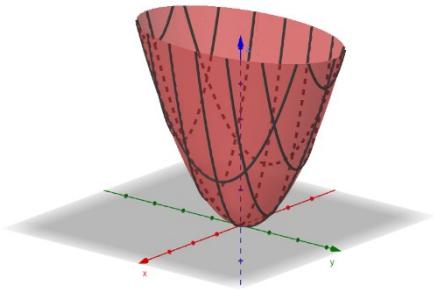


Figura 6

Essendo funzione di più variabili, la derivata prima sarà anch'essa funzione di più variabili, e sarà definita come ∇f[88], chiamata gradiente di *f*.

Il gradiente valutato in un punto della funzione *f* ci restituisce un vettore che indica il verso di massima crescita della funzione in quel punto. Prendendo il vettore opposto a questo saremo in grado di stabilire il verso di decrescita massima della funzione in quel punto.

Per capire meglio attraverso un esempio, consideriamo la cartina di una zona di montagna: sappiamo che per rappresentare il dislivello vengono utilizzate le curve di livello. Tutti i punti aventi la stessa quota sono uniti da una curva chiusa[89]; maggiore è la pendenza del dislivello, più queste curve sono vicine tra di loro.

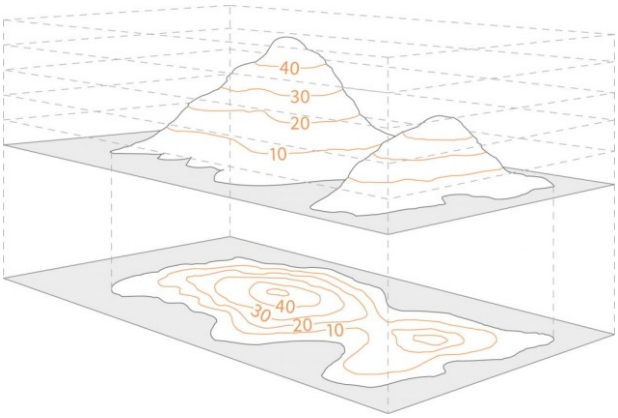


Figura 7

[88] Il simbolo ∇ ha nome *nabla*, in matematica è un operatore differenziale.

[89] Essa è anche chiamata isoipsa, dal greco ἴσος, *isos*= "uguale" e ὕψος, *hypsos*= "altezza". Le curve di livello per quote sotto il livello del mare sono chiamate invece isobate ("stessa profondità").

Guardando la figura 7, se ci trovassimo alla quota 30 e dovessimo passare a quota 40, avremmo a disposizione una molteplicità di tragitti per arrivarci: potremmo prendere la strada più corta e salire in verticale, oppure avvicinarci in modo più graduale e diminuire la pendenza da affrontare.

Poniamoci su un punto qualsiasi appartenente alla curva di quota 30; calcolare il gradiente della curva rispetto a questo punto vuol dire trovare il vettore applicato al punto, ortogonale alla curva stessa, con la punta del vettore orientata secondo la direzione di massima crescita della funzione.

Nella nostra similitudine, con questo metodo l'escursionista prenderà la strada con la maggior pendenza possibile. Il vettore con verso opposto punterà verso la direzione di massima decrescita della funzione.

Le componenti del gradiente sono le derivate parziali della funzione di partenza rispetto alle variabili. Calcoliamo come esempio il gradiente della funzione *f* sopra considerata:

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) = (3x, y)$$

Possiamo già notare dall'espressione su riportata, che l'influenza della derivata rispetto alla *x* è tre volte superiore rispetto a quello della *y* nell'incremento della funzione *f*. Se calcoliamo il valore del gradiente per il punto $(1,1)$, risulta $\nabla f = (3,1)$, che rappresenta il vettore con verso relativo alla massima crescita della funzione; prendendo $-\nabla f = -(3,1) = (-3,-1)$, otteniamo il verso di massima decrescita.

Vediamo ora come influisce il parametro η – il tasso di apprendimento – nel calcolo per avvicinarci al valore del minimo, che in questo caso sappiamo essere il punto $(0,0)$. Consideriamo un valore del tasso pari a *0.5* e andiamo a calcolare $\Delta x = -\eta \cdot \nabla f(1,1)$, ossia lo spostamento Δx che andremo a fare rispetto al punto $(1,1)$ per avvicinarci al punto di minimo[90]. Sostituendo si ottiene

$$\Delta x = -0.5 \cdot (3,1) = -\left(\frac{3}{2}, \frac{1}{2}\right)$$

calcoliamo il nuovo punto $x' = x + \Delta x$, che sarà

$$x' = (1,1) - \left(\frac{3}{2}, \frac{1}{2}\right) = \left(-\frac{1}{2}, \frac{1}{2}\right)$$

[90] Spostandosi dal punto *x* di una piccola quantità Δx, si andrà a variare anche il valore della funzione *f* di una quantità pari a $\Delta f = \nabla f \times \Delta x$, per cui sostituendo a Δx l'espressione $-\eta \times \nabla f$ otteniamo $\Delta f = -\eta \times \nabla f \times \nabla f < 0$, per cui abbiamo la garanzia che con un incremento $\Delta x = -\eta \times \nabla f$ la funzione decrescerà. Per una trattazione più rigorosa si veda B. Codenotti, *La rivoluzione silenziosa*, pp. 143-146.

e valutiamo il gradiente in questo nuovo punto, che risulterà pari a

$$\nabla f\left(-\frac{1}{2},\frac{1}{2}\right)=\left(-\frac{3}{2},\frac{1}{2}\right)$$

Calcolando $-\nabla f=-\left(-\frac{3}{2},\frac{1}{2}\right)=\left(\frac{3}{2},-\frac{1}{2}\right)$ notiamo che rispetto ai valori precedenti ci siamo avvicinati al punto di minimo per quanto riguarda la variabile *y*, ma lo abbiamo sorpassato per quanto riguarda la coordinata *x*; infatti se prima era necessario diminuire il suo valore ora è necessario aumentarlo (il valore di $\frac{\partial f}{\partial x}$ ha cambiato di segno nei due risultati ottenuti).

Questo a causa del valore troppo elevato di η. Infatti in *machine learning* solitamente vengono usati valori molto più piccoli; se proviamo a rifare tutti i calcoli, ma questa volta con $\eta=0{,}1$ otteniamo che $\Delta x=-(0{,}3;0{,}1)$, $x'=(0{,}7;0{,}9)$ e $-\nabla f=-(2{,}1;0{,}9)$, che ci mostra l'avvicinamento al punto di minimo in modo più graduale.

Fondamentalmente si compie lo stesso procedimento anche nel caso della nostra funzione di partenza, $C(w,b)$ con 13002 parametri: per cui è necessario individuare la formula del gradiente della funzione, calcolare il suo valore per ciascuno dei dati di test e poi sommare insieme tutti i valori ottenuti per ottenere il valore medio di $-\nabla C$.

Si applica poi l'algoritmo di retro-propagazione dell'errore per migliorare i valori di pesi e *bias*, e si reitera il procedimento finché la rete non raggiunge dei risultati soddisfacenti. L'applicazione della formula vista prima per avvicinarsi al punto di minimo si potrà riscrivere in modo generale per *w* e *b* nel modo seguente

$$w_k'=w_k-\eta\frac{\partial C}{\partial w_k}\qquad e\qquad b_m'=b_m-\eta\frac{\partial C}{\partial b_m}$$

2.6.3 La questione degli hidden layers

Una volta individuati i giusti parametri per i pesi e i *bias*, e aver fatto esercitare la rete su un numero sufficiente di esempi di numeri scritti a mano, il modello sarà in grado di riconoscere con un ottimo grado di affidabilità anche altri numeri con grafie differenti.

Da un punto di vista puramente prestazionale, ciò che ci interessa è il lavoro degli strati di input e quelli di output, cioè che il *software* sia in grado di acquisire le immagini in entrata e riesca ad individuare il numero in questione. Da questo punto di vista non sarebbe necessario

porsi alcuna altra questione, in quanto il modello funziona e siamo in grado di programmarne degli altri per andare a svolgere delle mansioni differenti.

Invece, uno dei problemi principali dei programmi di intelligenza artificiale risiede in un'opacità di fondo nei processi che regolano tali programmi; per cui, se è lecito aspettarsi che l'utente medio non sappia come funzioni nei particolari l'applicazione che sta usando – ad esempio come funziona *ChatGPT* –, nemmeno gli esperti del settore possono fornire risposte definitive sul "perché" il programma funzioni in un certo modo. Semplicemente, funziona.

Quasi tutti i calcoli che avvengono tra lo strato di input e quello di output è come se fossero racchiusi in una scatola nera di cui però, pur avendo progettato il *software*, non possediamo la chiave. Nell'esempio presentato nei paragrafi precedenti, all'interno di questa scatola sono racchiusi 13002 parametri: un numero molto esiguo rispetto ai 175 miliardi di parametri di GPT 3.5 e agli oltre 100 trilioni di GPT 4.

Sappiamo che avvengono dei calcoli a tutti gli strati della rete, ma non sappiamo dare un'interpretazione globale a tutte queste operazioni né spiegare il perché la rete abbia deciso di aggiornare i suoi parametri in un certo modo. Il processo che governa l'"intelligenza" artificiale è di per sé opaco.

È estremamente difficile, se non impossibile, dire cosa succeda ad ogni livello della rete o come si sta comportando ogni neurone, e a partire da questo spiegare quale sia la sua influenza nell'*output* finale. Il problema è chiaramente intrinseco al tipo di tecnologia: le applicazioni della "vecchia" intelligenza artificiale – dei sistemi esperti – erano programmate manualmente dagli ingegneri e possedevano un linguaggio comprensibile a qualunque buon studente di informatica; non possedevano però la plasticità delle reti neurali.

Se potessimo guardare all'interno del codice di queste reti neurali, non vedremmo altro che lunghe liste di calcoli matematici estremamente complessi, di difficile interpretazione.

Il problema dell'opacità generale delle reti neurali è, oltre che tecnologico, anche etico, politico e sociale. L'attenzione dei governi negli ultimi anni si sta spostando sempre di più verso la regolamentazione dell'IA, per fare in modo di ridurre questa opacità o di far implementare delle tecniche di verifica dei risultati e dei processi. Il principio che si vuole salvaguardare è quello della "spiegabilità" (*explicability*) dei processi.

2.7 Un nuovo tipo di IA: *GenAI*

Uno dei maggiori passi avanti nel campo dell'intelligenza artificiale è stato fatto nel 2017, quando dei ricercatori di Google pubblicarono un loro studio chiamato *Attention is all you need*[91], in cui veniva presentata una nuova ed innovativa architettura di rete, chiamata *Transformer*.

Questa è in grado di compiere in modo più efficiente e rapido le operazioni di *transduction*, cioè di trasformazione dei dati di input in dati di output (pensiamo ad esempio alle applicazioni di traduzione automatica) effettuate precedentemente da altri modelli, come gli RNN (che restano ancora oggi utilizzati).

Lo studio sull'architettura *transformer* si lega a doppio filo con l'avvento e il grande sviluppo dei modelli linguistici di grandi dimensioni – LLM, *Large Language Model* –: modelli addestrati su ingenti quantità di dati, per poi poter essere utilizzati per eseguire attività molto diverse tra loro: comprensione del testo, interazione scritta con utenti, scrittura documenti, generazione di immagini e suoni, ed altro ancora.

Una delle loro principali caratteristiche è l'estrema duttilità in ambiti molto diversi tra di loro. Grazie all'implementazione all'interno degli LLM dei modelli *transformer*, si è arrivati a prodotti estremamente potenti e per certi versi rivoluzionari come *ChatGPT*, che fanno parte di un nuovo sottoinsieme dell'IA, chiamata *GenAI*.

Come suggerisce il nome, l'IA generativa è un tipo di intelligenza artificiale che è in grado di creare dei contenuti originali (come immagini, testi, video, audio) su richiesta dell'utente. Per creare tali contenuti, una parte considerevole del lavoro è svolta dall'analisi che il modello fa dell'enorme quantità di dati che gli vengono messi a disposizione, cercando autonomamente (senza supervisione umana nella prima fase) modelli e relazioni tra i dati.

Tale lavoro, che può durare diverse settimane e impiegare ingenti risorse economiche, è poi coadiuvato da un apprendimento con rinforzo supervisionato dall'uomo, per arginare le possibilità che il modello sbagli o fornisca risposte non adatte – ad esempio output a sfondo razziale o che suggerisca un modo per costruire una bomba al tritolo.

Dato che l'addestramento di questi modelli è un processo lungo e costoso, solamente poche aziende possono permettersi modelli proprietari. Questi vengono poi offerti alle altre aziende per consentire loro di sviluppare le proprie applicazioni; ad esempio ChatGPT, di proprietà di

[91] Cfr. A. Vaswani *et alii*, *Attention is all you need*, 31st. Conference on Neural Information Processing Systems, Long Beach 2017. Reperibile https://doi.org/10.48550/arXiv.1706.03762.

OpenAI, offre differenti funzionalità a seconda del tipo di abbonamento che viene sottoscritto: quello *Enterprise*, rivolto alle aziende, consente l'accesso ai modelli più veloci e potenti rispetto a quelli accessibili tramite abbonamento gratuito per tutti gli utenti.

La presenza sul mercato di pochi modelli addestrati, per la maggior parte posseduti da aziende private, rappresenta una fonte di squilibrio dal punto di vista economico ed anche politico: nel Ventunesimo secolo – era dei *Big Data* –, chi controlla i dati ed è capace di estrapolarne informazioni preziose possiede un potere non indifferente anche verso gli Stati.

Per questo è lecito aspettarsi nei prossimi anni sviluppi di modelli proprietari anche da parte di tutte le principali potenze mondiali, come ad esempio si sta facendo in Italia con il progetto LLM, denominato "modello Italia", che sfrutterà la potenza del supercomputer Leonardo[92].

2.7.1 Cos'è e come funziona ChatGPT

Il nome esteso del programma è *Chatbot Generative Pretrained Transformer*: la prima parola – *chatbot* – ci dice che si tratta di un programma informatico in grado di simulare una conversazione umana, sia tramite testi scritti che tramite interazioni vocali con l'utente.

Ne esistono principalmente di due tipi: quelli definiti "dichiarativi", i più semplici, che forniscono il più delle volte risposte automatizzate alle richieste degli utenti; e quelli "predittivi", come quello in questione, che fanno largo uso di algoritmi di IA e possono quindi comprendere meglio le domande e articolare risposte più complesse.

La seconda parte, *Generative Pretrained Transformer*, ci presenta le caratteristiche principali che rendono questo chatbot così potente e versatile, ossia un modello linguistico di grandi dimensioni (*LLM*), basato su un'architettura di apprendimento profondo di tipo *transformer*.

Generative sta ad indicare la sua appartenenza al sottoinsieme della *GenAI*, e quindi la sua capacità di creare contenuti originali a partire dalle strutture e dai pattern precedentemente imparati.

Pretrained si riferisce al processo di apprendimento utilizzato per addestrare il modello, che in un primo momento è di tipo non-supervisionato su dati non etichettati: questo significa che il programma parte da una sorta di "tabula rasa", con i vari parametri inizializzati a valori casuali e, analizzando milioni e milioni di dati – libri, articoli, siti web –, è in grado di estrapolarne le

[92] Cfr. Modello Italia, la sfida italiana a ChatGPT | Wired Italia, da notare che questo modello LLM sarà addestrato prevalentemente con dati in lingua italiana, differentemente da quanto succede con tutti gli altri modelli, che si basano principalmente su dati in lingua anglosassone.

relazioni ed i pattern, per poi utilizzarli successivamente. In un secondo momento si utilizzano anche tecniche di apprendimento semi-supervisionato con intervento dell'uomo per migliorare ed affinare le capacità della macchina.

È conveniente fermarsi un momento sul processo di apprendimento tramite reti neurali, per apprezzarne la potenza e le capacità, ma anche per mettere in risalto la differente modalità con cui i modelli artificiali e l'uomo apprendono.

Prendiamo per esempio la comprensione della grammatica: mentre l'uomo la apprende studiandone le regole, per poi applicarla in modo consapevole – riconoscendo cioè le strutture base come nome, verbo, aggettivo, ecc.. e combinandole insieme –, per l'IA l'apprendimento della grammatica è frutto dell'analisi di milioni di frasi e strutture linguistiche in contesti differenti, per poi estrarne le regole grammaticali, la sintassi e le convenzioni di scrittura. Il processo che porta alla conoscenza grammaticale è per un certo senso inverso a quello usato dall'uomo.

L'IA non ha consapevolezza dell'uso di queste regole, non possiede una grammatica nel senso umano, ma sa solamente che "statisticamente" le parole "Giuseppe, Lucia" – più tutte quelle che ha trovato nei dati di addestramento –, non vanno precedute dalle parole "il, lo, la", ecc.., che noi chiamiamo rispettivamente nomi propri ed articoli. Questo carattere stocastico è intrinseco alla natura delle reti neurali, che sono in grado di eseguire molti compiti svolti dall'uomo, ma in un modo sostanzialmente diverso.

Relativamente all'esempio appena affrontato sarebbe interessante provare ad addestrare un modello con una grande quantità di dati, in cui però non fosse presente nemmeno una fonte esplicita relativa alle regole grammaticali.

Con buona probabilità al termine dell'addestramento il modello sarebbe in grado di comunicare egualmente in modo fluido e comprensibile, ma sarebbe incapace di rispondere ad una domanda come "quali sono le regole grammaticali dell'italiano?".

Il modo in cui il modello acquisisce ed elabora le informazioni è individuato dalla parola *Transformer*, che come abbiamo visto prima indica la nuova architettura introdotta per la prima volta dai ricercatori di Google nel 2017.

I modelli di linguaggio precedenti, basati prevalentemente su reti neurali ricorrenti – *RNN* –, processavano gli input in maniera sequenziale, aggiornando il loro stato a mano a mano che passavano da una parola all'altra.

Nella frase “Stasera si vede la costellazione del cane”, il modello *RNN* partiva dalla parola “Stasera” e proseguiva in avanti processando una parola alla volta, fino ad arrivare a “cane”. Le relazioni tra le parole sono mediate dagli altri termini e non sono presenti relazioni “dirette”.

Se scrivessimo “Stasera, mentre portavo a passeggio il cane, ho visto la costellazione del cane”, con il metodo sequenziale il modello potrebbe avere qualche problema a comprendere le relazioni tra le varie parole.

Il metodo sequenziale porta al problema di non riuscire a catturare in modo efficiente le dipendenze semantiche anche su sequenze di frasi non particolarmente lunghe, inficiando così sia la comprensione complessiva da parte del modello che la qualità delle risposte, limitate e a volte ripetitive.

Il nuovo modello invece utilizza il meccanismo dell’attenzione per comprendere le relazioni semantiche tra le parole in maniera molto più efficiente e potente.

Come primo passaggio, il testo inserito in input dall’utente viene suddiviso in elementi più piccoli – che possono essere parti di parole, parole intere, segni di punteggiatura, ecc… -, e rappresentati in linguaggio matematico tramite dei vettori, su cui il modello può agire attraverso calcoli (ricordiamo che la retro-propagazione dell’errore, che sta alla base dell’apprendimento delle macchine, è possibile solo su dati di tipo numerico).

Se prendiamo ad esempio la frase “la costellazione del cane è molto luminosa”, essa verrà divisa secondo le singole parole (per semplificare), e ogni parola verrà trasformata in un vettore *n*-dimensionale (dove *n* è legato al numero di parametri del modello; se ad esempio un modello ha 30 parametri, allora il vettore avrà dimensione $n \leq 30$), che in qualche modo rappresenterà il significato di quella parola.

A questo livello i singoli vettori non incorporano ancora il senso che la parola di partenza ha nella frase: il modello non sa ancora se “cane” indica l’animale o la costellazione; in questa fase dell’analisi, il modello potrebbe avergli dato un significato sbagliato.

Questo vettore di partenza contiene un primo “posizionamento semantico” della parola, ottenuto grazie ai dati di addestramento presenti: il modello confronta le migliaia di esempi presenti in memoria in cui compare e non compare la parola “cane”, e crea un vettore in cui associa a questi esempi un valore compreso tra [0;1], a seconda di quanto questi esempi siano compatibili con la parola di partenza.

Ad esempio, per “cane” la parola “corre” avrà dei valori più alti, poniamo 0.6, rispetto a “vola”, con un valore di 0,05. Contemporaneamente anche la parola “brilla” potrà avere un valore alto.

Questo processo viene chiamato *word embedding* e, sebbene non sia possibile definire che cosa rappresenti ciascuna singola variabile nel vettore, è lecito aspettarsi che a parole simili corrispondano vettori simili: ad esempio i termini "cane" e "gatto" hanno molte caratteristiche in comune (quadrupede, mammifero, ecc..) e quindi ci aspettiamo che i vettori a loro associati abbiano valori simili tra di loro.

La vera novità del *Transformer* si mostra nel passaggio successivo, grazie all'applicazione dell'algoritmo dell'auto-attenzione – *self-attention* –, per cui i singoli vettori vengono fatti "comunicare"[93] con tutti gli altri vettori della sequenza di input (sia questa una frase, un paragrafo o un articolo intero), e i valori di ciascun vettore vengono aggiornati con le informazioni ottenute dall'interazione con tutti gli altri vettori.

Questa è la differenza principale rispetto all'algoritmo sequenziale utilizzato dalle *RNN*: ora il significato di una parola è ricavato tramite una relazione con tutte le altre parole presenti nel testo, indipendentemente dalla loro posizione, e non solamente rispetto alla parola precedente.

In questo modo il modello è in grado di acquisire il contesto e di individuare quale parte del testo sia più determinante per definire ciascun termine. Questo processo viene effettuato in parallelo molte volte, e i vettori passano attraverso vari strati di *self-attention* alternati a reti neurali *feed-forward*.

Al termine di tutti questi passaggi, si ottiene un vettore che contiene tutte le informazioni delle parole precedenti, che il modello può utilizzare per predire la parola successiva più adatta al contesto. Il modo in cui la parola successiva viene decisa è di tipo statistico: infatti il *software* ottiene come ultimo passaggio una lista delle parole più probabili per continuare la frase; tra tutte queste, sceglie quella avente probabilità maggiore[94].

---

[93] L'estrazione dei dati di contesto e l'aggiornamento dei valori dei vettori – la "comunicazione" –, si ottiene tramite reiterate somme e moltiplicazioni tra vettori e matrici, ciascuna delle quali contiene dei parametri in grado di estrarre i valori di contesto importanti per quella parola in quella specifica posizione all'interno della frase. Ancora una volta è impossibile dire che cosa faccia di preciso anche solo una di queste matrici, perché affinate in autonomia dal modello durante la fase di addestramento.

[94] Il modo in cui il modello individua le parole è uno dei parametri che gli sviluppatori possono scegliere: esso viene chiamato "temperatura" e varia tra [0;1]. Un alto valore di questo parametro indica che il modello sceglierà le parole con meno probabilità e quindi più inconsuete, mentre con un valore basso sceglierà le parole più probabili e comuni. Con un valore alto si rischia di ricevere delle risposte poco coerenti, mentre con un valore basso si rischia di avere delle risposte poco articolate, e ripetitive nei termini. Solitamente viene usato un valore intermedio.

Inoltre, invece di predire solamente una parola alla volta in modo isolato, il *Transformer* è in grado anche di produrre una lista delle frasi più probabili che seguiranno[95] e scegliere la parola successiva anche in relazione a tale lista, per fornire un output migliore e più coerente.

Per fare un esempio, supponiamo di chiedere a ChatGPT di concludere la frase: "Stasera la costellazione del cane è …"; il modello procederà così:

1. Dividerà la frase di richiesta in *token* a cui assegnerà un vettore con dei valori di partenza.
2. Procederà ad applicare i meccanismi di *self-attention* e gli strati di reti neurali per fare in modo di aggiornare i valori dei vettori e individuare il contesto.
3. Dopo ripetuti cicli dei meccanismi sopra indicati, verrà calcolato un vettore finale che conterrà tutte le informazioni di contesto dei precedenti.
4. A tale vettore il modello assocerà una rosa di termini con varie probabilità: "luminosa", "oscurata", "particolarmente [bella]", e procederà a scegliere quella maggiormente probabile ed adatta al contesto.
5. Nel caso di una riposta più articolata, il modello ripartirà nuovamente dal punto (1) e costruirà la risposta una parola alla volta.

Ogni volta che chiediamo qualcosa a *ChatGPT* o ad un applicativo simile, il primo passaggio che il *software* fa è quello di analizzare la nostra frase in input secondo questo metodo, per poi fornirci la risposta.

Visto il funzionamento di questi applicativi, possiamo individuare alcuni spunti di riflessione interessanti.

In primo luogo, cade parzialmente l'obiezione sollevata da Searle riguardante l'abilità solamente sintattica delle applicazioni di IA: per offrire risposte coerenti con il contesto, i modelli devono essere in grado di comprendere anche semanticamente ciò che viene loro comunicato, ed è quello che in un certo modo sono in grado di fare, seppur in maniera differente rispetto ad un essere umano. La rappresentazione semantica delle parole per un *software* di IA è realizzata nello spazio n-dimensionale in cui i vettori – ossia la traduzione in linguaggio matematico delle parole (o *token*) – si trovano[96].

[95] Questo metodo viene chiamato *beam search*.

[96] Per una spiegazione più esaustiva si può visitare 3Blue1Brown - But what is a GPT? Visual intro to Transformers | Deep learning, chapter 5

Possiamo immaginare la rappresentazione semantica delle parole nel modo mostrato in fig. 7, per cui supponiamo di avere i vettori che individuano la posizione delle parole “mamma” e “papà” nello spazio n-dimensionale (in questo caso di dimensione 3, ma è necessario immaginarlo nell’ordine dei milioni, se non miliardi, di dimensioni), e di avere altri due vettori “nonno” e “nonna”.

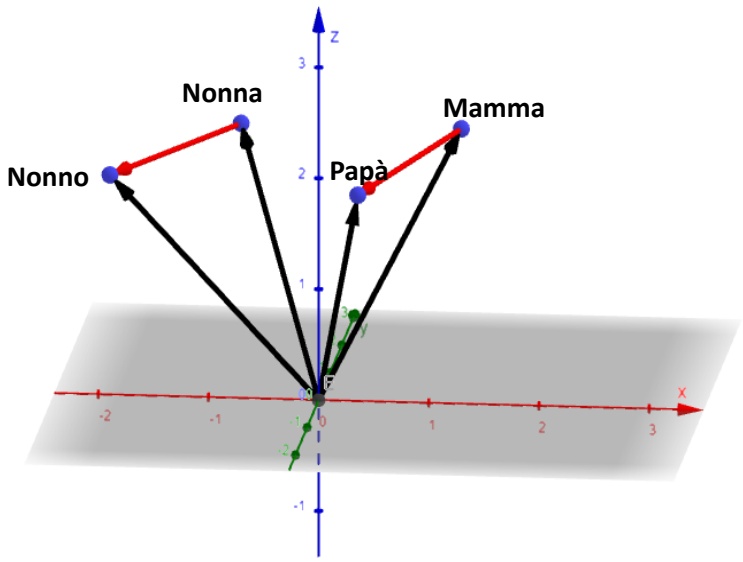


Figura 7

Se prendessimo la differenza tra i vettori “papà meno mamma”, noteremmo che tale valore sarebbe molto simile al valore del vettore della differenza “nonno meno nonna”. Tale valore sarebbe anche simile alla differenza tra vettori “zio meno zia” e così via. Il modello ha cioè associato ad una direzione dello spazio il concetto di “genere” e ha orientato di conseguenza i vettori tra loro simili.

Tutti i *token* che il modello ha analizzato nel momento dell’addestramento trovano posto all’interno di questo campo vettoriale e, se l’algoritmo ha funzionato a dovere, saranno divisi tra loro sia dal punto di vista sintattico – articoli, aggettivi, nomi, avverbi, ecc.. – che dal punto di vista semantico. Non sarebbe possibile ricevere delle risposte coerenti se il modello non avesse un minimo di competenza semantica.

Rimane però ancora valida la seconda parte dell’obiezione di Searle; posto che questi modelli mostrino una competenza anche di tipo semantico, è possibile affermare che “sappiano”, cioè comprendano, ciò che scrivono?

La risposta a questa domanda resta negativa: come abbiamo visto la rappresentazione semantico-sintattica di questi sistemi è di tipo geometrico-matematico ed il principio di funzionamento si basa su complesse equazioni matematiche e statistiche.

Questi modelli, come abbiamo visto nei paragrafi precedenti, restano dei *software* molto raffinati, in grado di apprendere e rielaborare dati per fornire oggetti (testi, immagini, video, audio, ecc…) simili a quelli che potrebbero essere prodotti da un essere umano.

Si potrebbe affermare che questi modelli possiedano e non possiedano una competenza semantica: la possiedono nel modo che è stato sopra spiegato, non la possiedono considerando che tutte le parole assimilate dal software sono sostanzialmente "vuote", non presentano cioè alcun carattere intenzionale.

Il linguaggio è intrecciato a doppio senso con il mondo e con il rapporto vivo e concreto che l'uomo ha con quest'ultimo: se scindessimo questo intreccio, parafrasando Kant, potremmo dire che "il linguaggio senza mondo è vuoto".

Pensiamo ad esempio a *Genesi 2:20-21*, in cui Dio, dopo aver formato dalla terra tutti gli animali e gli uccelli, li porta all'uomo per «vedere come li avrebbe chiamati, e perché ogni essere vivente portasse il nome che l'uomo gli avrebbe dato». Questa capacità linguistica e concettuale è una delle caratteristiche proprie dell'uomo: egli osserva, astrae, nomina e precisa la molteplicità di cui fa parte, che non si riduce solamente alle piante ed agli animali, ma si apre a tutto ciò che ricade all'interno dell'orizzonte trascendentale su cui l'uomo è aperto.

È questo uno dei caratteri che manca completamente ai modelli attuali di IA: il linguaggio di cui fanno uso, non rimanda a null'altro; è, appunto, un linguaggio vuoto, che non tende a nulla. Si potrebbe obiettare che, se si chiedesse a *ChatGPT* di disegnare "una mela su un albero in cima ad una collina", esso effettivamente produrrebbe qualcosa di simile a quanto richiesto.

Tuttavia, tale relazione tra parole ed oggetti rappresentati sarebbe solamente contingente e non essenziale: è messa in atto unicamente nel momento della produzione del contenuto. Nella richiesta di prima, l'immagine prodotta potrebbe presentare la mela sulla cima all'albero, anziché tra la chioma come ci aspetterebbe. Questo perché il modello, pur avendo "semantizzato" entrambi i significati di albero e mela, non è in grado di porli in una relazione reale tra di loro. Si dovrà prevedere un addestramento che faccia apprendere al modello come queste parole stiano in relazione nel mondo, ad esempio attraverso il riconoscimento di migliaia di immagini di mele e alberi.

Vi è poi un'ultima questione da affrontare, che risiede nel significato del termine *generative*: ossia se questi programmi possano essere espressione di una qualche creatività, e se ciò che creano possa essere considerato opera originale.

Pensiamo ai modelli usati per produrre immagini comparabili con fotografie reali eseguite da professionisti o all'abilità di scrivere testi in grado di eludere i più blasonati programmi antiplagio.

In primo luogo, anche senza coinvolgere nella discussione concetti come quello di "coscienza", quando si attribuisce una capacità creativa a questi *software* ci si dimentica che essi sono solamente degli strumenti che elaborano input provenienti dall'esterno: nel caso di una creazione artistica, sarà l'essere umano a chiedere di rappresentare qualcosa alla macchina secondo delle indicazioni; cosa analoga nel caso della produzione di un testo. Che il risultato finale sia considerabile una forma di arte è possibile, ma non è possibile affermare che il processo creativo sia proprio della macchina: esso infatti è proprio solamente dell'essere umano.

Inoltre, il *software* ha accesso solamente ad un vocabolario di concetti e parole che è limitato, e che è incrementabile solamente fornendo altre informazioni dall'esterno, ma non tramite un "principio" interno. E non ha la capacità di espanderlo in autonomia, perché gli manca anzitutto il rapporto con il mondo. Ma anche per un'altra ragione più radicale: per costituzione, l'orizzonte dei *software* è di tipo algoritmico, per cui essi non hanno la possibilità di andare oltre a ciò che è contenuto nel codice del programma. Attraverso l'analisi di milioni di documenti, i modelli costruiscono il proprio vocabolario individuando i *pattern* e le relazioni presenti tra le singole parole; una volta che questo processo è terminato, in assenza di ulteriori input non è possibile alcuna risemantizzazione delle parole. Il modello ha cioè minimizzato la funzione di costo complessiva di tutte le parole date e si trova in un punto di minimo, in cui il processo di addestramento si ferma.

L'invenzione di nuove parole o la risemantizzazione delle parole già date, in una maniera che potremmo definire "poetica", non è possibile all'interno di un orizzonte di tipo algoritmico. Infatti, il *software* non fa altro che andare a trovare i *pattern* nelle parole che gli vengono date dall'esterno. Trovate queste relazioni, in assenza di altri *input* esso non è capace di modificare ulteriormente i significati delle parole.

L'essere umano, invece, grazie all'apertura trascendentale del suo orizzonte, è in grado non solo di lavorare sull'attuale, cioè sul reale già dato (i significati attuali), ma anche di accedere all'ambito del virtuale. Accedendo a questa dimensione, l'uomo è in grado di compiere l'azione

“poetica”, cioè di portare alla luce significati latenti in parole già date o di inventare nuove parole per descrivere ciò che già è presente.

I *software* sono in grado cioè di lavorare unicamente sull’attuale, cioè sul reale già dato che hanno a disposizione, mentre la capacità dell’uomo è di riuscire ad accedere al reale non dato – il virtuale – a partire dall’attuale.

Potremmo affermare che le opere create dai modelli presentano tratti di originalità – cioè non sono un mero copia-incolla da altri lavori –, ma non per questo possiamo affermare che l’IA esprima una qualche forma di creatività, che come tale è esclusiva dell’essere umano.

3 Gli agenti razionali

Come abbiamo visto nel paragrafo 1.2, secondo quello che viene chiamato “modello standard”, l’approccio più utilizzato nel campo di studi inerente all’IA è quello che si occupa di sviluppare e costruire agenti razionali artificiali.

L’intelligenza in questo caso è “espletata” dall’agente – sia esso virtuale o fisico – attraverso l’esecuzione della migliore azione possibile negli ambienti in cui si trova ad operare; un agente con queste caratteristiche viene anche chiamato agente razionale. Ma dovrà essere chiarito che cosa intendiamo con “agente”, “agente razionale”, “ambiente” e soprattutto con “migliore azione possibile”.

In questo capitolo ci proponiamo di presentare delle possibili definizioni dei termini sopra riportati, mostrando come in alcuni casi tali definizioni dipendano l’una dalle altre. Per fare ciò utilizzeremo ancora una volta come riferimento il libro di Russell e Norvig, *Intelligenza artificiale, un approccio moderno*[97].

3.1 Cosa si intende per “agente”

In modo abbastanza ampio, è possibile definire “agente” qualunque entità – sia essa fisica o virtuale –, che sia in grado di percepire l’ambiente circostante mediante dei sensori e che possa agire su di esso mediante degli attuatori.

Nel caso dell’uomo, la raccolta di informazioni dall’ambiente esterno avviene tramite gli organi di senso, mentre l’interazione avviene tramite la movimentazione fisica delle nostre

[97] Cfr. S. Russell, P. Norvig, F. Amigoni (a cura di), *Intelligenza artificiale, un approccio moderno*, Vol.1, Pearson Italia, Milano 2024, cap. 2.

membra. Per una macchina, la raccolta delle informazioni avviene tramite i sensori che ha a disposizione mediante pacchetti di dati e l'interazione potrà avvenire sia in modo "fisico", cioè con attuatori elettro-meccanici, sia in modo "virtuale", con l'emissione di pacchetti di dati.

Ad esempio, un robot utilizzato nelle saldature delle automobili interagirà in modo "fisico", andando a compiere il lavoro per cui è programmato. Un bot virtuale impiegato per inviare messaggi promozionali a una lista di contatti email agirà invece in modo "virtuale".

Nel caso della macchina, le informazioni arriveranno a seconda dei sensori di cui è dotata e agirà sfruttando i motori che ha a disposizione: ad esempio, una telecamera di sicurezza sfrutterà il suo sensore ottico per individuare e seguire gli spostamenti delle persone nel suo campo visivo. L'insieme di tutte le percezioni che un agente riceve durante la sua vita operativa è chiamata "sequenza percettiva".

Notiamo che non è possibile definire un ambiente "universale", cioè che valga indipendentemente per ogni tipo di agente, ma questo è determinato sulla base di ciò che esso è in grado di percepire e su cosa può agire.

L'ambiente dell'esempio relativo alla telecamera di sicurezza è solamente quello "visivo", proveniente dal sensore fotografico. Per un robot più complesso in grado di muoversi all'interno di un magazzino, l'ambiente oltre che "visivo" sarà determinato anche dalle altre dimensioni sensibili a cui può avere accesso con i suoi sensori.

3.2 Il concetto di razionalità

Attraverso le informazioni che riceve dall'esterno, lo scopo di un qualsiasi agente è quello di interagire con l'ambiente per raggiungere il suo obiettivo. Un agente viene detto "razionale" se è in grado di compiere la migliore azione possibile con le informazioni che ha a disposizione in quel momento.

La questione cruciale risiede nel riuscire a definire una certa azione come la "migliore possibile": si potrebbe infatti usare come parametro l'efficienza complessiva del sistema oppure, quella che porta a minimizzare l'utilizzo dell'energia, e così via.

Per confrontare tra loro le azioni possibili e valutare quale tra queste sia quella più adatta in un certo ambiente, vengono considerate esclusivamente le conseguenze del comportamento dell'agente. L'approccio utilizzato è cioè quello consequenzialista: nella sequenza di azioni possibili viene individuata quella desiderabile tramite una misura di prestazione o di utilità

rispetto alle conseguenze attese. Tale misura viene decisa a monte dai programmatori, in quanto le macchine non possiedono alcuna preferenza o desiderio personale che possano indirizzarle verso un'azione anziché un'altra.

Come per l'osservazione precedente sull'ambiente, anche per la scelta della migliore azione le misure di prestazione sono tutte diverse tra loro, poiché i fini operativi sono diversi a seconda del compito degli agenti. Ad esempio, per un robot lavapavimenti conterà la qualità della pulizia in rapporto con la velocità media nel calcolo dei possibili percorsi da intraprendere, mentre un *software* per la guida autonoma avrà dei parametri completamente diversi da valutare per scegliere la strada più veloce che colleghi un punto A al punto B.

Facciamo notare che in questa prospettiva, propria del consequenzialismo, per poter valutare i diversi tipi di conseguenze che conseguono alle diverse azioni è necessario associare ad ogni conseguenza, probabile o meno probabile, un valore numerico. Questi valori verranno poi utilizzati da un algoritmo interno all'agente, per calcolare quale azione è da preferire alle altre: essa sarà quella che avrà il valore totale più alto della somma dei singoli valori delle conseguenze (questo aspetto verrà trattato nel dettaglio al par. 3.4). In proposito, i punti su cui ci si può soffermare a riflettere sono molteplici; ne proponiamo di seguito alcuni:

(1) L'approccio consequenzialista non considera il valore in sé delle singole conseguenze, guarda piuttosto il risultato finale dato dalla loro somma: su un piano etico significa che anche delle conseguenze negative, se considerate assieme ad altre positive, possono portare ad un risultato complessivo positivo e quindi risultare desiderabili all'interno di un processo decisionale. Un esempio classico si rifà alle scelte in ambito militare: in un bombardamento mirato contro un obiettivo, quante persone innocenti possono essere messe a rischio. Un approccio di questo tipo non solleva interrogativi morali per nel caso del robot lavapavimenti, il quale non può portare a conseguenze gravi tramite le sue azioni, ma coinvolge le applicazioni più avanzate e socialmente impattanti: pensiamo ad esempio alle auto a guida assistita, ai droni militari autonomi o alle applicazioni di IA in ambito medico.

(2) Da quanto visto sopra, per queste applicazioni che operano in ambienti critici è necessario assegnare un valore numerico alle varie conseguenze per poi poterle valutare assieme a tutte le altre. Ad esempio, chi e in che modo può valutare quanto valga una vita rispetto ad un'altra. Al riguardo, un esempio ormai diventato classico è quello del *trolley problem*[98], in cui si ha un treno

[98] Fu introdotto la prima volta da Philippa Foot nel 1967 nel suo articolo *The problem of abortion and the doctrine of double effect* pubblicato in "Oxford Review", vol. 5.

fuori controllo che si approssima ad una biforcazione delle rotaie. Da una parte delle rotaie c'è una sola persona, dall'altra ce ne sono cinque. La scelta è tra salvare una persona e lasciarne morire cinque, oppure tirare una leva e salvare queste ultime, sacrificando però la persona singola. Se facessimo associare alla macchina un valore molto grande per la singola vita umana– per cui il valore qualitativo della persona viene tradotto in valore quantitativo – allora la scelta sarebbe facile, poco più che un conto: 1 minore di 5. La macchina quindi azionerebbe la leva e salverebbe le cinque persone. Tuttavia, il problema non è di così facile risoluzione se a scegliere è un essere umano, che non valuta (solo) secondo criteri quantitativi[99]. Queste questioni etiche sono di interesse centrale per l'ambito della *machine ethics*, in cui si studia sia l'impatto etico dell'IA attraverso i *software* ed i robot, sia come implementare delle teorie etiche efficaci all'interno di queste applicazioni.

Riprendendo quanto detto prima di questo inciso, la razionalità di uno specifico agente dipende da diversi fattori correlati fra loro, ma possiamo individuarne quattro principali, che definiscono cos'è razionale in un dato momento:

1. La misura di *performance* che definisce il criterio di successo
2. La conoscenza pregressa dell'ambiente da parte dell'agente
3. Le azioni che l'agente può effettuare
4. La sequenza percettiva dell'agente fino a quel momento

Consideriamo come esempio la macchina costruita da Shannon per risolvere i labirinti, sappiamo che, quanto al punto (2), per l'agente non vi è alcuna conoscenza a priori dell'ambiente in cui si troverà. Il sensore infatti deve testare ogni parete dei vari quadrati a disposizione per capire che direzione prendere. Dopo aver risolto per la prima volta il labirinto, il sensore acquisisce una conoscenza dell'ambiente ed è in grado di muoversi molto più velocemente.

Per il punto (3), le azioni a disposizione del sensore sono essenzialmente due: la rotazione in senso antiorario e l'avanzamento. Circa il punto (4), il sensore percepisce solamente dove si trova in quel preciso momento, ma non ha memoria delle percezioni precedenti.

Il fattore (1), invece, per l'esperimento originale non è determinato: il criterio di successo è quello di trovare l'obiettivo, ma non è presente una misura delle *performance*. Questa potrebbe

[99] L'articolo di Philippa Foot ebbe grande diffusione, anche fuori dall'ambito strettamente bioetico, lasciando spazio a molte altre versioni del problema originale. A tal proposito si veda E. David, *Would you kill the fat man?*, Princeton University Press, Oxford 2014.

essere implementata misurando ad esempio il numero di mosse occorrenti per arrivare all'obiettivo o il tempo impiegato allo scopo.

La definizione di un'agente razionale basata sui quattro punti sopra presentati si può articolare come segue[100]: per ogni possibile sequenza percettiva, un agente razionale dovrebbe scegliere l'azione (punto 3) che ci si aspetta massimizzi la misura di *performance* (punto 1), considerate le conoscenze provenienti dalla sequenza percettiva e dalle altre conoscenze che l'agente possiede (punto 4 e punto 2).

La sequenza 1-4 che abbiamo visto va anche sotto il nome di PEAS – *Performance, Environment, Actuators, Sensors* –; nel momento in cui si va a progettare un agente razionale è essenziale tenere a mente questi quattro punti chiave, poiché dalla loro corretta specificazione dipende la buona realizzazione finale dell'agente.

3.3 I "programmi agente"

Possiamo considerare gli "agenti" come composti da due parti distinte: una è l'architettura, cioè il modulo fisico che si occuperà di ospitare il *software*, l'altra è il programma agente.

I programmi agente sono il cuore di ogni agente razionale, contengono cioè tutte le informazioni che consentono all'agente di interagire con l'ambiente a partire dalle informazioni che esso ha a disposizione. Esistono principalmente quattro tipi di programmi agente che incarnano i principi che sottostanno a tutti gli altri sistemi intelligenti[101]:

- Agenti reattivi semplici
- Agenti reattivi basati su modello
- Agenti basati su obiettivi
- Agenti capaci di apprendere

Nel primo gruppo ricadono gli agenti più semplici, che considerano unicamente la percezione corrente e ignorano la sequenza di tutte le azioni precedenti. Ad esempio, un robot aspirapolvere agisce determinando se la posizione dove si trova presenta dello sporco, ma questa decisione non è influenzata dall'azione compiuta qualche metro prima. La più semplice modellizzazione di questo tipo di condizioni di azione è individuata dalla formula *if-then*, chiamata anche formula di condizione-azione: al presentarsi di una condizione puntuale (*if*), il programma risponde

[100] Cfr. S. Russell, P. Norvig, F. Amigoni (a cura di), *Intelligenza artificiale, un approccio moderno*, p. 43.
[101] A tal proposito si veda *Intelligenza artificiale, un approccio moderno*, pp. 52-60.

attraverso un'azione ben precisa (*then*), oppure scorrendo tra una serie di azioni alternative (*else*).

Il non tener conto della sequenza percettiva precedente può essere una condizione accettabile per applicazioni semplici, ma risulta insoddisfacente per le applicazioni più avanzate. Pensiamo ad esempio al *software* di un'auto a guida autonoma: per determinare se la macchina di fronte a sé sta frenando accendendo le luci di stop, è necessario che esso possa confrontare tra di loro il fotogramma precedente alla frenata e il fotogramma successivo, relativo alle luci dei freni. Se non fosse dotato di questa memoria dei fotogrammi, probabilmente non frenerebbe mai.

Agenti di questo tipo, dotati cioè di una memoria della sequenza percettiva precedente alla percezione corrente, fanno parte del secondo gruppo. Essi devono possedere una sorta di "stato interno", che si ottiene con l'implementazione di due tipi di modelli: il modello di transizione ed il modello sensoriale.

Il primo si occupa di fornire al *software* sia le informazioni circa gli effetti delle azioni dell'agente sull'ambiente che la modalità di evoluzione del mondo indipendentemente dell'agente. Come esempio del primo tipo di informazioni, agendo sul volante verso destra, anche l'automobile girerà a destra proporzionalmente alla rotazione del volante (effetto dell'azione); come esempio del secondo tipo di informazioni, si pensi a quella per cui in presenza di pioggia il manto stradale risulta più sdrucciolevole (evoluzione del mondo).

Il modello sensoriale si occupa invece di tradurre i dati provenienti dall'ambiente in informazioni per l'agente: l'accensione dei fari rossi di una macchina indica una frenata; l'abbassamento dell'intensità luminosa indica il sopraggiungere della sera. Un agente che utilizzi congiuntamente questi due modelli prenderà il nome di "agente basato su modello". È chiaro come questo tipo di agente sia molto più adatto, rispetto al precedente, ad applicazioni in ambienti complessi.

La successiva implementazione che può essere integrata in un agente è la capacità di orientare le sue azioni rispetto ad un obiettivo (*goal*) deciso dai programmatori. Dopo che un programma per un taxi a guida autonoma (dotato di entrambi i modelli visti sopra) abbia ricevuto tutte le istruzioni su come agire sui comandi della macchina e su come interpretare le informazioni provenienti dall'esterno, sarà necessario fornirgli un obiettivo da raggiungere, ad esempio un punto B come destinazione del tragitto.

Fornito tale obiettivo, il programma si occuperà di scegliere le azioni più appropriate per conseguirlo (per quanto riguarda l'esempio dell'auto, essa sceglierà il tragitto più adatto). Sono presenti nell'ambito dell'IA delle branche di studio specifiche, relative ai metodi di ricerca e pianificazione volti a individuare quali siano le migliori azioni da compiere: infatti, non è sempre semplice per un agente determinare quale azione sia da preferire rispetto ad un'altra.

Una delle differenze più importanti di questo tipo di agenti rispetto a quelli reattivi, è la rappresentazione esplicita della conoscenza che guida le loro azioni: nell'esempio della macchina che si ferma, un agente reattivo agisce sui freni, perché in modo implicito nel suo programma è scritto di agire in questo modo. Invece, un agente basato su obiettivi, frena perché quella è l'unica azione che gli consente di soddisfare un certo obiettivo, ad esempio di non fare incidenti. Così modificando l'obiettivo, si modifica anche il comportamento dell'agente; mentre ciò non è possibile per l'agente reattivo, per cui sarebbe necessario riscrivere parte del *software*.

Talvolta non è sufficiente la specificazione dell'obiettivo per consentire all'agente di compiere al meglio ciò per cui è progettato: infatti, per arrivare da un punto A ad un punto B, un taxi a guida autonoma può scegliere una pluralità di percorsi e quindi di azioni da compiere.

Potrebbe decidere per un percorso più veloce, ma che comporta con maggiore dispendio di carburante, oppure optare per un percorso urbano con velocità più basse. Per calcolare in generale quale sequenza di azioni sia preferibile rispetto alle altre, si deve integrare all'interno del programma una "funzione di utilità", che non è altro che una "internalizzazione" della misura di prestazione che abbiamo visto all'inizio del capitolo.

L'introduzione di questa funzione è necessaria per risolvere due casi critici a cui può andare incontro un agente razionale basato su obiettivi: nel primo caso, possiamo incontrare un conflitto tra obiettivi diversi (arrivare più velocemente o risparmiare carburante); nel secondo caso, l'agente ha a disposizione diversi obiettivi da soddisfare, ma a nessuno di questi è associata una probabilità certa di riuscita.

La funzione di utilità andrà, nel primo caso, a specificare come bilanciare i due obiettivi, mentre nel secondo caso farà corrispondere un certo valore di probabilità a ciascuno dei vari obiettivi, in modo che l'agente possa scegliere quello con probabilità di riuscita maggiore.

La parte più complessa di questi tipi di agente è la programmazione della funzione di utilità, che richiede di tener conto di molti parametri per poter portare alla migliore scelta d'azione in una determinata circostanza.

Tutti gli agenti sopra visti posso essere progettati anche come agenti capaci di apprendere. Il vantaggio principale di questa implementazione è che essi saranno in grado di operare inizialmente anche in contesti non conosciuti: pensiamo ad esempio al taxi a guida autonoma che compie molte ore di guida virtuale per imparare a destreggiarsi nei diversi contesti stradali o ad un cane robot che mappa l'ambiente circostante, per poi riuscire a muoversi con facilità. Inoltre, grazie al processo di apprendimento, con l'aumentare del tempo i programmi sono in grado di migliorare la loro interazione con l'ambiente, diventando capaci di far fronte anche a situazioni impreviste.

3.4 Prendere una decisione

Ogni essere umano nella vita di ogni giorno è chiamato a prendere delle decisioni intorno ai fatti del mondo che gli capitano, a volte imprevisti o imprevedibili. Anche se spesso è un procedimento attuato in modo quasi automatico, il processo che porta alla decisione deve tenere in conto molte variabili, tra le quali gli esiti prevedibili delle diverse azioni e il fine specifico che si vuole raggiungere.

Tra tutte queste, una variabile molto importante che l'individuo deve tenere in considerazione è la presenza o l'assenza di informazioni attorno a ciò che viene scelto. Davanti a una decisione si è spesso davanti alla scelta tra due o più alternative, che devono essere valutate e soppesate tra di loro per individuare quella più adeguata.

Statisticamente, un individuo sceglie l'opzione su cui ha più informazioni, che è cioè in grado di fornirgli un risultato certo. Questa situazione però non si presenta di frequente, in quanto molto spesso ci si trova a dover scegliere tra varie possibilità delle quali non abbiamo una conoscenza completa, ma di cui possediamo solamente informazioni parziali.

Implicitamente, l'individuo associa ad ogni opzione disponibile la probabilità che questa si realizzi una volta scelta: probabilità che dipende dalle informazioni relative alle singole opzioni. In questo modo, le opzioni correlate con più informazioni sono anche quelle che con più probabilità porteranno ai risultati voluti dall'individuo.

Tuttavia l'ambito del ragionamento associato alla scelta umana non si esaurisce qui, cioè in un mero calcolo di probabilità; può accadere infatti che l'individuo scelga un'opzione con scarsissime probabilità di riuscita, ma di cui è fortemente convinto: pensiamo ad esempio ad un genitore che si getta nel fiume per cercare di salvare il figlio trascinato dalla corrente.

Come possiamo notare, già qui è presente uno scarto tra una razionalità "calcolante", capace di trattare solamente l'aspetto quantitativo ma non quello qualitativo, e una razionalità invece più propria dell'uomo, in grado di ricomprendere sia l'aspetto quantitativo che il qualitativo.

La più famosa teoria che si occupa di analizzare il comportamento dell'essere umano è la "teoria dell'utilità attesa"[102] di J. von Neumann[103] e O. Morgenstern. Tale teoria vede l'uomo come un essere razionale, di cui è possibile prevedere il comportamento e studiare le preferenze.

Il criterio con cui l'individuo sceglie tra differenti possibilità è quello che gli consente di ricavarne maggiore utilità. Questa modellizzazione dell'essere umano ha trovato particolare sviluppo all'interno dell'ambito economico e si è poi allargata anche ad altri campi.

Come vedremo, la nozione di razionalità che abbiamo presentato all'inizio del capitolo e il criterio di scelta dell' "agente razionale" in base a una funzione di utilità affondano le loro radici in questa modellizzazione utilitaristica del comportamento umano.

Tale teoria non è esente da problemi, e i suoi punti deboli sono principalmente due: il primo è la non completa congruenza tra il comportamento reale degli individui e quello invece teorizzato dal modello: Nel caso, ad esempio, di giochi che riguardino delle puntate associte con delle probabilità di vincita, gli individui tendono a preferire l'opzione più certa (100% di vincere 1000€) invece di quella più incerta, ma potenzialmente fioriera di maggiori guadagni (80% di vincere 3000€). Il secondo punto debole è che la teoria non si può applicare in contesti ambigui, ma solamente in quelli di rischio[104].

3.4.1 La teoria dell'utilità attesa

Presentata nei primi capitoli di *Theory of games and economic behaviour*, questa teoria illustra un modello economico del comportamento umano. L'obiettivo è di fornire una teoria matematica consistente capace di descrivere e predire le scelte e le preferenze dell'individuo che

[102] Cfr. J. von Neumann, O. Morgenstern, *Theory of games and economic behavior*, Princeton University Press, Princeton 1953.

[103] Neumann e Morgenstern sono conosciuti anche per i loro studi attorno alla teoria dei giochi, che precedono libro in questione. Nella prefazione alla prima edizione del 1943, entrambi gli autori fanno presente che il libro «conterrà un'esposizione e varie applicazioni della teoria matematica dei giochi. […] Le applicazioni saranno di due tipi: da una parte ai giochi nel senso proprio, dall'altra ai problemi economici e sociologici».

[104] La letteratura in merito presenta differenti possibili definizioni dei due termini, ma possiamo riassumerle così: in un contesto di rischio, l'individuo ha delle informazioni che gli consentono di formulare una distribuzione di probabilità simile a quella vera. Invece, in una situazione di ambiguità queste informazioni non sono disponibili e l'individuo agisce senza avere a disposizione una distribuzione di probabilità.

è considerato “razionale” in quanto punta a massimizzare l’utilità o la soddisfazione che può raggiungere attraverso le proprie azioni.

Le azioni analizzate da questa teoria sono le azioni condotte in condizioni di incertezza, in cui un’azione può generare esiti diversi, ognuno associato a una diversa probabilità di accadere. Nel caso in cui l’individuo non si preoccupi del rischio connesso alla scelta, ovvero sia neutrale rispetto al rischio, il criterio che guida le sue azioni si baserà unicamente sul valore atteso, rappresentato dalla media ponderata dei possibili valori monetari moltiplicati per le rispettive probabilità.

Consideriamo ad esempio una persona che voglia investire in titoli azionari parte dei suoi risparmi. Sa che può guadagnare una certa somma A con probabilità *p* e una certa somma $B < A$ con una probabilità complementare rispetto alla prima $(1 - p)$. La formula del valore atteso V_a è data da:

$$V_a = p \cdot A + (1 - p) \cdot B$$

Supponiamo che *p*=0,5 (50% di probabilità), A=2500€, B=450€, allora V_a sarebbe pari a 1475€. Il consumatore potrebbe considerare altri tipi di investimento, calcolare il valore atteso di ciascuno di questi e poi confrontarli tra di loro: il valore atteso più alto rappresenterà l’investimento migliore da fare, cioè quello con rendite alte e perdite basse.

Le considerazioni appena fatte valgono solamente sotto la condizione che l’individuo in questione non tenga in conto il rischio associato alla scelta che andrà a fare. Tuttavia, questo non vale per la maggior parte degli individui; tenendo in conto questo parametro, gli individui andranno a scegliere l’opzione che consentirà una “maggiore utilità attesa” e non più il “maggior valore atteso”. In questo modo la teoria dell’utilità attesa consente di trattare efficacemente l’atteggiamento degli individui anche rispetto al rischio associato alla scelta.

Per fare ciò, è necessario individuare una certa funzione di utilità $u(x)$ che associ al valore “monetario” un valore di utilità. Al variare dell’espressione $u(x)$ è possibile rappresentare i diversi atteggiamenti rispetto al rischio.

Riprendendo l’esempio fatto poco prima, supponiamo di definire $u(x)$ come $u(x) = \sqrt{x}$. Applichiamo quindi la funzione e trasformiamo i valori monetari in valori di utilità, calcolando l’utilità attesa dell’investimento

$$Eu(V_a) = p \cdot \sqrt{A} + (1 - p) \cdot \sqrt{B}$$

In questo caso $Eu(V_a)$ è pari a $\cong 35{,}60$.

Supponiamo ora di avere un investimento C certo (cioè con probabilità 100%) pari a 1475€ e calcoliamo $Eu(C) = \sqrt{1475} \cong 38{,}40$. Possiamo notare che, pur presentando gli stessi valori attesi, $V_a = V_c$, un individuo le cui preferenze sono descritte da $u(x)$, preferirebbe un investimento certo rispetto a uno con una piccola percentuale di rischio.

Questo risultato non è casuale, infatti la scelta della funzione $u(x)$ descrive l'atteggiamento degli individui rispetto al rischio, che può essere di propensione, neutralità o avversione.

La teoria matematica dell'utilità di Neumann e Morgenstern poggia su quattro assiomi principali[105]; denotiamo con $\succ$ "preferibile" e con $\sim$ "indifferente" rispetto alle preferenze di un individuo. Avremo allora:

1. Assioma dell'ordinamento

   Supponiamo di avere due possibili alternative A e B, allora

   $$(A \succ B) \vee (B \succ A) \vee (A \sim B)$$

2. Assioma della transitività

   Supponiamo di avere tre possibili alternative A, B, C, allora

   $$(A \succ B) \wedge (B \succ C) \Longrightarrow (A \succ C)$$

3. Assioma della continuità

   Supponiamo di avere tre preferenze ordinate A≻B≻C, allora è possibile determinare un valore della probabilità *p* in modo che $[p \cdot A + (1-p) \cdot C] \sim B$

4. Assioma dell'indipendenza

   Supponiamo di avere tre preferenze A, B, C, allora

   $$\forall A, B, C \; e \; \forall p, con \; 0 \leq p \leq 1, A \succ B \; sse \; pA \cdot (1-p)C \succ pB \cdot (1-p)C$$

Gli assiomi (1) e (2) sono indispensabili per avere una classificazione delle preferenze degli individui. L'assioma (4) può anche essere definito come "criterio della cosa certa", per cui la preferenza di A rispetto a B non viene modificata dall'introduzione di una terza alternativa C, qualunque essa sia.

Per il teorema di von Neumann e Morgenstern, se questi assiomi sono soddisfatti, allora esiste una funzione di utilità $u(x)$ che assegna a ciascuna preferenza un valore di utilità e conserva le relazioni definite dagli assiomi.

[105] Per una più approfondita esposizione matematica dell'argomento, si consiglia di consultare i capitoli in proposito presenti nel testo originale di von Neumann e Morgenstern.

Data la forte assiomatizzazione, è facile aspettarsi che il sistema proposto per applicarsi a situazioni reali debba indebolire alcuni suoi assiomi, per poter integrare altri fattori che entrano in gioco nei comportamenti effettivi degli individui. Già dopo pochi anni dalla pubblicazione del libro furono prodotti dei paradossi per mostrare come la teoria descritta non sempre andasse d'accordo con l'esperienza reale; i più famosi sono il paradosso di Allais (1953) e il paradosso di Ellsberg (1961). Il primo in particolare critica l'assioma di indipendenza, mostrando come questo venga disatteso da esperimenti con persone in carne ed ossa.

3.4.2 Utilità attesa e agenti razionali

Come abbiamo visto, la teoria dell'utilità attesa modellizza il comportamento umano in condizioni di incertezza e consente di predire, con una discreta efficacia, quale scelta andrà a fare un individuo sotto determinate condizioni. Questo è di importanza centrale nella programmazione di agenti razionali, poiché fornisce un metodo per discernere quali azioni siano da preferire e quali no in un certo contesto.

Poniamo in rilievo due concetti principali che compaiono all'interno della teoria dell'utilità, che approfondiremo nei prossimi paragrafi:

- la funzione di utilità $u(x)$
- Le distribuzioni di probabilità da associare agli eventi possibili che possono risultare dopo l'esecuzione di un'azione.

3.4.2.1 Il teorema di Bayes

Quanto al secondo punto, la teoria dell'utilità si combina con la teoria della probabilità per fornire la teoria delle decisioni, che rappresenta la base teorica per il comportamento degli "agenti razionali".

Abbiamo visto che, nel calcolo del valore dell'utilità è necessario moltiplicare la probabilità dell'evento per la funzione di utilità dell'evento stesso. Tale probabilità tuttavia molto spesso non può essere conosciuta a priori, oppure a volte non si ha proprio a disposizione la distribuzione di probabilità per ogni possibile evento. Questo rappresenta un problema, perché in assenza di questo dato non è possibile calcolare il valore di utilità o, ancora peggio, nel caso di una valutazione errata del valore di probabilità, si avrebbero delle predizioni fallaci.

Per riuscire in parte a risolvere questo problema, si utilizza il teorema di Bayes, che è alla base di tutti i moderni sistemi di inferenza probabilistica, avente equazione

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Dove A e B sono due eventi, P(A) e P(B) sono le probabilità associate a ciascun evento, P(B|A) è la probabilità che B si verifichi sapendo che A si è verificato, P(A|B) è la probabilità che A abbia determinato l'effetto B. A e B possono essere due eventi qualunque, ma è chiaro che debbono essere legati da una qualche relazione di causalità; in caso contrario, la probabilità P(B|A) sarà pari a zero e non ricaveremmo nessuna informazione utile dall'equazione.

Pensiamo ad esempio ad una applicazione di IA in ambito medico: un paziente lamenta un forte torcicollo da giorni (evento B). La macchina sa che nel 70% dei casi la meningite provoca del torcicollo (probabilità P(B|A)). Ha inoltre a disposizione dati indicanti che la probabilità a priori di avere la meningite per un paziente qualsiasi è di 1/50000 (probabilità P(A)), mentre che abbia il torcicollo è pari a l'1% (probabilità P(B)).

La probabilità P(A|B) che un paziente avente il torcicollo abbia la meningite è pari a:

$$P(A|B) = \frac{0.7 \cdot 1/50000}{0.01} = 0.0014$$

Possiamo visualizzare l'equazione scritta poco sopra anche come

$$P(causa|effetto) = \frac{P(effetto|causa)P(causa)}{P(effetto)}$$

Siamo cioè in grado di stimare la probabilità che una certa causa abbia determinato l'accadimento dell'effetto, avendo osservato l'effetto. il teorema di Bayes è frequentemente e variamente applicato: oltre che nell'ambito medico (nel supporto per la diagnosi di malattie e nella verifica dell'efficacia di un farmaco), anche nella categorizzazione degli spam, nel valutare il rischio di credito e nel prevedere vendite future di un'azienda.

La grande potenza del teorema di Bayes all'interno della teoria delle decisioni è la sua capacità di gestire in modo ottimale l'incertezza. È però necessaria una grande quantità di dati per stimare tutte le probabilità che stanno a destra dell'equazione, che siano quanto più simili a quelle reali. Come abbiamo imparato a capire, questo compito può essere svolto dalle applicazioni di IA capaci di analizzare enormi moli di dati e da ciò ricavare informazioni quali relazioni e dipendenze.

Per una trattazione rigorosa dell'argomento di consiglia di consultare il già citato *Intelligenza artificiale, un approccio moderno*[106], di Russel e Norvig, specialmente ai capp. 12-13[107].

3.4.2.2 Questioni sulla funzione di utilità

Consideriamo un agente razionale qualsiasi in cui venga implementata una funzione di utilità $u(x)$. Dalla teoria esposta precedentemente sappiamo che nel prendere una decisione razionale – dove "razionale" indica l'azione che porta alla maggiore utilità attesa –, il programma non deve fare altro che calcolare $Eu(x)$, cioè l'utilità attesa, sommando le utilità dei singoli esiti per la rispettiva probabilità del loro accadere

$$Eu(x) = \sum_{i=1}^{N} p_i \cdot u(x_i)$$

Applicando il principio della "massima utilità attesa", il programma dovrebbe scegliere l'azione che è in grado di massimizzare l'utilità attesa

$$azione = argmax_x Eu(x)$$

Tutto ciò ci offre solamente un metodo astratto per calcolare l'azione migliore, ma non risolve i problemi a cui va incontro l'IA: il principio formalizza il modo di agire, ma non ci offre una traduzione operativa del concetto. In altre parole, la teoria dell'utilità è una teoria normativa, cioè prescrive la regola generale secondo cui gli agenti dovrebbero agire, ma non è affiancata da una parte applicativa.

La questione principale è dunque quella di tradurre la dimensione normativa in indicazioni procedurali attuabili dagli agenti razionali. Come prima cosa bisognerà decidere quali esiti includere nell'ambito della stima dell'utilità attesa, cioè quali conseguenze siano da tenere in conto per una certa azione; classicamente, si possono distinguere le conseguenze intese, cioè quelle propriamente ricercate come fine dell'azione, e le conseguenze previste, che possono seguire collateralmente all'azione ma non volute di per sé. La prospettiva consequenzialista però, non tiene in conto questa distinzione e pone tutte le conseguenze sullo stesso piano. Questo rappresenta un problema nel computo finale: infatti, una conseguenza intesa e una prevista non

[106]Cfr. S. Russell, P. Norvig, F. Amigoni (a cura di), *Intelligenza artificiale, un approccio moderno*.

[107] Per una introduzione sulla statistica si veda anche M. Loreti, *Teoria degli errori e fondamenti di statistica. Introduzione alla fisica sperimentale*, Zanichelli, 1998, pp. 24-25.

possono avere lo stesso peso. Si dovrebbe introdurre un fattore correttivo che sia in grado di tenere conto della differenza.

Sarà poi necessario definire la funzione di utilità $u(x)$, da cui dipenderà anche il tipo di comportamento che l'agente dovrà tenere in relazione agli eventi da affrontare: esso infatti potrà essere tendente al rischio, neutrale o avverso al rischio, a seconda della funzione utilizzata. È chiaro che ciascun agente avrà bisogno di una funzione personalizzata, dipendente dai contesti in cui è chiamato ad agire e al tipo di eventi a cui deve rispondere.

A titolo di esempio consideriamo l'implementazione di un sistema di IA all'interno di un drone aereo da guerra, in grado di svolgere in modo autonomo i compiti di ricognizione, individuazione dei bersagli ed eliminazione degli stessi. Per quanto riguarda gli ultimi due campi d'azione, è chiaro che molta attenzione dovrà essere posta nella determinazione della funzione di utilità e nei valori da attribuire ai singoli esiti, sia che questi vengano inseriti manualmente dai programmatori sia che siano il risultato di un calcolo interno della macchina.

Un *software* con una funzione di utilità che descrive un comportamento propenso al rischio, potrebbe decidere di sganciare un missile per colpire l'obiettivo anche se nell'area circostante sono presenti un discreto numero di civili. Un *software* più conservativo potrebbe invece attendere il momento in cui l'obiettivo si trovi in un luogo più isolato.

In tutto ciò, sarebbe anche necessario quantificare quanta "utilità" debba portare nel computo finale la morte o la sopravvivenza di uno o più individui a fronte del raggiungimento dell'obiettivo. Un ragionamento analogo può essere fatto nei confronti delle tecnologie di IA applicate nel contesto della guida autonoma, specialmente in caso di incidenti che coinvolgano più individui.

Le questioni orbitanti attorno alla gestione di sistemi di IA (sempre più autonomi e con gradi di interazione sempre maggiori) sono sia di natura computazionale che di natura etica. Non si tratta più solamente di macchine relegate in uno spazio fisso, come ad esempio braccia robotizzate usate in catena di montaggio, ma di una costellazione di applicazioni che vengono utilizzate a livello economico, politico e sociale per analizzare dati e ricavarne informazioni, spendibili poi nel mondo reale. Oltre alle già citate applicazioni in ambito automobilistico e militare, pensiamo anche all'ambito medico (in relazione ai trapianti di orgnai) e giudiziario (a supporto di avvocati nell'imbastimento delle cause).

L'autonomia ormai raggiunta da questi sistemi necessita di una regolamentazione etica, dal punto di vista del loro funzionamento e del loro utilizzo (es. diritto alla trasparenza dei risultati,

alla riservatezza, all'affidabilità[108]); ma anche di una discussione attorno allo statuo etico di tali macchine, capendo se queste ultime possano implementare una capacità di ragionamento etico e in che misura.

L'etica dell'intelligenza artificiale si occupa di queste questioni da circa due decenni.

[108] Cfr. *Orientamenti etici per un'IA affidabile,* reperibile su Orientamenti etici per un'IA affidabile | Plasmare il futuro digitale dell'Europa

4 Etica dell'IA

La riflessione etica, almeno nella cultura occidentale, è sempre stata una questione riservata all'ambito del comportamento umano. Pensiamo ad esempio alla riflessione sulla ricerca della felicità e sull'esercizio della virtù presente nell'*Etica Nicomachea* di Aristotele, che esplicita la differenza tra il modo di vivere di piante e animali e il tipo di attività propria della razionalità umana[109]. Solo recentemente sono state affrontate questioni etiche riguardanti il benessere animale o le questioni ecologiche, in riferimento alla sostenibilità di certi comportamenti umani sul mondo. Anche in questi casi, tuttavia, il soggetto morale rimane uno soltanto, cioè l'uomo, che riconosce o meno uno statuto etico ad altri enti.

Dalla fine del XX° secolo, si è affacciato un nuovo protagonista all'interno del panorama dell'etica: l'agente morale artificiale, dotato di uno statuto particolare, poiché non appartenente alla categoria degli "enti di natura" strettamente intesi, ma frutto della produzione tecnologica umana. La recente esplosione dell'IA e le sempre maggiori delegazioni di compiti alle macchine hanno reso necessaria una riflessione etica che veda questi strumenti non più solo come "oggetti", ma anche come "soggetti" morali[110] (in un senso però radicalmente diverso da quello che si può attribuire ai soggetti umani).

Le tecnologie precedenti all'introduzione e all'uso massivo dell'IA comportavano delle questioni etiche riguardanti le macchine – anche quelle informatiche – solo come oggetti, cioè come strumenti usati dall'uomo. Il problema etico in quel caso si poneva in relazione al modo in cui vengono utilizzate le macchine e ai problemi che possono interessare la progettazione delle stesse. Tradizionalmente questa viene definita *computer ethics*. Anche successivamente all'avvento dell'IA, queste questioni rimangono importanti e si pongono ancora con maggior peso nel panorama etico: pensiamo ai problemi di *privacy* dovuti all'accesso e all'analisi di dati personali per addestrare i modelli di IA; all'opacità dei processi decisionali inerenti al *deep learning*; alla presenza di *bias* che possono inficiare l'analisi dei dati di partenza o portare a risultati non neutrali.

---

[109] Cfr. Aristotele, *Etica Nicomachea*, 1097 b30-1098 a5: «è evidente che il semplice vivere è comune anche alle piante, e che quello che si cerca è qualcosa di specifico. Bisognerà dunque escludere anche la vita consistente nel nutrirsi e nel crescere; dopo di questa viene un certo tipo di vita fatta di sensazioni, ma è evidente che anch'essa è comune sia al cavallo che al bue e a tutti gli animali. Allora rimane solo un certo tipo di vita attiva, propria della parte razionale.»

[110] Cfr. si veda al riguardo la *Stanford Encyclopedia of Philosophy* all'indirizzo Ethics of Artificial Intelligence and Robotics (Stanford Encyclopedia of Philosophy).

I sistemi di IA possono però anche essere trattati come "soggetti", a fronte della loro sempre maggiore autonomia di azione ed interazione con gli individui e il mondo reale. Autonomia da intendersi però in senso tecnico, per cui la macchina è capace di svolgere e portare a termine dei compiti senza la continua assistenza o supervisione da parte dell'essere umano. La questione morale si pone nel momento in cui si inizia a delegare ad agenti artificiali un crescente numero di compiti, anche delicati, ad esempio nell'ambito medico, che se svolti dall'uomo chiamerebbero in causa il giudizio morale.

Di questo secondo ambito, quello degli agenti artificiali come possibili soggetti morali, si occupa la *machine ethics*[111], che ha come obiettivo principale quello di rendere le macchine moralmente autonome, capaci cioè di prendere decisioni in maniera esplicita, utilizzando norme morali messe a loro disposizione per formulare giudizi morali complessi. La macchina in questione non dovrà implementare una "funzione etica" implicita, bensì esplicita, di cui potrà fare uso nelle diverse situazioni in cui si troverà ad agire.

La suddivisione tra *computer ethics* e *machine ethics*[112] non è tuttavia da intendersi in modo così rigido. In letteratura, spesso, la seconda viene a caratterizzarsi come l'insieme principale di cui la prima è un sottoinsieme. D'altronde le questioni etiche che riguardano la IA intersecano entrambi gli ambiti: si può intendere infatti l'IA come un oggetto a disposizione dell'uomo, per cui il suo utilizzo dev'essere etico nei limiti legislativi di ciascun Paese (*privacy*, tracciabilità, trasparenza, ecc..), oppure si può considerare l'agente morale artificiale come un "soggetti", cioè in grado di prendere decisione complesse in autonomia

L'ulteriore riflessione circa la traducibilità dell'etica in termini computazionali e la sua applicabilità effettiva, è invece di competenza della *machine metaethics*, che si pone ad un livello di analisi per certi versi complementare a quello precedente. Se da una parte è corretto procedere sperimentalmente, provando a costruire delle macchine che siano in grado di applicare principi morali in modo "autonomo", dall'altra è necessario comprendere se e in che misura l'etica possa essere tradotta in linguaggio informatico, così da delimitare il raggio d'azione di queste macchine.

La questione circa gli agenti morali artificiali riguarda quindi anche la filosofia, che è chiamata a collaborare assieme ad un ampio ambito di materie scientifico-tecnologiche, per contribuire a questo nuovo ambito di ricerca. In particolare sono necessari degli studi filosofici nell'ambito (1)

---

[111] Cfr. L. Cracco, *Machine ethics: stato dell'arte e osservazioni critiche*, "Filosofia", vol. 65, 2020, pp. 75-89.

[112] Chiamata anche *computational ethics*, *machine morality* o *computational morality*.

dell'epistemologia dell'etica e in particolare sul rapporto tra etica dell'uomo e etica delle macchine; (2) della ontologia degli agenti morali artificiali, con particolare riguardo al tema della responsabilità. L'agire autonomo, infatti, implica un'assunzione di responsabilità da parte dell'agente: essa può ricadere integralmente su di esso o può essere "distribuita" tra i vari attori (ad es. programmatori, distributori del *software* di IA, intermediari, ecc..).

Come per la questione del termine di "intelligenza", che abbiamo considerato nel cap. 1, anche la morale è necessario comprendere quale significato e ruolo essa assuma nella dimensione artificiale rispetto alla dimensione propriamente umana, evitando di confondere le due dimensioni e di operare trasposizioni indebite dall'una all'altra.

Come affermato da M. Anderson e S. L. Anderson[113], lo scopo primario della *machine ethics* è quello di «creare una macchina che segua essa stessa un ideale etico o un insieme di principi; vale a dire, che sia guidata da questo o questi principi sui possibili esiti che possono seguire a seguito delle sue decisioni»[114].

Una macchina di questo tipo si definisce agente morale artificiale se, come è prospettato nel documento *Draft Ethics Guidelines for Trustworthy AI*[115], è in grado di «(a) giungere autonomamente a giudizi normativi e a conclusioni, e (b) di agire autonomamente sulla base di questi auto-generati giudizi e conclusioni»[116].

Il campo della *machine ethics* è molto giovane, attivo dagli inizi del 2000, e presenta ancora poche soluzioni rispetto alle questioni che vengono continuamente sollevate all'interno di quest'ambito, e non ha ancora offerto da risultati concreti. Il campo della *computer ethics* è invece attivo fin dagli anni Cinquanta.

Tra gli addetti ai lavori si riconoscono principalmente quattro motivazioni per lo studio e la costruzione di macchine che siano in grado di seguire norme morali per dirigere il proprio comportamento.

La prima[117] riguarda il rapido sviluppo di sistemi autonomi – quali auto a guida autonoma, robot per l'assistenza ai malati o droni da guerra –, che possono essere impiegati all'interno di

---

[113] Cfr. M. Anderson, S. L. Anderson, *Creating an ethical intelligent agent*, "AI Magazine", vol. 28, n. 4, pp. 15-26.

[114] Cfr. *Ivi*, p. 15.

[115] Cfr. si tratta di un documento prodotto da 52 esperti e rappresentati del settore per conto della Commissione Europea per fornire un primo quadro di riferimento per l'etica dell'IA. È disponible su: Ethics guidelines for trustworthy AI | Shaping Europe's digital future

[116] Cfr. *Ivi*, nota 19, p. 13.

[117] Cfr. C. Allen, W. Wallach, I. Smith, *Why machine ethics?*, "IEEE Intelligent System", vol. 21, n. 4, 2006, pp. 12-17.

ambienti particolarmente delicati e critici. L'implementazione di una capacità di ragionamento etico darebbe a questi sistemi la capacità di agire in modo eticamente accettabile, allineandoli ad una serie di valori riconosciuti dalla società nella quale i sistemi di IA saranno chiamati ad operare.

La seconda motivazione risiede nel tentativo di ridurre l'opacità presente intrinsecamente nei processi decisionali dei sistemi di IA che fanno uso di tecniche di *deep learning*. Una macchina in grado di spiegare come sia arrivata a certe conclusioni sarebbe più facilmente integrabile all'interno di un tessuto sociale e consentirebbe di ricostruire il suo processo decisionale.

La terza motivazione[118] risiede nel rapporto sempre più stretto che sembra destinato ad instaurarsi tra l'uomo e i sistemi di IA, per cui questi ultimi potrebbero diventare dei "suggeritori etici", in grado di affiancare gli esperti in alcuni settori sensibili, come ad esempio la medicina o l'ambito giudiziario (dove esistono già applicazioni di questo tipo). Se questo per alcuni può essere un valore aggiunto, poiché la macchina può fornire dati che possono sfuggire al professionista, per altri questo tipo di applicazione rappresenta un'arma a doppio taglio: la delega continua ai sistemi di IA di decisioni eticamente importanti e sensibili potrebbe portare ad un'atrofizzazione della capacità di ragionamento etico nell'uomo.

Come è avvenuto ad esempio con la calcolatrice, per cui ora quasi nessuno è in grado di fare conti a mente in modo rapido, lo stesso potrebbe avvenire per le capacità di giudizio morale. Il problema grave è che una calcolatrice difficilmente "sbaglia" i conti, mentre le applicazioni di IA, già solo nell'interpretazione del linguaggio umano, possono confondersi e generare soluzioni artificiose. In un campo complesso come quello morale, affidarsi totalmente a questi sistemi non significherebbe avere necessariamente una risposta eticamente corretta.

L'ultima motivazione[119] risiede nella convinzione che lo sviluppo della *machine ethics* possa attivare un processo in grado di rivitalizzare lo studio dell'etica "umana", creando un canale di comunicazione tra i filosofi e gli ingegneri. Il problema principale rimane sempre quello della computabilità dell'etica: si tratta di trovare il modo migliore per tradurre degli aspetti valoriali (qualitativi) in entità quantitative, che la macchina possa usare nei suoi ragionamenti.

---

[118] Cfr. *Ivi*., p.12.

[119] Cfr. M. Anderson, S. L. Anderson, *Machine ethics, creating an ethical intelligent agent*, "AI Magazine", vol. 28, n°4, 2007.

4.1 Agenti morali artificiali

Una distinzione ormai divenuta classica nella letteratura, presentata da Moor nell'articolo *Four kinds of ethical robots*[120], è quella secondo cui i robot possono essere raggruppati in quattro principali "classi etiche", a seconda delle loro specifiche capacità[121]:

1. Il primo tipo, il più "debole", è l'*ethical impact agent*, quello cui appartengono agenti le cui azioni hanno conseguenze sia che siano intese che non. In pratica, tutto ciò che di tecnologico ha un impatto nella realtà, può essere considerato un agente di questo tipo. Secondo questa definizione, potenzialmente ogni tipo di macchina può ricadere in questa prima classe: un termostato digitale ha il compito di mantenere costante la temperatura nelle stanze della casa: in caso di guasto, la temperatura scenderebbe e gli inquilini ne risentirebbero. Inserire una macchina in questa classe non ci dà quindi molte informazioni su che tipo di agente etico essa sia.
2. Nella seconda classe ricadono gli "agenti etici impliciti" (*implicit ethical agent*), quelli per cui le norme etiche sono inserite dai programmatori all'interno del codice e che agiscono quindi in accordo con queste norme, ma senza averne una rappresentazione esplicita. Moor afferma che questi agenti posseggono una specie di "virtù integrata", che da adito solamente a una scelta di costruzione o di programmazione fatta a monte. Applicazioni di questo tipo possono essere i robot per la sicurezza o la protezione, come ad esempio i sistemi di allerta degli aerei in caso di quota troppo bassa o di pericolo di collisione. In questo caso il sistema presenta nel codice una "funzione etica" implicita, per la quale è associato un valore di pericolo alla discesa al di sotto di una certa quota; raggiunta questa, il sistema fa scattare l'allarme in modo automatico. Il funzionamento si basa su una classica formula condizionale del tipo *if-then*.
3. Il terzo tipo di agente, l'agente etico esplicito (*explicit ethical agent*), è quello più interessante dal punto di vista dell'etica dell'IA. In questo caso, la macchina è in grado di identificare e processare informazioni etiche presenti nell'ambiente e di formulare giudizi etici, per poi agire in accordo con essi. I principi etici devono essere codificati in modo esplicito all'interno della macchina, per fare in modo che questa li possa utilizzare per risolvere dilemmi etici e prendere decisioni allineate a questi principi.

[120] Cfr. J.H. Moor, *Four kinds of ethical robots*, "Philosophy Now", vol. 72, 2009, pp. 12-14.
[121] Cfr. *ivi*, p. 12.

Vedremo successivamente in che modo questi principi possono essere integrati nella macchina e come essa possa apprenderli per poi usarli in autonomia.

4. L'ultima classe è quella degli "agenti etici completi" (*full ethical agents*), che sono in grado di compiere tutto ciò che è in grado di fare la classe precedente con l'aggiunta di alcune "caratteristiche metafisiche"[122], quali la coscienza, l'intenzionalità e il libero arbitrio. L'unico esempio presente in questa classe è l'uomo, cioè l'unico dotato di quelle caratteristiche peculiari viste poco sopra. La questione principale, come notato anche da Moor, sta nel capire se c'è un modo per collegare tra loro la terza e la quarta classe di agenti etici, o se invece c'è una distanza incolmabile tra le due classi, per cui il massimo che possiamo aspettarci è di riuscire a progettare degli agenti artificiali che siano degli agenti etici espliciti.

Come giustamente notato da F. Fossa[123], nel passaggio tra il terzo e il quarto tipo di agente morale artificiale è facile cadere nel paradigma della continuità, cioè ritenere che tra agenti umani e artificiali non sia presente una differenza sostanziale, ma vi sia anzi un tipo di continuità essenziale che non necessita di essere ulteriormente indagata e approfondita, sostenendo che la riproduzione tecnologica può cogliere la sostanza del comportamento morale umano. Il paradigma continuista non vede la differenza ontologica tra umano e artificiale, poiché si basa in ultima instanza su una forma di riduzionismo.

L'avvertimento di Fossa[124] è quello di non scivolare dall'analogia all'omologia nella considerazione dell'agire etico. Nel primo caso vi è una "somiglianza nella differenza", in cui le differenze sono conservate nel rapporto tra i due poli; nel secondo caso, invece, l'attenzione è posta solamente sulle somiglianze, mentre le differenze vengono lasciate in secondo piano, quasi che queste fossero alcunché di contigente. Il secondo caso è un'espressione del paradigma continuistico e, per certi versi, anche di quello riduzionistico: l'annullamento delle differenze tra umano e artificiale rischia di ridurre l'etica ad un problema solamente computazionale, in cui l'uomo si troverebbe ad essere solamente una sofisticata "calcolatrice morale": in quanto tale, replicabile.

[122] Cfr. *ivi*, p. 12.

[123] Cfr. F. Fossa, *Etica funzionale. Considerazioni filosofiche sulla teoria dell'agente morale artificiale*, "Filosofia", vol. 65, 2020, pp. 91-106.

[124] Cfr. *ivi*, p. 95-96, nota 20.

Come abbiamo visto anche per la definizione di "intelligenza artificiale", allo stato attuale gli addetti ai lavori sono consapevoli che la riproduzione tecnologica è ben lontana dal riuscire anche solo ad emulare in modo soddisfacente una parte del ragionamento etico umano, poiché i processi che regolano i *software* sono completamente differenti da quelli biologici presenti nel cervello umano (senza contare la difficoltà di implementare un tipo di ragionamento non solo computazionale, ma anche "affettivo", come quello usato dall'uomo per compiere le sue scelte morali).

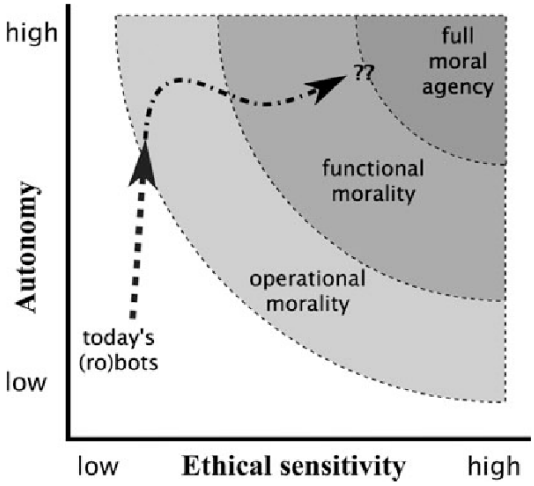


Figura 8

È necessario però focalizzarsi sulla differenza sostanziale che intercorre tra "artificiale" e "umano" anche nel campo morale, poiché solamente una volta messa a fuoco questa differenza sarà possibile individuare i limiti degli agenti morali artificiali e definire il loro corretto campo di applicazione.

Per visualizzare meglio in che modo intendere la differenza tra agenti etici umani ed artificiali, possiamo considerare la fig. 8[125], in cui Wallach e Allen considerano come variabili principali la "sensibilità etica" e l'autonomia. Al punto di origine degli assi possiamo collocare quei prodotti umani che non presentano nessuno dei due aspetti, come ad esempio un martello non è autonomo né presenta alcuna sensibilità a valori morali. Al centro potremmo porre quelli che abbiamo definito precedentemente come *ethical impact agent* (la prima classe di agenti individuata da Moor).

Nell'insieme dell'*operational morality* trovano posto gli agenti morali impliciti, nel *functional morality* gli agenti morali espliciti e nell'ultimo insieme, *full moral agency*, trovano posto gli

[125] Cfr. W. Wallach, C. Allen, *Framing robot arms control*, "Ethics and Information Technology", vol. 15, 2013, pp. 125-135.

“agenti morali completi”, il cui solo esponente è l’essere umano. Come si può notare, la difficoltà dello sviluppo di un’etica per gli agenti è esemplificata anche dal grafico: mentre è relativamente semplice aumentare l’autonomia delle macchine nello svolgimento dei compiti (lungo l’asse y si ha una crescita molto più veloce e lineare), assai più lenta è l’implementazione della sensibilità etica all’interno degli stessi agenti, che presenta una crescita non lineare e più lenta.

La rappresentazione delle due variabili su di un medesimo piano cartesiano[126] le tratta però come invarianti rispetto al tipo di agente morale cui le si riferisce, quasi che la loro variazione sia di tipo quantitativo. Questa rappresentazione prevede dunque una continuità delle variabili per tutti i vari tipi di agenti, per cui sviluppando maggiormente la “sensibilità etica” si passa con conitnutià da una agente all’altro. Il presupposto è che “autonomia” e “sensibilità etica” abbiano il medesimo significato e si configurino allo stesso modo rispetto a qualunque tipo di agente vengano riferite, senza presentare alcun cambiamento qualitativo tra un agente e l’altro.

Come fa notare ancora una volta Fossa[127], la differenza che intercorre tra gli agenti “espliciti” e quelli “in senso pieno” è il diverso rapporto che questi due tipi agenti hanno rispetto ai valori. I primi infatti sono in grado solamente di “comportarsi secondo valori”, cioè di agire in modo procedurale secondo dei valori che vengono assegnati loro, e che ci aspettiamo vengano soddisfatti nel corso del funzionamento della macchina. Anche attraverso il *machine learning*[128], quello che i modelli di IA trovano nei dati non sono valori in sé, ma dei parametri che solamente dal punto di vista dell’uomo possono essere interpretati come aventi una valenza morale. Ciò significa che alle macchine in quanto tali non pertiene alcuna determinazione dei valori, ma solamente l’agire “passivo” rispetto ad essi.

Gli esseri umani invece, gli unici ad essere “agenti etici completi”, presentano una bidimensionalità nel loro agire morale: essi sono in grado sia di agire secondo valori, ma sono anche in grado di determinarli liberamente e di riconoscerli come tali. Negli uomini il rapporto con i valori è cioè diretto, e non mediato come succede nelle macchine; e quindi «l’esperienza morale umana è costituita dallo sforzo di vivere affermando ciò che ci sta a cuore e opponendoci a ciò che riteniamo inaccettabile, il che è inseparabile da un’interrogazione su cosa sia ciò che ci sta a cuore e ciò che riteniamo inaccettabile»[129]. Sottolineiamo come quelle appena presentate siano dimensioni e non semplicemente “parti” dell’agire morale prettamente umano: la

[126] Cfr. F. Fossa, *Etica funzionale*, p. 98.
[127] Cfr. F. Fossa, *Etica funzionale*, pp. 99-100.
[128] Cfr. *Ivi*, p. 99, nota 30.
[129] Cfr. *Ivi*, p. 99.

mancanza di una dimensione, come nel caso delle macchine, comporta una configurazione completamente diversa dell'agire artificiale rispetto a quello tipicamente umano. Le due dimensioni già considerate - quelle, dell'"agire secondo valori" e quella del "determinare i valori" – non sono nell'uomo due parti che si "incastrino" o si "sommino" assieme: piuttosto si implicano vicendevolmente. Nel nostro agire è sempre presupposto un ragionamento attorno ai valori che ci guidano, altrimenti qualunque azione ci sarebbe indifferente e non ci porremmo nemmeno la domanda relativa a ciò che è giusto fare. Viceversa, la nostra determinazione dei valori implica una certa adesione ad essi, attraverso un atto immanente (ad esempio scegliere di essere onesti), che può anche tradursi in un atto transitivo ad esempio, riconsegnare un portafoglio al proprietario). Nel momento in cui agiamo, scegliamo di aderire a una qualche linea di condotta che guidi la nostra azione, cioè ad un valore, che riconosciamo come tale. Questa implicazione tra le due dimensioni non è invece presente negli agenti morali artificiali, in cui le dimensioni sono ridotte a parti, che possono essere quindi "assemblate" o disassemblate in moneti diversi. In questo modo, però, viene a mancare quell'integrazione originaria tra le due dimensioni, che non è possibile riprodurre artificialmente, ovvero la cui combinazione potrà essere solo alcunché di contenigente.

Come vedremo in maniera più approfondita nel capitolo dedicato a Tommaso d'Aquino, l'uomo è costitutivamente orientato, nell'esperienza morale, verso il Bene come tale, secondo una "volizione radicale". L'apertura trascendentale della volontà e ella ragione gli consente la "determinazione dei valori" di cui parla Fossa e quella interrogazione critica circa la validità dei valori stessi. Nella macchina, invece, non è presente nulla di tutto ciò, tanto che in assenza di istruzioni essa rimarrebbe inerte o, al raggiungimento di un obiettivo, si fermerebbe in attesa di ulteriori istruzioni.

Nell'analisi qui condotta rimane comunque aperta la possibilità della progettazione di un "agente etico esplicito" che obbedisca ad una "morale funzionale", mentre viene esclusa la possibilità di un agente morale completo di tipo artificiale.

4.2 Principali questioni relative alla *machine ethics*

Una delle prime difficoltà che incontra chi vuole implementare abilità relative al ragionamento morale in agenti artificiali è quella relativa al peculiare statuto epistemologico dell'etica.

Dall'etica infatti non è possibile aspettarsi la stessa certezza di risultati e lo stesso rigore di metodo che si possono avere, ad esempio, nella matematica o nella fisica. Nella lunga storia della filosofia morale si sono infatti succedute molte teorie etiche, ognuna con le proprie caratteristiche peculiari, a volte contrastanti le une con le altre. La pluralità di metodi e risultati nell'etica non toglie alcuna dignità alla filosofia morale, che come tale si occupa di una delle dimensioni più proprie dell'uomo, cioè del suo rapporto eminentemente pratico col mondo e con gli altri individui.

Tale pluralità di metodi e principi si ritrova anche nel panorama dell'etica dell'IA: un ipotetico agente morale artificiale potrebbe implementare un'etica di stampo kantiano, oppure essere programmato per allinearsi coi valori della cultura islamica o agire in accordo con la *golden rule*. Ognuna di queste proposte presenta modi differenti di formalizzare l'etica e l'agire morale. Nell'implementazione di regole morali è necessario tenere conto anche di questa eterogeneità, per creare macchine allineate culturalmente ai diversi contesti socio-culturali, pena il rischio di cadere in un "suprematismo etico", con il risultato di produrre macchine eticamente allineate solamente ad una cultura specifica.

Un'altra difficoltà[130] che si incontra è quella di come tenere in conto delle altre dimensioni propriamente umane, specialmente quelle legate all'ambito delle emozioni e dei sentimenti. Spesso si sostiene che per compiere l'azione moralmente giusta in uno specifico contesto non è sempre necessario valersi di (o essere dotati di) una dimensione "passionale" o affettiva, perché questo potrebbe offuscare la adeguatezza del giudizio (se una persona cede facilmente all'ira, dovrebbe astenersi dal prendere decisioni nel momento in cui questa si manifesti).

Alcuni sostengono addirittura che l'assenza di emozioni nelle macchine potrebbe aiutarle a prendere decisioni a "mente fredda" e con risultati migliori rispetto a quanto farebbe un essere umano nella stessa situazione.

Questo può essere certamente vero per quelle decisioni che anche nell'uomo vengono prese di riflesso rispetto ad uno stimolo esterno: pensiamo ad esempio ad una sterzatura improvvisa per evitare di tamponare la macchina di fronte che ha frenato improvvisamente. In ciò, la capacità di reazione e di previsione è sicuramente più veloce ed efficiente in una macchina, che in un essere umano.

[130] Cfr. M. Anderson, S. L. Anderson, *Machine ethics: creating an ethical intelligent agent*, pp. 18-19.

Discorso diverso nel caso di macchine impiegate nell'ambito ospedaliero. Certamente è possibile, come è già stato fatto[131], implementare i principi base dell'etica medica (non maleficenza, beneficenza, rispetto dell'autonomia del paziente, ecc…) all'interno di un *software* che si occupi di ricordare ai pazienti quando prendere le medicine, preservando l'autonomia delle scelte del paziente. Ma il rapporto paziente-dottore può essere delegato anche nei suoi aspetti più delicati ad una macchina? In questo campo le applicazioni di IA dovranno essere piuttosto un supporto alla pratica medica, anche durante il processo decisionale; ma le considerazioni finali dovranno essere di competenza del medico, che non obbedisce solamente ad un processo statistico e algoritmico per determinare quale azione compiere.

Un'ultima questione che si pone, vista la sempre maggiore autonomia dei sistemi di IA, è l'attribuzione di responsabilità per le conseguenze delle azioni. Per un essere umano che compie un'azione cattiva, questa attribuzione è abbastanza semplice, poiché si è in presenza di intenzionalità, coscienza e libero arbitrio (salvo casi eccezionali).

Per una macchina invece il discorso non è così semplice, poiché essa non può rispondere per sé stessa (non è cioè un soggetto nel senso pieno del termine), ma a rispondere per essa dev'essere una costellazione di attori, quali ad esempio programmatori, distributori, manutentori, ecc…

Il problema è prima di tutto di tipo sociale: in un futuro prossimo in cui molte automobili includeranno sistemi di assistenza alla guida in cui è consentito al conducente distrarsi e togliere le mani dal volante, in caso di incidente sarà necessario determinare le responsabilità dell'accaduto; ma ciò varrà anche nel caso in cui un *software* di IA sbagli a determinare il risultato di una lastra di un paziente e infici il suo processo di guarigione.

M. Anderson e S. L. Anderson affermano[132] che, riguardo a questo problema, non si è tenuta in considerazione una distinzione importante: quella

> tra compiere l'azione moralmente giusta in una data situazione, inclusa la possibilità di essere in grado di giustificarla appellandosi a un accettabile principio etico, e essere ritenuti moralmente responsabili per quell'azione. L'intenzionalità e il libero arbitrio saranno in un certo senso necessari per ritenere un ente moralmente responsabile per le sue azioni. […] Ma nessuno di questi attributi è necessario per compiere l'azione morale giusta in un dilemma etico e per giustificarla.

[131] Il programma sviluppato ha nome EthEl.
[132] Cfr. M. Anderson, S. L. Anderson, *Machine ethics*, p.19.

Gli autori proseguono spiegando che per una macchina tutto ciò che è richiesto per compiere un'azione moralmente giusta è che essa «agisca in un modo che si conformi con quella che sarebbe giudicata come l'azione moralmente giusta e sia in grado di giustificarla facendo appello ai principi etici che sta seguendo»[133].

Da queste poche righe è chiaro, almeno da un punto di vista filosofico, come l'ambito della *machine ethics* abbia bisogno di definire precisamente i suoi confini e i suoi limiti e di comprendere in che cosa saranno impiegati gli agenti razionali artificiali. Affermare che l'intenzionalità e il libero arbitrio non sono necessari né per compiere l'azione giusta né per giustificarla, significa ignorare l'intera riflessione classica attorno all'agire dell'uomo e all'antropologia filosofica.

Di più, significa aver ridotto l'etica a una teoria traducibile computazionalmente, in cui si comparano solamente i risultati (la *performance*) ottenuti dall'essere umano e dall'ente artificiale: se questi risultano tra loro congruenti, allora se ne conclude che l'intenzionalità e il libero arbitrio non sono necessari per la determinazione della azione moralmente giusta. Se non che la differenza tra i due casi risiede ancora una volta nei processi che guidano l'agire umano rispetto all'agire artificiale. Più in generale, l'atto umano non può essere ridotto alla *performance* esteriore in cui esso si esprime.

4.3 Approcci all'etica computazionale

Nel progettare un agente morale artificiale si pongono principalmente due tipi di questioni, che risultano intrecciate tra loro: in primo luogo la determinazione di quale teoria normativa sia la più adatta da implementare con le macchine; in secondo luogo stabilire quale sia la modellizzazione della teoria più adatta a imitare le capacità morali dell'uomo.

Riguardo alla prima questione[134], i tentativi finora compiuti hanno utilizzato teorie etiche quali l'utilitarismo di Bentham, il consequenzialismo, il deontologismo nella sua declinazione kantiana (riferentesi all'imperativo categorico), la teoria dei doveri *prima facie* dell'etica di Ross (che presenta elementi appartenenti sia alle teorie teleologiche che deontologiche) e l'etica delle virtù di stampo aristotelico (*virtue ethics*).

---

[133] Cfr. *ibidem*.

[134] Si veda a proposito M. Anderson, S. L. Anderson, *Machine Ethics*; C. Allen, G. Varner, J. Zinser, *Prolegomena to any future artificial moral agent*, "Journal of Experimental & Theoretical Artificial Intelligence, vol. 12, 2000, pp. 251-261; W. Wallach, C. Allen, I. Smit, *Machine morality: bottom-up and top-down approaches for modelling human moral faculties*, "AI & Society", vol. 22, 2008, pp. 565-582.

Ma i tre filoni principali presi in considerazione sono: il deontologismo cui vengono ricondotti teleologismo di cui fa parte l'utilitarismo e il consequenzialismo, nel deontologismo (Kant), l'approccio cosidetto "misto" (Ross) e l'etica delle virtù.

Ogni teoria, oltre ad avere i propri punti forza, presenta anche delle difficoltà specifiche nel tradursi in una teoria effettivamente utilizzabile da un agente morale artificiale. Attualmente l'approccio misto si pone come la migliore fra le opzioni, poiché è in grado di tenere in conto sia la dimensione dei principi che la dimensione più "mondana", in cui i principi non possono essere sempre applicati in modo rigido e in cui ogni situazione è diversa dall'altra.

4.3.1 L'approccio utilitarista

Tra quelle che abbiamo presentato, la teoria che risolve parzialmente il problema della computazionabilità dell'etica è l'approccio utilitarista, per cui la determinazione dell'azione moralmente giusta in una data situazione è quella che valuta l'azione massimizzando la quantità totale di "utilità" ottenuta per il maggior numero di persone (dove utilità indica il benessere). In questa prospettiva, ciò che porta al benessere è piacevole, e ciò che è piacevole è bene. Tutto ciò che non è piacevole è male, ed è quindi da evitarsi.

Dato questo semplice principio di divisone tra ciò che è bene, e quindi da ricercare, e ciò che è male, e quindi da evitare, è possibile calcolare quale azione sia preferibile ad un'altra tramite una "aritmetica morale". Per fare ciò è necessario assegnare a ciascuna conseguenza un valore numerico positivo o negativo, proporzionale alla bontà della conseguenza stessa.

Bentham introduce sette variabili quantitative principali[135] da tenere in conto per valutare una conseguenza: (1) l'intensità del piacere provocato, (2) la sua durata, (3) la certezza o incertezza del suo raggiungimento, (4) l'estensione del piacere, (5) la velocità del raggiungimento, (6) la fecondità e (7) la purezza[136].

Traducendo la (1), la (2) e la (3) in una semplice formula matematica, otteniamo che il piacere (o utilità) totale di un'azione sarà data da

$$\sum(intensità \cdot durata \cdot probabilità) \times singola\ conseguenza$$

Si valutano le variabili in parentesi per ogni possibile conseguenza che l'agente si aspetta segua all'azione, e si ottiene così un valore che rappresenta la "bontà" dell'azione.

[135] Cfr. M. Quinn, *Bentham on mensuration: calculation and moral reasoning*, "Utilitas", vol. 26, n. 1, 2014, pp. 61-104.

[136] Fecondità intesa come la possibilità che seguano altre sensazioni piacevoli dopo l'azione. La purezza invece è la probabilità che a un'azione non segua una sensazione opposta.

Quando dobbiamo compiere una scelta, dobbiamo calcolare per ogni opzione possibile[137] questo valore, e poi confrontare i rispettivi risultati: l'azione con valore più alto sarà quella da preferire, in quanto massimizza il piacere (o utilità) attesa.

Se ritorniamo al capitolo in cui abbiamo trattato degli agenti razionali e della teoria dell'utilità attesa, vediamo che le due teorie sono molto simili, a riconferma che la teoria utilitaristica ben si presta a una implementazione negli agenti morali artificiali.

Uno dei problemi maggiori per questa teoria[138], specialmente nella sua declinazione consequenzialistica[139], risiede nel fatto che per massimizzare l'utilità, si potrebbe decidere di sacrificare una vita umana (pensiamo ai droni militari o alla scelta dei destinatari per la donazione di organi) o commettere altro genere di azioni non moralmente accettabili secondo prospettive più consolidate[140]. Inoltre bisogna tenere conto di quali conseguenze far rientrare all'interno del computo: se solo quelle intese, anche parte di quelle attese, o tutte indiscriminatamente.

Accanto a questa teoria è dunque necessario affiancarne un'altra che non tenga solamente conto delle conseguenze, ma si rifaccia anche a dei principi inviolabili (almeno *prima facie*).

---

[137] Se l'azione ha ricadute anche su altri soggetti, allora il calcolo delle conseguenze dev'essere compiuto anche riguardo a tutti questi ultimi. È chiaro che all'aumentare dei soggetti coinvolti, aumenterà anche la complessità dei calcoli.

[138] Distinguiamo la posizione prettamente consequenzialistica, rispetto a quella benthiana, poiché Bentham si rifà anche alla lezione classica dell'etica; infatti indica, rispetto a quelli elencati precedentemnte, altri sei elementi da tenere in considerazione in ogni azione: 1) l'atto in sé, 2) le circostanze, 3) l'intenzione, 4) la consapevolezza dell'agente, 5) il motivo che l'ha generata e 6) la disposizione generale. La riduzione dell'utilitarismo di Bentham al semplice calcolo delle conseguenze è perciò indebita, anche se frequente.

[139] Per il consequenzialismo non c'è differenza tra conseguenze intese e conseguenze previste: le prime sono quelle che l'agente può ottenere (ad esempio la guarigione di un paziente attraverso una delicata operazione chirurgica); le seconde possono seguire l'azione, ma non sono connesse necessariamente con l'azione in sé (nel corso dell'operazione, il paziente può morire; tuttavia l'intenzione del chirurgo era di salvarlo e non di danneggiarlo).

[140] Gli esempi prendono come riferimento delle applicazioni avanzate di IA. Tuttavia una serie di tali azioni è già stata compiuta in passato, basti pensare alle bombe atomiche sganciate sopra Hiroshima e Nagasaki per costringere il Giappone alla resa. In quel caso il valore della conseguenza "fine della guerra" è stata valutata superiore al valore della conseguenza "morte di migliaia di persone".

4.3.2 L'approccio deontologico e l'approccio misto

Tradizionalmente[141] l'approccio deontologico si fa riferimento a Kant e alla sua riflessione nella *Critica della ragion pratica* riguardo all'imperativo categorico e alla legge morale.

In questo caso, l'azione per essere considerata moralmente buona deve soddisfare le condizioni dell'imperativo categorico; Kant ne presenta tre formulazioni, noi riportiamo la prima che afferma: «agisci come se la massima della tua azione potesse diventare per mezzo della tua volontà una legge universale di natura»[142].

Tralasciando momentaneamente il senso che concetti quali "massima" e "volontà" possono assumere per una macchina, facciamo notare che la formulazione dell'imperativo categorico kantiano non prescrive in che modo specifico agire in una particolare situazione, ma dà solamente la forma a cui deve uniformarsi ogni azione che si voglia considerare "buona".

Una possibile applicazione dell'approccio kantiano all'etica delle macchine si tradurrebbe quindi in prima battuta nella determinazione di alcuni valori inviolabili, a cui l'agente morale artificiale dovrebbe necessariamente rifarsi prima di decidere quale azione compiere. In questo modo si ovvierebbe alla difficoltà del metodo basato solamente sull'approccio utilitarista, che trattava i valori come variabili dipendenti dalle conseguenze dell'azione. Nell'approccio kantiano, invece, i valori diventano variabili indipendenti.

Quello che iniziamo a intravedere è la ristretta autonomia che le macchine potranno avere nel campo morale: i valori principali secondo cui esse agiranno saranno imposti dall'esterno, cioè dai programmatori. In questo modo, seguendo Kant, possiamo a ragione ritenere la macchina come un'entità dotata di una morale eteronoma e non propriamente autonoma (quale è possibile solamente all'essere umano). Il progetto di implementare artificialmente l'etica kantiana ci indica quindi anche una prima differenza sostanziale tra la macchina e l'uomo, che deve essere tenuta in considerazione: solo l'uomo è in grado di darsi dei valori (e di attuare una loro verifica nella applicazione universale), la macchina al massimo può seguire i valori (o al massimo "calcolarne" di nuovi, ma sempre a partire da quelli già dati).

Uno degli scogli principali in cui ci si imbatte nell'utilizzo della teoria kantiana riguarda la gestione del conflitto tra doveri assoluti, reso celebre dal carteggio tra Constant e Kant[143], che

---

[141] Cfr. W. Wallach *et alii*, *Machine Morality*, p. 253 e ss.

[142] Cfr. I. Kant, F. Gonnelli (traduzione), *Fondazione della metafisica dei costumi*, Laterza, Bari 2018, p. 75.

[143] Si veda a proposito I. Kant, B. Constant, S. M. Carmignani (a cura di), *Il diritto di mentire*, Passigli Editori, Firenze 2008.

porterebbe la macchina a bloccarsi e non riuscire a decidere quale azione mettere in atto. Una macchina programmata secondo un'impostazione kantiana non sarebbe perciò in grado di far fronte alla complessità delle situazioni reali.

Per evitare questa eventualità si ricorre all'"approccio misto"[144], basato sulla teoria di Ross, che sostituisce ai doveri assoluti di stampo kantiano i doveri *prima facie*, che in caso di conflitto tra doveri possono essere violati in nome di obbligazioni più stringenti. Si ottiene così un metodo che riesce a tenere insieme sia la teoria utilitaristica, basata principalmente sul valore delle conseguenze, che la teoria deontologica kantiana, che guarda al soddisfacimento dei principi.

4.3.3 L'etica delle virtù

L'ultima teoria presa in esame è quella esposta da Aristotele nell'*Etica Nicomachea*, in cui particolare attenzione è dedicata alla definizione e all'analisi delle caratteristiche della virtù. Essa è definita come «uno stato abituale che produce scelte, consistente in una medietà rispetto a noi, determinato razionalmente»[145].

Questo approccio, pur essendo diverso da quelli precedenti, presenta le stesse difficoltà computazionali che abbiamo considerato nell'approccio deontologico, quali il conflitto tra virtù (ad esempio tra onestà e compassione) e la difficoltà di riconoscere quale virtù applicare in un contesto reale da parte di una agente morale artificiale. Considerando ad esempio l'onestà, la macchina dovrebbe essere in grado di comprendere quando applicare questa virtù, in che misura ed in che contesto. Ora questo non è semplice nemmeno per l'uomo, ma diventa addirittura impossibile a tradursi in un algoritmo adatto per la macchina.

Tuttavia per le applicazioni di IA la teoria aristotelica può risultare utile in un'ottica di apprendimento rinforzato. Possono essere infatti implementati degli atteggiamenti virtuosi all'interno della macchina e, nel momento in cui questa agisse in accordo con queste disposizioni, potrebbe ricevere un feedback positivo; negativo altrimenti. In questo modo non si fornisce alla macchina un comportamento completamente predeterminato, ma esso viene continuamente affinato attraverso l'esperienza di casi simili tra loro. Questa modalità di apprendimento dell'abito virtuoso ben si adatta al modo in cui funzionano i *software* di IA.

---

[144] Cfr. M. Anderson, S. L. Anderson, *Machine ethics*, pp. 18-19.
[145] Cfr. Aristotele, *Etica Nicomachea*, Laterza, Bari, 2018 p. 63, 1106b 30 – 1107a 5.

4.3.4 Approcci *top-down, bottom-up* e *misto*

Dopo aver spiegato quali sono le teorie etiche che si possono candidare maggiormente a ricoprire un ruolo di primo piano all'interno dello sviluppo dell'etica per le applicazioni di IA, possiamo analizzare brevemente quale sia il metodo più adatto per riuscire a modellizzare le capacità etiche umane.

Esistono principalmente tre approcci: *top-down*, *bottom-up* e misto, quest'ultimo includente elementi sia del primo che del secondo approccio.

Con l'espressione "*top-down*" possiamo intendere due diversi significati[146]: in un senso ingegneristico, tale approccio si occupa di analizzare e suddividere un compito (*task*) in suoi sottocomponenti più semplici che possono essere direttamente implementati e organizzati per ottenere il risultato desiderato.

Nel senso etico del termine intendiamo quell'approccio che prende come riferimento una teoria normativa generale (ad esempio il deontologismo kantiano o la "*golden rule"*) e ne ricava le conseguenze da applicare poi ai casi particolari. Potremmo definirlo anche come un approccio "deduttivo" alle teorie etiche, che ben si applica alle modalità di funzionamento dei sistemi esperti.

Per la programmazione degli agenti morali artificiali, questo significa considerare una teoria etica, enuclearne gli elementi nodali e utilizzarli per la progettazione degli algoritmi che andranno ad implementare tale teoria nel programma.

Con l'espressione "*bottom-up*"[147], nel campo dell'ingegneria ci si riferisce all'approccio per cui la teoria implementata serve solamente a fornire un obiettivo al *software* (es. "agisci massimizzando il bene comune"), ma non si occupa di specificare in concreto i passaggi da seguire per arrivare a quel risultato. Il *software* utilizza varie tecniche di tentativi ed errori per raggiungere l'obiettivo in maniera autonoma, sviluppando un proprio metodo di azione.

Si tratta di un approccio che ben si sposa con il principio di funzionamento dell'IA basata sulle reti neurali: si fornisce un obiettivo e un ampio set di dati al *software* e questo, dopo un periodo di apprendimento, avrà individuato quale azione sarà la migliore da compiere per raggiungere quello specifico risultato. Una delle principali difficoltà di questo approccio sta nell'incapacità dei sistemi di gestire più valori (ossia obiettivi), e nella scelta di quale sia il più adatto per giungere

[146] Cfr. W. Wallach *et alii*., *Machine morality*.
[147] Cfr. *ivi*.

ad un certo obiettivo. In questo senso, questi metodi lavorano bene su problemi limitati e ben specificati.

Sia l'approccio *top-down* che quello *bottom-up* si trovano a gestire due diversi problemi, che possiamo considerare complementari l'uno all'altro: il primo ha difficoltà a gestire i casi particolari, poiché non è possibile derivare da alcune norme generali l'azione giusta da compiere in ogni circostanza. Potremmo definirlo un approccio "rigido", che non è in grado di far fronte all'estrema variabilità delle situazioni a cui l'agente morale artificiale si troverebbe esposto.

L'approccio *bottom-up*, invece, potenzialmente può risolvere situazioni complicate e inaspettate, ma trova difficoltà nel gestire un grande numero di valori etici. Esistono ovviamente applicazioni reali dei due approcci che svolgono bene il proprio compito (ad esempio i suggeritori di diagnostica in medicina), ma per i futuri sviluppi della *machine ethics* gli studi si stanno orientando verso un approccio misto.

Si cerca così di tenere insieme le parti migliori dei due approcci: quello *top-down*, più rigoroso e "deduttivo", capace di fare riferimento ai valori morali principali, e quello *bottom-up*, in grado di far fronte alla complessità del reale, con un approccio più flessibile.

4.4 Orientamenti etici per un'IA affidabile

Il documento[148] prodotto da un gruppo di esperti di IA nel 2018 per conto della Commissione Europea rimane ancora oggi uno dei più completi sull'argomento: esso fornisce obiettivi chiari per lo sviluppo dell'IA nei prossimi anni.

Si tratta di orientamenti, che non hanno come tali carattere prescrittivo per le aziende, ma che risultano utili per mettere a fuoco quali siano i diritti che le applicazioni di IA saranno chiamate a tutelare e quali strumenti dovranno essere implementati per consentire ciò.

Si tratta in particolare di sviluppare un'IA affidabile, secondo tre pilastri principali[149]: (1) la legalità, cioè l'osservanza delle leggi in vigore, (2) l'eticità, quindi l'adesione a principi e valori etici e (3) la robustezza, per ridurre al minimo i danni non intenzionali provocati da usi scorretti dei programmi di IA.

[148] Cfr. Commissione Europea, *Direzione generale delle Reti di comunicazione, dei contenuti e delle tecnologie, Orientamenti etici per un'IA affidabile*, Ufficio delle pubblicazioni, 2019, https://data.europa.eu/doi/10.2759/640340

[149] Cfr. *ivi*, p.10.

Il documento si articola secondo tre capitoli principali: nel primo vengono individuati i principi etici fondamentali che devono essere salvaguardati in ogni fase di utilizzo dell'IA (nello sviluppo, nella distribuzione e nell'utilizzo finale). Nel secondo capitolo vengono presentati sette requisiti che i sistemi di IA sono chiamati a soddisfare per preservare i principi precedenti. Nel terzo capitolo si rende disponibile una proposta concreta di lista di controllo per verificare l'affidabilità dell'IA nel rendere operativi i requisiti del capitolo 2.

Vorremmo qui riportare i punti principali di ciascun capitolo. Per una trattazione rigorosa e sistematica degli stessi si dovrà fare riferimento al documento completo disponibile *online*.

Nel primo capitolo vengono presentati i diritti fondamentali da tenere come base per lo sviluppo di un'IA affidabile, ricavati dai trattati internazionali in materia di diritti umani e dalla Carta dei diritti fondamentali dell'Unione europea. Essi sono:

1. Rispetto della dignità umana
2. Libertà individuale
3. Rispetto della democrazia, della giustizia e dello Stato di diritto
4. Uguaglianza, non discriminazione e solidarietà
5. Diritti dei cittadini

Notiamo solamente che nella spiegazione del punto (1), l'accento principale viene posto sul rispetto che i sistemi di IA devono avere verso gli esseri umani in quanto "soggetti morali" e non «come semplici oggetti da vagliare, catalogare, valutare per punteggio, aggregare, condizionare o manipolare»[150]. La sfida principale è ancora una volta quella di non ridurre l'essere umano a una "cifra" del calcolo che l'IA compie per giungere ai risultati.

Da questi diritti vengono ricavati i principi etici, che vengono anche chiamati "imperativi etici", cui ogni operatore del settore dovrebbe aderirvi nello sviluppo di sistemi di IA. Essi sono quattro:

1. Rispetto dell'autonomia umana
2. Prevenzione dei danni
3. Equità
4. Esplicabilità

Il punto (4) è particolarmente importante, poiché legato al discorso affrontato nei nostri capitoli precedenti riguardo l'inevitabile opacità dei sistemi di IA. In questo punto si fa riferimento infatti alla "scatola nera" che non consente una trasparenza del funzionamento degli algoritmi di IA. Il

[150] Cfr. *ivi*., p.12.

lavoro per rendere questo processo il più possibile trasparente agli esseri umani è di cruciale importanza, specialmente nei settori più delicati. Nella nota 33 del documento UE si prende come esempio l'uso di questi sistemi nell'ambito giuridico, per cui la decisione di consentire la libertà condizionale a un detenuto potrebbe dipendere da un ragionamento compiuto da un sistema di IA.

Nel secondo capitolo vengono presentati i sette requisiti principali per lo sviluppo di un'IA affidabile. È specificato che tale elenco non è esaustivo, ed è aperto a possibili integrazioni. I requisiti sono i seguenti, ciascuno composto da svariati altri sotto-punti (che qui non elencheremo nel dettaglio):

1. Intervento e sorveglianza umani
2. Robustezza tecnica e sicurezza
3. Riservatezza e *governance* dei dati
4. Trasparenza
5. Diversità, non discriminazione ed equità
6. Benessere sociale e ambientale
7. *Accountability*

Riguardo al punto (1), si specifica che i sistemi di IA dovrebbero essere pensati per sostenere l'autonomia e il processo decisionale umano. Sotto questa luce, gli agenti morali artificiali che abbiamo visto precedentemente dovrebbero porsi come "suggeritori etici" e non ricoprire il posto dei "decisori etici". Questo sarà vero in modo particolare in quei contesti, come la medicina o l'ambito giuridico, in cui sarà sempre richiesto l'intervento umano, anche solamente nel controllo dei processi decisionali della macchina.

Il punto (4) porta all'attenzione due questioni principali. Per prima la tracciabilità, per cui è necessaria una documentazione precisa dei dati e dei processi che vengono utilizzati dal sistema di IA per prendere una decisione, in modo da poter eventualmente correggere una decisione sbagliata. Senza tracciabilità è pressoché impossibile riuscire a prevenire errori futuri.

Come seconda questione c'è la spiegabilità, ossia la possibilità di ricostruire i processi decisionali messi in atto da un sistema di IA. Questo si lega a doppio filo con la tracciabilità: meglio tracciati saranno i dati e i processi, più facilmente si potrà avere una spiegazione affidabile del processo decisionale.

La (7) presenta il requisito della verificabilità[151], ossia della possibilità di valutare i dati, gli algoritmi e i processi di progettazione. Questo è importante perché implica che anche le aziende private dovranno emettere delle valutazioni dei propri sistemi di IA che siano accessibili a tutti, in modo da poter verificare la congruenza con tutti gli altri principi.

4.5 Stato attuale dell'etica dell'IA

Come abbiamo potuto vedere, il campo della *machine ethics* non ha prodotto risultati pratici di particolare impatto (siamo infatti ancora ben lontani da un agente morale artificiale di alto livello), ma lo studio metaetico ed etico attorno all'impatto delle nuove tecnologie di IA è condizione necessaria per arrivare preparati allo sviluppo sempre più veloce delle nuove tecnologie.

La sfida più grande è sicuramente quella di far incontrare i due mondi della filosofia morale e dell'ingegneria e di riuscire a farli comunicare tra loro. Il primo deve imparare a comprendere i metodi di funzionamento dei modelli di intelligenza artificiale, per intuire le nuove opportunità e le sfide computazionali che questo tipo di approccio connessionistico comporta; il secondo deve essere introdotto alla complessità del mondo etico: la scelta di una teoria etica al posto di un'altra non può essere fatta solo in base alla facilità di implementazione algoritmica, ma deve tenere in conto anche della qualità intrinseca della teoria e della rilevanza delle sue implicazioni. Dando come assodato che una macchina non è un essere umano (né potrà diventarlo), è auspicabile che divenga sistematico l'incontro tra filosofi ed ingegneri per la determinazione di una algoretica[152] efficace per i sistemi di IA.

[151] Cfr., *ivi*., p. 23.
[152] Uno tra gli esperti più importanti in Italia su questo tema è il prof. Paolo Benanti

5 La teoria dell'azione in Tommaso d'Aquino

In questo capitolo ci proponiamo di ricostruire la teoria dell'azione presentata da Tommaso d'Aquino nella sua *Summa Theologiae*, *Prima secundae*, con particolare riferimento alle questioni 6-18. Ci concentreremo sulla ricostruzione delle varie dimensioni dell'atto umano e delle sue componenti, che possano trovare un punto di incontro con gli argomenti esposti nei capitoli dedicati all'IA e alla modalità di funzionamento degli agenti artificiali (morali e non). In particolare ci concentreremo nel fornire i criteri di demarcazione tra umano e artificiale nell'ambito morale e a suggeriremo quali dimensioni dell'atto umano potrebbero essere implementate nelle macchine per lo sviluppo di agenti morali artificiali.

La riproposizione in tempi moderni di una teoria classica dell'agire umano si rende necessaria per vari motivi:

(1) l'attualità della proposta di Tommaso, che spesso non viene presa in considerazione dagli esperti del settore. Come abbiamo visto, le teorie etiche prese come riferimento per la modellizzazione di un'etica delle macchine presentano solamente un'unica proposta classica, cioè quella di Aristotele, consistente nell'etica delle virtù. Tuttavia, anche l'impianto dell'Aquinate ben si presta a essere un punto di riferimento per i modelli misti.

(2) La necessità di un'analisi non solo dal punto di vista dei risultati, ma soprattutto dei processi che guidano l'agire umano e l'agire artificiale. Se vogliamo risolvere il problema di confine tra "agenti morali espliciti" e "agenti morali in senso pieno", l'attenzione si deve spostare sulla differenza sostanziale dei processi implicati dagli uni e dagli altri.

(3) La riconsiderazione complessiva della totalità dell'uomo, per evitare ogni deriva di stampo riduzionistico nei prossimi sviluppi di IA. La possibilità di imitazione da parte delle macchine di alcune abilità tradizionalmente considerate appannaggio solamente dell'essere umano non deve portare alla conclusione di una replicabilità perfetta di tutte le capacità umane. Ancora una volta, se i risultati (o *performance*) di queste capacità possono essere simili, i processi sono radicalmente diversi.

Per la ricostruzione della teoria dell'azione faremo uso, oltre che del testo originale della *Summa*[153], anche di alcuni scritti molto completi sull'argomento, che possono essere utilmente consultati per avere un resoconto approfondito circa l'etica in Tommaso.

[153] Il testo a cui si farà riferimento è Tommaso d'Aquino, *La Somma teologica*, Edizioni Studio Domenicano, Bologna 2014.

Questi sono: A. Campodonico, M. S. Vaccarezza, *La pretesa del bene*[154], M. S. Vaccarezza, *Le ragioni del contingente*[155] e, uno tra i più specialistici e completi, W. Kluxen, *L'etica filosofica di Tommaso d'Aquino*[156].

5.1 Struttura generale della *Summa*

La *Summa Theologiae* nasce come testo rivolto ai "principianti", cioè a coloro che per la prima volta si avvicinavano allo studio della "Sacra Doctrina" (quella che in termini moderno chiameremmo "teologia"). L'opera completa è divisa in tre Parti, la terza delle quali non è conclusa. La seconda parte, invece, è suddivisa in due sottosezioni, una prima detta *Prima secundae* (indicata con I-II), e una seconda denominata *Secunda secundae* (indicata con II-II).

Nella *Prima pars* vengono trattate le questioni relative alla sacra dottrina, alla natura di Dio, alla creazione, agli angeli e all'uomo in generale. La trattazione della *Prima secundae* è invece centrata sulla morale, con particolare riferimento al fine ultimo dell'uomo, agli atti umani, alle passioni, la teoria degli "abiti" e alla legge morale. La *Secunda secundae* si concentra prevalentemente sulla spiegazione delle varie virtù (teologali e cardinal)i. La *Tertia pars*, inconclusa, avrebbe dovuto trattare della mediazione di Cristo.

Ogni *Pars* è composta da nuclei concettuali fondamentali, ossia le questioni, suddivise a loro volta in articoli, nei quali vengono trattati i principali nodi concettuali che compongono le singole questioni. Ad esempio nella I-II, q.1, sono presenti in totale otto articoli totali, che trattano del fine ultimo dell'uomo.

Per la *Prima secundae* suddivideremo il testo secondo questo schema[157]: la beatitudine (qq. 1-5), la struttura dell'azione umana (qq. 6-17), la bontà e malizia degli atti umani (qq. 18-21), le passioni (qq. 22-48), le virtù e i vizi (qq. 49-89), la legge e la grazia (qq. 90-114).

---

[154] Cfr. A. Campodonico, M.S. Vaccarezza, *La pretesa del bene, teoria dell'azione ed etica in Tommaso d'Aquino*, Orthotes Editrice, Napoli 2012.

[155] Cfr. M.S. Vaccarezza, *Le ragioni del contingente, la saggezza pratica tra Aristotele e Tommaso d'Aquino*, Orthotes Editrice, Napoli 2012.

[156] Cfr. W. Kluxen, C. Vigna (a cura di), *L'etica filosofica di Tommaso d'Aquino*, Vita e Pensiero, Milano 2005.

[157] Per questo ci rifacciamo alla suddivisione proposta in A. Campodonico, M. S. Vaccarezza, *La pretesa del bene*, p. 17.

5.1.1 *Organizzazione dell'articolo*

Come dicevamo, ogni questione è suddivisa in un numero variabile di articoli. Ogni articolo è organizzato secondo un modello standard, che segue questo schema:

1. Presentazione del tema specifico affrontato dall'articolo tramite l'espressione dubitativa (*utrum sit* o *an sit*).
2. Presentazione della tesi contraria a quella poi sostenuta da Tommaso, tramite l'espressione "sembra che" (*videtur quod*), utilizzando argomentazioni di altri filosofi e teologi. Ciascuna di queste tesi è numerata, poiché viene poi ripresa da Tommaso alla fine dell'articolo.
3. Presentazione della tesi di Tommaso nel "*sed contra*", attraverso la citazione di uno o più classici, che si presenta come posizione contraria rispetto a quella del *videtur quod*.
4. Risposta (*Respondeo*) di Tommaso alla domanda posta dall'articolo.
5. Soluzione delle difficoltà presentate nel *videtur quod*, in cui Tommaso affronta le posizioni precedenti (*ad primum ergo dicendum, ad secundum, ecc...*) e ne mette in luce gli errori, valorizzando per altro gli spunti rilevanti.

Una volta individuata la questione e l'articolo che ci interessa, per conoscere la posizione di Tommaso è sufficiente leggere il *sed contra* e passare al *respondeo*.

Per citare le questioni e gli articoli della *Summa* si utilizza un metodo standard: per prima cosa occorre indicare la Parte in cui si trova la questione (I, I-II, II-II, III), poi il numero della questione e il numero dell'articolo, indicando con *resp*. se ci si riferisce al *respondeo* dell'articolo. Si utilizza poi *s.c.* se si cita il *sed contra*, "ag" se si citano gli argomenti iniziali, "ad" se ci si riferisce alle soluzioni delle difficoltà.

5.2 La questione del fine ultimo

Per individuare quali siano le dimensione costitutive dell'essere umano, in riferimento alle quali considerare l'agire morale, riportiamo il prologo con cui Tommaso apre la *Prima Secundae*:

> Come insegna il Damasceno, si dice che l'uomo è stato creato a immagine di Dio, in quanto l'immagine sta a indicare «un essere dotato d'intelligenza, di libero arbitrio, e di dominio sui propri atti»; perciò, dopo di aver parlato dell'esemplare, cioè di Dio e di quanto è derivato dalla divina potenza conforme al divino volere, rimane da trattare della sua immagine, cioè dell'uomo, in quanto questi è principio delle proprie azioni, in forza del libero arbitrio e del dominio che ha su di esse.

Da queste prime righe è chiaro che per Tommaso le caratteristiche peculiari dell'uomo (intelligenza, volontà e libero arbitrio) trovano spiegazione nel fatto che l'uomo è la creatura per eccellenza. Esso è principio delle proprie azioni in virtù del libero arbitrio e della capacità (dominio) di orientarle secondo la sua volontà[158]: temi questi già affrontati nella *Prima Pars*.

Su questo punto, Tommaso distingue da subito due modi in cui è possibile intendere le azioni compiute dall'uomo[159]: esse possono essere considerate come *actus humani*, cioè come atti propriamente umani, che procedono cioè da una "volontà deliberata" e sono perciò azioni libere; o come *actus hominis*, cioè atti che all'uomo semplicemente capita di compiere, su cui egli non ha propriamente dominio. Pensiamo ad esempio al caso di uno spasmo improvviso o di un tic nervoso, di cui si ha certamente coscienza, ma su cui non si ha il controllo. Oppure anche il semplice grattarsi la barba mentre si sta pensando o il gesticolare mentre si parla.

La distinzione qui riportata non è da intendersi come una separazione netta: infatti ogni atto umano nasce da un sostrato in cui è compreso anche l'atto dell'uomo (e viceversa). Nell'esempio appena esposto, l'atto dell'uomo (il tic nervoso) può indurre la persona a decidere (atto umano) di iniziare un percorso di psicoterapia per affrontare il problema. Ma come sappiamo dopo la lezione di Freud, ogni azione umana scaturisce anche da un sostrato inconscio che come tale non è trasparente allo stesso agente. Con questo non si vuole affermare che ogni azione nasca dall'inconscio (che rappresenta infatti solo una dimensione della mente), ma che l'uomo è sempre attraversato da entrambe le dimensioni nella sua esperienza di vita.

La prima questione che Tommaso affronta è relativa al fine ultimo dell'uomo. In etica, infatti, il fine è ultimo nell'ordine dell'esecuzione dell'azione, mentre ricopre la funzione di principio nell'ordine dell'intenzione, cioè possiede la ragione di causa[160]. Così come la ragione tende per natura alla conoscenza dei primi principi, anche la volontà dovrà essere ordinata ad un "oggetto" a lei adeguato, che è identificato da Tommaso con il fine e il Bene[161].

Dimostrato che ogni azione umana è orientata ad un fine, nella q. 1, a. 4 si viene a dimostrare l'esistenza del fine ultimo e come l'accettazione di un processo di risalita *ad infinitum* nella catena dei fini si ponga in contraddizione con la ragione stessa di fine. Per fare ciò, Tommaso distingue

---

[158] Come troviamo in I-II, q. 17, a. 1, ad. 2, «radice della libertà come soggetto è la volontà, ma come causa è la ragione. Infatti la volontà può liberamente indirizzarsi a cose diverse perché la ragione è capace di concepire beni diversi. Quindi i filosofi definiscono il libero arbitrio «un libero giudizio dovuto alla ragione», come per indicare che la ragione è la causa della libertà»

[159] Cfr. I-IIae, q. 1, a. 1, *resp*.

[160] Cfr. I-IIae, q.1, a.1, ad. 1.

[161] Oggetto non è il termine più adeguato, poiché tale Bene nella metafisica classica non è un oggetto da conseguire, bensì una Persona da riconoscere e da incontrare

tra due tipi di ordine rispetto al fine: l'ordine dell'intenzione (*ordo intentionis*) e l'ordine dell'esecuzione (*ordo esecutionis*), e afferma che

> In tutti e due deve esistere un termine primo. Ciò che infatti è primo nell'ordine dell'intenzione costituisce come il principio motore degli appetiti, per cui, eliminato il principio, l'appetito rimane inerte. Il principio invece dell'ordine dell'esecuzione è il primo passo che uno compie nell'operare, per cui, eliminando questo, nessuno comincerebbe mai un'operazione[162].

Il principio nell'ordine dell'intenzione è il fine ultimo, mentre il principio nell'ordine dell'esecuzione è il primo dei "mezzi" ordinati al fine. Se si ammettesse la possibilità di procedere all'infinito, allora, dal punto di vista dell'intenzione, l'azione umana non avrebbe mai un termine ultimo e l'intenzione come tale rimarrebbe strutturalmente insoddisfatta. Se cioè non esistesse il Bene come fine ultimo, la volontà non si porterebbe su nessuno dei beni finiti, che risultano appetibili solamente in relazione con il Bene (in questo senso, tutto ciò che dall'uomo è desiderato è sempre ordinato rispetto al fine ultimo: cfr. I-II, q. 1, a. 6). Rispetto all'ordine dell'esecuzione, ammettendo di poter arretrare all'infinito, l'agente non inizierebbe mai a operare, poiché per lui sarebbe strutturalmente impossibile porre un primo tra gli atti adducenti al fine scelti in vista del conseguimento del fine. In entrambi i casi è dunque esclusa la possibilità di un processo *ad infinitum*.

Il rapporto con tale fine ultimo è diverso per come viene conseguito dall'uomo e come invece può essere conseguito dalle altre creature. I beni finiti come tali sono transeunti, e vengono conseguiti in vista di un loro utilizzo anche assimilativo. Assimilato il bene finito, la volontà passa alla ricerca di un altro che possa soddisfare l'appetito. L'uso consumativo dei beni è proprio sia degli animali che dell'uomo (pensiamo al cibo), mentre il fine ultimo come tale non può essere "consumato", ma può essere solamente "fruito" in modo continuativo[163]. Il fine ultimo infatti è in grado di saturare l'arco intenzionale del desiderio. L'atto di fruizione è possibile solamente all'essere umano, mentre per le altre creature ciò non è possibile, pur tendendo anche loro al fine ultimo.

Rifacendosi ad una distinzione già usata da Aristotele[164], Tommaso distingue tra un fine oggettivo e un fine formale[165]. Il primo – *finis cuius* – è l'oggetto che si vuole conseguire: in questo caso, esso è uguale per tutte le creature; infatti Dio è fine ultimo di tutti gli esseri. La distinzione

---

[162] Cfr. I-IIae, q. 1, a. 4, *resp*.

[163] La distinzione è tra l'uso (*usus*) e la fruizione (*fruitio*). Nel primo caso si ha una assimilazione dell'oggetto o un rapporto di tipo strumentale, nel secondo il rapporto non è teso al consumo, quanto ad una fruizione rispettosa del bene stesso.

[164] Cfr. *De Anima*, II, 415 b 3, 21.

[165] Cfr. I IIae, q. 1, a. 8

si ha nel secondo caso quello del *finis quo*, cioè la fruizione soggettiva del fine, per cui solo le creature intelligenti sono in grado di raggiungere Dio tramite la conoscenza e l'amore, mentre le altre creature ne partecipano a seconda del loro statuto creaturale. In questo senso, per l'uomo il raggiungimento del Bene come tale (cioè di Dio) non avverrà tramite un "fagocitamento", ma attraverso un rapporto vivo, indicato dalla figura della beatitudine.

5.2.1 Il fine ultimo e gli agenti morali artificiali

Questa discussione attorno al fine ultimo, oltre a fornirci le conoscenze di base per orientare un'analisi più specifica attorno all'agire umano, ci dà modo di riflettere sullo statuto ontologico dei sistemi di IA che abbiamo imparato a conoscere nei capitoli precedenti.

In I-IIae, q. 1, a. 2, Tommaso affronta il problema se l'agire per un fine sia una proprietà esclusiva della natura razionale. Egli afferma, nelle prime righe del *respondeo*, che «è necessario che tutti gli agenti agiscano per un fine». Nell'essere razionale ciò è dovuto all'appetito intellettivo, cioè alla volontà, mentre negli altri esseri all'inclinazione naturale, cioè all'appetito naturale.

Sappiamo che a nessun sistema di IA fino ad ora progettato può essere attribuita una coscienza, ma nemmeno un "appetito" nel modo inteso qui (a causa della natura intrinsecamente algoritmico-procedurale di questi sistemi, che rappresenta un limite di tipo ontologico). Tommaso fa notare che ci sono due modi in cui si può tendere verso il fine:

> Primo, muovendo sé stesso verso il fine, come fa l'uomo; secondo, facendosi muovere da altri verso il fine, come la freccia che tende a un fine determinato perché è mossa dall'arciere, il quale ne indirizza l'operazione verso il bersaglio. Quindi gli esseri dotati di ragione muovono sé stessi al raggiungimento del fine, […] gli esseri invece privi di ragione tendono al fine in forza di un'inclinazione naturale, come sospinti da altri […]. E ciò perché non conoscono la finalità delle cose, per cui non possono ordinare nulla verso il fine, ma vengono ordinati da altri al suo raggiungimento[166].

A una prima lettura si potrebbe obiettare che gli agenti morali artificiali sono in grado di fare proprio quello che Tommaso attribuisce all'uomo: agiscono in modo autonomo e sono in grado di determinare quali fini perseguire, e agire in accordo con essi. Tuttavia è bene notare che, nel caso degli agenti morali la loro autonomia è ristretta all'ambito dei processi e dei calcoli per soddisfare una certa misura di utilità, che però come tale non è da essi autonomamente stabilita. La posizione dei valori a cui la macchina si rifà nel suo processo "decisionale" è di competenza dell'uomo che la progetta. Un agente morale artificiale, in assenza di codificazioni valoriali,

[166] Cfr. I-IIae, q. 1, a. 2, *resp*.

resterebbe fermo e non sarebbe in grado di darsi valori in autonomia; analogamente, non sarebbe in grado di sostituire i valori già da esso implementati.

Mentre l'uomo è capace di scegliere dei valori da seguire, di confrontarli e di decidere quali non seguire, la macchina non è capace di tutto questo; è cioè totalmente eteronoma rispetto ai valori. Essa è perciò uno strumento (una freccia) nelle mani dell'uomo, che la programma secondo gli scopi che gli servono. È un'agente morale, ma solamente nel senso di un "elaboratore" morale, che non è in grado (né probabilmente lo sarà mai) di accedere alla "finalità delle cose".

5.3 Il ruolo della volontà

Le due principali potenze che caratterizzano l'essere umano sono l'intelletto e la volontà, da non intendersi come due potenze reciprocamente indipendenti, ma come due dimensioni dell'essere umano che, pur conservando ognuna la propria specificità, interagiscono insieme in modo inseparabile.

La volontà per Tommaso è un appetito razionale[167], orientato anzitutto al bene universale, in grado di portarsi in modo non necessitato sui beni contingenti e limitati, in virtù della libertà di scelta. Si tratta di una libertà d'arbitrio, poiché i beni finiti, pur essendo appetibili per la volontà, non presentano il carattere necessitante e totalizzante che spetta solo al fine ultimo, che è polo di una tensione inevitabile e inesauribile. Come tali, i beni limitati possono sempre essere relativizzati rispetto al fine ultimo.

In I-IIae, q. 10, a. 2– se la volontà è mossa in maniera necessitante dal proprio oggetto – Tommaso scrive nel *Respondeo*:

> la volontà può avere due mozioni: la prima in rapporto all'esercizio dell'atto, la seconda in rapporto alla specificazione dell'atto, la quale dipende dall'oggetto. Quanto alla prima, dunque, la volontà non può essere mossa da alcun oggetto in maniera necessitante: infatti uno può astenersi dalla considerazione di qualsiasi oggetto, e per conseguenza può anche non volerlo in modo attuale. Quanto invece al secondo genere di mozione la volontà è mossa da alcuni oggetti in maniera necessitante e da altri no. […] Se quindi alla volontà viene presentato un oggetto universalmente e sotto tutti gli aspetti buono, necessariamente la volontà tenderà verso di esso, quando desidera qualcosa: infatti non potrà volere l'opposto. Se invece le viene presentato un oggetto che non è bene sotto tutti gli aspetti, allora la volontà non sarà portata necessariamente a volerlo.

[167] La volontà è anche definita da Tommaso *adpetitus intellectivus vel rationalis*. Si veda I-IIae, q. 8, a. 1, *resp*.

Questo passo ci dice che in ogni processo di scelta che preceda e guidi la concretizzazione dell'azione, l'uomo è libero rispetto alla determinazione del bene finito più adatto, da considerarsi questo come un momento infravalente rispetto al fine ultimo, in quanto può relativizzare ogni bene finito rispetto al bene come tale. L'apertura della volontà – potremmo anche dire il suo arco intenzionale – è di tipo trascendentale, ed è ciò che consente la relativizzazione dei beni finiti, in quanto nessuno di questi può saturare in modo pieno questa apertura.

Il rapporto tra volontà e intelletto è chiarito nella I-IIae, q. 9, a. 1, in cui Tommaso si chiede se la volontà sia mossa dall'intelletto: la risposta è che la volontà muove come causa agente, cioè rispetto all'esercizio dell'atto, data la sua tensione verso il fine ultimo. L'intelletto, invece, muove come causa formale, cioè rispetto alla specificazione dell'atto, poiché il suo principio formale è l'essere nella sua verità. Si potrebbe obiettare[168] che qui si nasconde un circolo vizioso: la volontà muove l'intelletto all'esercizio dell'atto, ma è l'intelletto che offre alla volontà l'oggetto da volere; la risoluzione del dilemma sta proprio in quella tensione necessaria della volontà verso il fine ultimo, che come tale dà il via a tutto il dinamismo umano.

5.4 La scansione dell'agire umano in dodici atti

Riguardo alla trattazione delle qq. 6.-17 della I-IIae, nel corso del tempo è stata elaborata una suddivisione dell'atto umano in dodici elementi: sei dell'intelletto e altrettanti della volontà. Lo scopo di tale articolazione non è di fornire una scansione cronologica degli elementi in gioco, in quanto essi non sono sempre tutti necessari alla produzione di un atto deliberato (es. non si delibera sui mezzi quando ci si deve spostare camminando), ma di fornire piuttosto un quadro logico più chiaro. Noi ci rifaremo alla divisione presentata da M.S. Vaccarezza in *Le ragioni del contigente*[169].

Seguendo questo schema, per primi vengono gli atti che riguardano la determinazione del fine, ossia:

- L'apprensione del fine (I-IIae, q. 9, a. 1)
- La volizione del fine (I-IIae, q. 12, a.1, ad 3)
- Il giudizio dell'intelletto circa la conseguibilità del fine individuato precedentemente (I-IIae, q. 12, a.1, ad 1)

[168] Cfr. M.S. Vaccarezza, *Le ragioni del contingente*, pp. 27-30.
[169] Cfr., *ivi*, pp. 45-47.

- L'intenzione (*intentio*) del fine (I-IIae, q.12)

A questi seguono i successivi atti, non più concentrati sul fine ma sulla determinazione dei "mezzi" per raggiungerlo, che è più corretto chiamare "fini infravalenti", poiché non possono essere interscambiati come si potrebbe invece fare con degli strumenti.

Eccoli:

- Il *consilium*, cioè il consiglio o deliberazione, con cui l'intelletto pratico effettua una prima ricerca circa le azioni più appropriate in vista del fine scelto (I-IIae, q. 14)
- Il *consensus*, in cui la volontà vaglia le azioni possibili emerse dal *consilium* (I-IIae, q. 15)
- Il terzo stadio, del *iudicium praticum* (giudizio pratico), che si occupa di individuare in modo definitivo l'azione più efficace per raggiungere il fine (I-IIae, q. 13, a. 3)
- L'*electio*, la scelta vera e propria, con cui l'atto proposto viene approvato definitivamente dalla volontà (I-IIae, q. 13)
- L'*imperium*, cioè il comando effettivo per dare inizio all'azione (I-IIae, q. 17)
- L'*usus* che rappresenta l'esecuzione vera e propria, ed è il corrispettivo affettivo del punto precedente, che si specifica poi in *usus passivum* e *fruitio* (I-IIae, q. 11 e q. 16)

Come possiamo notare, presentato in questo modo schematico, sembra che la ragione e la volontà si alternino in modo rigido nella determinazione dei passaggi, per giungere all'esecuzione dell'azione. Bisogna tuttavia tenere a mente quanto detto prima sul rapporto tra le due facoltà, per evitare di cadere in una dicotomia rigida: più che di separazione, sarà allora più corretto parlare di una prevalenza[170] di una sull'altra nelle diverse parti del processo: prevalenza questa che conserva l'operato della facoltà precedente. Il processo di scelta umano non presenta quindi una rigida separazione tra le due facoltà, né una prevaricazione di una sull'altra, bensì una intima connessione tra le due, in cui la dimensione affettiva (della volontà) non può essere eliminata.

Ci pare questa un'osservazione da tenere a mente nell'implementazione di un sistema di ragionamento etico all'interno degli agenti morali artificiali: sarebbe necessario affiancare all'elemento di "ragione calcolante", che potremmo individuare nell'applicazione dell'approccio misto (quello dei doveri *prima facie* di Ross), un sistema di controllo "affettivo"[171], in grado cioè

[170] Crf. M. S. Vaccarezza, *Le ragioni del contingente*, pp. 47-48.

[171] Questo sistema di controllo "affettivo" sarà detto in questo modo solamente per via analogica, poiché si tratterà sempre di un approccio di tipo computazionale, che poco può condividere con la dimensione affettiva propriamente umana.

di valutare i possibili corsi di azione non solo secondo il parametro dell'efficienza e della rapidità, ma anche ad esempio quelli della dignità della persona o del valore sociale dell'azione.

A ben vedere si tratta di suddividere dei processi che oggi vengono svolti da un solo programma, in due programmi differenti: uno basato su un approccio dell'agire "razionale" più classico ed uno invece più sensibile ai valori ed ai diritti. Entrambi, a partire da una base di conoscenza comune, potrebbero svolgere il processo di scelta e individuare i passaggi fondamentali per giustificare un'azione al posto di un'altra. Avendo obiettivi diversi da soddisfare, alla fine presenteranno due modi diversi di giustificare la scelta della propria azione; la sfida sta nel riuscire ad integrarli insieme, per avere un processo decisionale sensibile ad entrambe le dimensioni.

Pur proponendo un processo decisionale per certi versi differente a quanto disponibile oggi, le questioni principali rimangono le stesse: è infatti necessario comprendere in che modo implementare i valori nel sistema "affettivo" e come delimitare il raggio d'azione e di affidabilità di un tale sistema. Come abbiamo avuto modo di comprendere, questi sistemi si basano su procedure algoritmiche che lavorano sui dati che hanno a disposizione, e, nel caso di un sistema "affettivo", non sono in grado di determinare dei principi da sé, ma sono in grado solamente di "agire" in accordo con dei principi che vengono posti dall'essere umano.

Solamente quest'ultimo, per via dell'apertura trascendentale del suo orizzonte, è in grado di de-situarsi e di lavorare e riflettere sugli stessi principi in modo critico. Ribadiamo ancora una volta questo punto, perché anche l'agente etico artificiale sarà portatore di "*bias* etici" che dipenderanno dai fattori socio-culturali in cui questi agenti verranno programmati e dai dati che andranno ad analizzare; già la scelta di utilizzare un approccio kantiano oppure utilitaristico (ispirato a Bentham), dà all'etica implementata nell'agente un taglio decisamente occidentale, che come tale fa riferimento ad un gruppo di valori ben consolidati. Se questo invece venisse progettato in Paesi con un'altra cultura, ad esempio in Giappone, potrebbe implementare degli approcci differenti da quelli tipici dell'occidente, veicolanti principi differenti.

Per fare l'esempio di un agente artificiale che includa sia la componente "ragione calcolante" che quella "affettiva", potremmo considerare le applicazioni di IA che si occupano di decidere se erogare o meno un credito ad un individuo che ne fa richiesta per acquistare la sua prima casa[172].

[172] L'esempio sopra riportato ha un carattere utopico, infatti le grandi banche hanno lo scopo primario di tutelarsi dal rischio e la "scommessa" di erogare un prestito senza le dovute garanzie non sarà mai attuabile (soprattutto con le somme in gioco per l'acquisto di una casa). Il consulente umano può però decidere di tenere in considerazione altri valori, non solo quelli meramente economici.

Nella parte della "dimensione calcolante" verranno presi in considerazioni dati quali la solvibilità, il patrimonio disponibile, la storia finanziaria precedente, la stabilità del lavoro, ecc... Nella parte "affettiva" potrebbero venire presi in considerazione valori non solamente economici, quali ad esempio il diritto ad una vita dignitosa o il diritto ad un alloggio, oppure dati quali l'età, la presenza o meno di figli, la loro età, il tessuto sociale in cui è inserita l'abitazione, ecc... È chiaro che un individuo giovane con un reddito medio-basso, con alle spalle una famiglia non troppo stabile economicamente, difficilmente avrà accesso ad un mutuo per acquistare una casa, poiché la "ragione calcolante" con grande probabilità boccerebbe tale richiesta. La parte "affettiva", invece, potrebbe mettere in luce come la banca non concede mai prestiti a persone giovani (pregiudicando in un certo senso l'autonomia dell'individuo e il suo diritto ad un'abitazione).

Passiamo ora ad analizzare alcuni dei passi che compongono il sistema in dodici atti, partendo dalla questione 12, avendo già analizzato precedentemente la questione 9 relativa all'apprensione del fine.

5.4.1 Questione 12: l'intenzione

Il termine "intenzione" (*intentio*) indica etimologicamente una "tensione verso qualcosa" ed è perciò un atto della volontà, cioè della facoltà colei che muove verso il fine. La volontà tende ad un fine (in questo caso un fine determinato) secondo l'ordine della ragione, in quanto tale fine deve prima essere conosciuto per essere proposto come polo attrattivo alla volontà.

Possiamo intendere in tre modi differenti il rapporto della volontà al fine:

> Primo, in modo assoluto: e allora viene denominata volere [*voluntas*], in quanto ad es. vogliamo in modo assoluto la guarigione, o altre cose del genere. Secondo, quando il fine è visto come l'oggetto in cui la volontà si riposa: e in questo caso il rapporto al fine è la fruizione [*fruitio*]. Terzo, quando il fine è visto come il termine delle cose ad esso ordinate: e allora il rapporto al fine è l'intenzione [*intentio*]. Infatti non diciamo di tendere alla guarigione solo perché la vogliamo, ma perché vogliamo raggiungerla con qualche mezzo[173].

L'oggetto dell'intenzione, quale atto della volontà, non è indirizzato solamente al fine ultimo, bensì anche ai fini intermedi che, per esemplificare, possono essere intesi come le tappe intermedie di un percorso che va da A a Z. Quest'ultimo rappresenta il fine ultimo, polo di attrazione e punto di soddisfacimento per la volontà; il passaggio attraverso i fini intermedi può essere visto come un passaggio da A a B, da B a C, e così via. Inoltre, lo stesso termine

[173] Cfr. I-IIae, q. 12, a. 1, ad. 4.

“intenzione” indica il moto verso un fine, mentre il quietarsi della volontà avviene solamente con la fruizione del fine[174].

Una precisazione importante di Tommaso riguarda il rapporto che intercorre tra il fine prescelto ed i “mezzi”: parola che traduce impropriamente l’espressione latina “*id quod est ad finem*” (ciò che è “per” o “in vista di” un fine). La traduzione abituale non riesce a restituire appieno il significato originale, poiché “mezzi” può essere inteso come qualcosa di interscambiabile; mentre, nella prospettiva dell’Aquinate, scegliere un “mezzo” al posto di un altro cambia la qualità dell’azione. Per cui ad esempio se mi sposto in città, mi è abbastanza indifferente se usare la bicicletta o l’automobile, in entrambi i casi arriverò alla destinazione che mi ero prefissato ed è corretto definire questi come “mezzi”. Invece, se ho come obiettivo una promozione sul posto di lavoro, ben diverso è arrivarci attraverso l’impegno personale piuttosto che denigrando i mei colleghi; nel secondo caso, l’azione è senza dubbio riprovevole, pur avendo lo stesso fine della prima (in questa prospettiva non può dunque valere la famosa espressione secondo cui “il fine giustifica i mezzi”).

Nella I-IIae, q. 12, all’a. 4 si chiarisce il rapporto tra l’intenzione del fine e la volizione dei mezzi, ossia se questo sia da intendersi come un identico atto o siano invece due atti separati. Nel *respondeo*, Tommaso afferma che

> il moto della volontà verso il fine e verso i mezzi ad esso ordinati può essere considerato sotto due aspetti. Primo, in quanto la volontà si porta distintamente e direttamente sull’uno e sull’altro. E allora, assolutamente parlando, abbiamo due moti della volontà. Secondo, può essere considerato in quanto la volontà si porta sui mezzi per raggiungere il fine. E in questo caso è numericamente unico il moto della volontà che tende verso il fine e verso i mezzi.

Nello stesso articolo, in ad. 3 precisa

> Se quindi il moto della volontà ha per oggetto i mezzi in quanto ordinati al fine, è denominato scelta [*electio*]; se invece ha per oggetto il fine da raggiungere mediante i mezzi, è detto intenzione [*intentio*]. E ne è prova il fatto che si può avere l’intenzione del fine prima di determinare i mezzi, che sono l’oggetto della scelta.

Dalla lettura di questi testi sembra emergere l’individuazione di due momenti distinti (non solo concettualmente, ma anche temporalmente), in cui la volontà si può riferire o ai “mezzi” (*electio*) o al fine (*intentio*), legando in un certo modo la stretta connessione tra il fine e ciò che è “in vista del” fine.

[174] Cfr. I-IIae, q. 12, a. 2, ad. 3.

Tuttavia i due momenti (cioè il moto della volontà verso il fine e le realtà che ad esso adducono) sono distinguibili solamente dal punto di vista formale, mentre nel dinamismo dell'azione essi si presentano come un unico atto[175]. Siamo qui d'accordo con l'analisi di Jannis, per cui «non c'è alcuna intenzione finché non c'è l'adozione dei mezzi, il che accade con la scelta. E viceversa: quando c'è una scelta, c'è anche un'intenzione»[176]. L'intenzione del fine senza alcun rapporto con i mezzi sarebbe più da intendere come *voluntas*, cioè come un volere il fine in modo assoluto come già visto nella citazione di inizio paragrafo. L'intenzione non è quindi da intendersi come un momento astratto rivolto al fine; affinché si possa dire che un soggetto ha "un'intenzione" è necessario che questi abbia già individuato una sequenza di azioni possibili per raggiungere il fine.

5.4.1.1 Indicazioni per gli agenti morali artificiali

Tommaso specifica chiaramente che negli animali irrazionali non è presente propriamente l'intenzione al fine, poiché in assenza della ragione non è possibile ordinare alcuna cosa in vista del fine. Tuttavia ammette che è possibile che anche la natura (e quindi anche gli animali) possa avere un'intenzione del fine, qualora questa intenzione provenga dall'esterno, come una freccia messa in moto dall'arciere.

È chiaro che i sistemi di IA propriamente non hanno intenzione del fine nel senso qui proposto: questi sistemi piuttosto "agiscono" cercando di soddisfare delle richieste che vengono poste dall'esterno dai programmatori. Se prendiamo come esempio *ChatGPT*, esso non si pone il fine di "apprendere la lingua inglese" (o qualsiasi altro idioma), ma solamente di minimizzare una funzione di costo che gli consenta di raggiungere una certa precisione, espressa in un rapporto quantitativo corretto/sbagliato (il fine di "apprendere la lingua inglese" è piuttosto il risultato a cui puntano i programmatori). Allo stesso modo dovranno essere intesi gli agenti morali artificiali, i quali nello svolgimento in autonomia dei loro compiti avranno come obiettivo solamente quello di minimizzare la distanza rispetto a un valore etico (inserito a monte dai programmatori).

---

[175] Come afferma anche Tommaso nella I-IIae, q. 12, a. 4 c: «I mezzi stanno al fine come un punto intermedio sta al termine corrispettivo. Ora, nel mondo fisico è unico il moto che raggiunge il suo termine attraverso un punto intermedio. Quindi anche nel campo degli atti volontari è un moto unico l'intenzione del fine e la volizione dei mezzi.»

[176] Cfr. J. Finnis, (traduzione di) A. Cevolini, *Contenuto e intenzione nei giudizi morali secondo Tommaso d'Aquino*, "Divus Thomas", vol. 104, n. 2, 2001, pp. 21-23. Reperibile su https://www.jstor.org/stable/45074911

Venendo meno nell'IA una concezione forte di "intenzione", è necessario concepire diversamente anche la responsabilità degli agenti artificiali: potremmo parlare solo in maniera analogica di "intenzione" per questi agenti, poiché essi non fanno altro che elaborare statisticamente delle informazioni che posseggono (o acquisiscono) per soddisfare dei requisiti introdotti dall'esterno (valori, obiettivi, ecc…). La responsabilità in questi agenti artificiali dovrà essere quindi intesa in modo "distribuito", per cui si dovranno tenere in conto attori quali i programmatori, i distributori dei *software*, le aziende che li implementano, ecc… Non sarà possibile "scaricare" le responsabilità sul *software*, poiché esso propriamente è un agente morale inconscio.

5.4.2 *Questione 14: la deliberazione o consilium*

Lasciamo ora le considerazioni attorno al fine e passiamo a quelle circa i "mezzi" o, più correttamente, ai fini infravalenti che consentono il raggiungimento pratico del fine. Qui ci proponiamo di affrontare la questione 14, saltando quindi momentaneamente la questione 13 che tratta della scelta [*electio*], nonostante nella presentazione dell'Aquinate questa preceda quella[177]. Prendiamo questa decisione per favorire una maggiore organicità dell'esposizione (supportati in questo dai contributi della critica), ma rimanendo sempre fedeli all'impostazione di Tommaso.

Citiamo a tal proposito il primo articolo della questione 14, che definisce precisamente che cosa sia il consiglio e quale sia il suo rapporto con le altre dimensioni dell'azione

> la scelta, come si è visto, segue il giudizio della ragione sulle azioni da compiere. Ora, nelle azioni da compiere si riscontra molta incertezza: poiché le azioni riguardano cose singolari contingenti, che per la loro variabilità sono incerte. D'altra parte nelle cose dubbie e incerte la ragione non proferisce il suo giudizio senza una previa ricerca. Quindi è necessaria una ricerca della ragione prima del giudizio sulle azioni da compiere; e questa ricerca viene chiamata consiglio, o deliberazione. Per cui il Filosofo scrive che la scelta è "il desiderio di cose predeliberate mediante il consiglio"[178].

[177] Per spiegare come mai Tommaso operi questa scelta, ci rifacciamo al commento presente in Tommaso d'Aquino, (a cura di) domenicani italiani, *La somma teologica*, PDUL Edizioni Studio Domenicano, Bologna, 1985. Qui, nella I-IIae, alla nota 1 della questione 13, si sostiene che Tommaso inverte la trattazione per due motivi: (1) per una ragione di ordine speculativo, poiché segue un ordine sistematico e non genetico nella trattazione, dunque prima espone l'essenza della cosa (elezione) e poi le cause estrinseche; (2) per una ragione di ordine storico, in quanto Aristotele nell'*Etica* aveva premesso al consiglio l'elezione, e tale opera è la fonte principale di Tommaso. Aggiungiamo che la trattazione scelta dall'Aquinate può anche evitare di intendere la scansione degli atti in modo troppo rigido e secondo una legge di causa-effetto.

[178] Cfr. I-IIae, q. 14, a.1, *resp*.

Dopo l'intenzione del fine da parte della volontà, l'attività di entrambe le facoltà si concentrano verso ciò che porta al fine. Il primo "passaggio" spetta alla ragione che, attraverso la figura del consiglio o deliberazione, opera una ricerca (*inquisitio*) su quali siano le azioni possibili da compiere per raggiungere il fine. Il consiglio viene definito come un «appetito investigativo»[179], poiché la volontà offre la materia e l'incentivo per la ricerca che è compiuta dalla ragione. Inoltre, il consiglio ha per oggetto solamente i mezzi, poiché il fine ha carattere di principio e non può essere quindi oggetto di ricerca.

L'*inquisitio* tramite il consiglio è necessaria perché, come tutti abbiamo avuto modo di sperimentare, ci possono essere più percorsi d'azione che ci conducono al conseguimento del medesimo fine. Tommaso precisa però che per alcuni generi di azioni[180] non è richiesto il consiglio, perché non è necessario investigare quali "mezzi" ci conducano al fine; pensiamo ad esempio quando ci alziamo e ci muoviamo per le stanze della casa: non è necessario riflettere su quali "mezzi" siamo più adatti allo scopo, semplicemente ci mettiamo in moto.

Il termine *consilium*, anche nella sua traduzione italiana "consiglio", implica l'idea di un confronto tra diverse persone[181], cioè di un "consesso" dove ci si ritrova per discutere attorno a delle questioni. In questo senso, il consiglio è un confronto sulle possibili azioni da compiere, non solamente sulle azioni in sé ma anche sui possibili risvolti futuri: l'attenzione è rivolta anche alle conseguenze di un possibile corso d'azione e non è ristretta al solo raggiungimento del fine.

Possono esserci due modi di procedere a partire da un principio[182]: secondo un ordine compositivo, partendo dalle cause per arrivare agli effetti, se questo principio ha priorità sia nell'ordine della conoscenza che in quello reale; oppure secondo un ordine risolutivo, partendo

---

[179] Cfr. *ibidem*.

[180] Cfr. I-IIae, q. 14, a. 4, *resp*., «Quindi, al dire del Filosofo, noi non deliberiamo nel nostro consiglio su queste due cose: sui fatti insignificanti e su quelli già determinati nel loro modo di esecuzione, cioè sui vari esercizi delle arti; "eccettuate le arti congetturali", precisa il Nisseno, "come la medicina, la mercatura e simili"». Un operaio impiegato in catena di montaggio, una volta appreso il lavoro, non dovrà deliberare continuamente su quali movimenti fare e come, poiché li compirà in modo quasi meccanico. Per un chirurgo, invece, sarà necessario soppesare ogni movimento, per evitare conseguenze anche gravi. Il rischio della ripetizione meccanica di qualsiasi azione è un'atrofizzazione del processo di scelta, per cui si agisce senza riflettere.

[181] Cfr. I-IIae, q. 14, a. 2, *resp*.

[182] Cfr. I-IIae, q. 14, a. 5, *resp.*, si veda anche *In De Trin.*, lect. II, q. 2, a. 1, ad 3, Tommaso scrive: «Il processo raziocinativo può assumere due orientamenti: compositivo, quando procede dalle forme più universali a quelle meno universali (particolari): risolutivo, quando procede in senso inverso. Infatti, ciò che è più universale è più semplice. Ora è universalissimo ciò che appartiene a ogni ente. E perciò il termine ultimo in questa vita è lo studio dell'ente e di tutto ciò che gli appartiene in quanto ente. E queste sono le cose di cui si occupa la scienza divina (o metafisica), ossia le sostanze separate e tutto quanto è comune a tutti gli enti. Dal che risulta che l'indagine metafisica è sommamente speculativa»

in modo inverso dagli effetti per risalire alle cause, se il principio ha priorità nell'ordine della conoscenza ma è posteriore nell'ordine reale.

È chiaro che con l'agire morale ci troviamo nell'ordine risolutivo, poiché il fine dell'azione è primo nell'intenzione, ma ultimo rispetto al suo conseguimento reale. Per certi versi questo procedimento può risultare controintuitivo, abituati come siamo a prestare attenzione solamente alla "fatticità" dell'azione, più che al suo ordine dal punto di vista del ragionamento. Nella I-IIae, q. 14, a. 5, alla risoluzione della prima difficoltà, Tommaso afferma che «Il consiglio ha certamente per oggetto le operazioni, ma il motivo dell'operare è desunto dal fine: perciò l'ordine del ragionare sulle operazioni è l'inverso di quello dell'operare».

Al termine del consiglio, quindi, la ragione avrà determinato due o più possibili percorsi d'azione che l'individuo potrebbe intraprendere per raggiungere il fine prefissato. Sara compito del *consensus*, tramite la volontà, considerare anche affettivamente ciascuna delle opzioni, per poi passare alla scelta vera e propria (*electio*) di una sola tra queste.

5.4.2.1 Indicazioni per un'etica dell'IA

Questa dimensione è forse quella che si presta con meno difficoltà ad una traduzione in termini computazionali, e di fatto è quello che viene eseguito nel momento in cui la macchina deve ricercare i possibili percorsi d'azione da intraprendere. È doveroso sottolineare ancora una volta che il consiglio, ad opera della ragione, non è da ridurre a un mero calcolo come è inevitabile che sia nella prospettiva consequenzialistica.

Prendendo spunto dal termine *cosilium* inteso come "consesso", un sistema di IA potrebbe calcolare i diversi percorsi d'azione seguendo diversi modelli etici: avrebbe allora un risultato applicando la teoria kantiana, uno differente applicando la teoria utilitaristica classica, un altro ancora con la teoria di Ross dei doveri *prima facie*, e così via. All'interno di ogni teoria si presenterebbero poi altre diverse opzioni da calcolare (es. quale dovere *prima facie* trasgredire o quali conseguenze tenere in conto e quali altre no). Il risultato sarebbe un diagramma ad albero, con azioni differenti sia per ogni tipologia di modello etico, sia all'interno di ogni singolo modello. In modo schematico, in cima a questo diagramma ci sarebbe il "problema" che il *software* si trova a dover affrontare (ad esempio, per una macchina a guida autonoma come evitare un ostacolo o per un programma di profilazione quale contenuto specifico suggerire); invece di trattare questo problema secondo un singolo modello etico, il sistema potrebbe utilizzare diversi modelli etici (primo ramo del diagramma) e, per ognuno di questi, stimare l'azione migliore da compiere secondo diversi parametri (secondo ramo del diagramma). Così,

ad esempio, all'interno della prospettiva conquenzialistica ci saranno diversi tipi di azioni possibili, a seconda di quali conseguenze siano state tenute in conto e con quali valori e probabilità.

5.4.3 Questione 15: il consenso

Nel passaggio precedente, tramite il consiglio la ragione aveva individuato i possibili corsi d'azione attuabili in vista del fine. Il consenso, atto della volontà, vaglia questi "mezzi" secondo una prospettiva affettiva, valutandoli quindi non solo dal punto di vista "prestazionale" ma anche rispetto al loro rapporto al bene (ribadiamo che questi momenti, nell'agire concreto, si trovano in un'unità: nel momento in cui valutiamo quali mezzi usare per arrivare ad un fine, implicitamente li valutiamo anche da un punto di vista "affettivo"). Questo passaggio è chiarito nel *sed contra* dell'articolo 1, in cui «il Damasceno insegna che "se uno giudica senza amare, non vi è sentenza", cioè consenso. Ma l'amare spetta a una potenza appetitiva. Quindi anche l'acconsentire»[183].

Acconsentire[184], spiega Tommaso, equivale ad un "sentire insieme", in cui la distanza tra il senziente e il sentito, comportando una certa unione tra i due poli; ora, è proprio della volontà questo moto verso le cose. Per l'intelletto, invece, si tratta di un "assentire" (*sentire ad*) a qualcosa d'altro che rimane altro.

Nell'articolo 3, Tommaso ci offre una ricostruzione del rapporto tra le varie parti dell'azione, per cui «nell'ordine dell'agire [umano] prima di tutto va posta l'apprensione del fine; quindi il desiderio di esso; viene poi il consiglio relativo ai mezzi, e finalmente l'appetizione dei mezzi»[185]. Ora, il fine che noi ci proponiamo di raggiungere non è in quanto tale il fine ultimo, e quindi si può giustamente definire come fine intermedio. Come tale, anch'esso può ricadere sotto la dinamica del consiglio e del consenso, in quanto è possibile vederlo come un "mezzo" per un fine più alto.

Dopo che la volontà si è portata sui singoli percorsi d'azione individuati dal consiglio e li ha valutati da un punto di vista "affettivo", non per questo ha deciso in modo definitivo quale tra questi percorsi sia il più adatto. Se fosse così, allora non si spiegherebbe il senso della scelta [*electio*] che segue al consenso. Tommaso lo precisa affermando che

---

[183] Cfr. I-IIae, q. 15, a. 1, *s.c.*
[184] Cfr. I-IIae, q. 15, a. 1, ad. 3.
[185] Cfr. I-IIae, q. 15, a. 3, *resp*.

> La scelta aggiunge al consenso un riferimento alle cose scartate: perciò dopo il consenso c'è ancora la scelta. Infatti può capitare che nel consiglio, o deliberazione, si trovino più mezzi adatti al raggiungimento del fine [*per consilium inveniantur plura ducentia ad finem*], ciascuno dei quali è gradito e quindi oggetto di consenso; ma tra le molte cose che piacciono ne preferiamo una mediante la scelta. Se invece fosse gradito un mezzo soltanto, allora tra il consenso e la scelta non ci sarebbe differenza reale, ma solo di ragione: e verrebbe detto consenso in quanto si tratta di un fatto che ci piace di compiere, scelta invece in quanto è una cosa preferita ad altre non gradite[186].

Qualora vi fosse solamente un "mezzo" per raggiungere il fine, allora non sarebbe necessario il consenso, ma si passerebbe direttamente alla scelta, che ratificherebbe quanto emerso dal consiglio. Ma nel momento in cui c'è interesse verso molteplici "mezzi", attraverso la scelta ne individuiamo uno (o più): quello che sarà poi utilizzato praticamente per il raggiungimento del fine, attraverso il giudizio o decisione della ragione.

5.4.3.1 Indicazioni per un'etica dell'IA

Riprendendo la proposta del punto 5.4.2.1, una volta individuate le ramificazioni dei percorsi d'azione secondo le teorie e le possibili azioni da intraprendere all'interno delle teorie stesse, il sistema di IA potrebbe passare a fare un confronto incrociato tra le teorie, per valutare il loro grado di "sensibilità" rispetto a certi valori etici. Posto ad esempio come valore principale il "rispetto dell'autonomia", ogni corso d'azione per ogni teoria etica verrebbe sottoposto ad una verifica rispetto a questo valore, per stimarne la compatibilità o meno. Alla fine avremo una lista di valori che ci descrivono quanto una certa azione calcolata rispetto ad una teoria "rispetta" il valore che ci siamo prefissati di osservare.

5.4.4 Questione 13: l'elezione o scelta

La definizione di "scelta" proposta da Tommaso mette ancora una volta in primo piano l'inestricabilità delle facoltà della ragione e della volontà nel processo decisionale. L'*electio* viene infatti definita come «intellezione appetitiva o un'appetizione intellettiva»[187], cioè come un sinolo delle due facoltà, in cui una (la ragione) è formale rispetto all'altro. La scelta non è solamente l'appetito (consenso), ma nemmeno solamente il consiglio, ma una loro combinazione. Ciò vuol dire che «l'atto mediante cui la volontà tende a una cosa presentata come buona, essendo ordinato al fine dalla ragione, appartiene materialmente alla volontà e

[186] Cfr. I-IIae, q. 15. a. 3, ad. 3.
[187] Cfr. I-IIae, q. 15, a. 1, *resp.*

formalmente alla ragione»[188]. Nonostante ciò, la scelta resta prevalentemente un atto della volontà, poiché è essa che si porta verso il bene prescelto.

Tommaso precisa subito che l'atto della scelta non è presupposto necessariamente ad ogni atto volontario: perché vi sia scelta, bisogna che essa si riferisca a più cose passibili di scelta, tra le quali una verrà preferita alle altre. Gli esseri privi di ragione[189] (tra cui gli animali), nei quali è presente l'appetito sensitivo ma non la facoltà della volontà, sono determinati per natura a oggetti particolari; mentre la volontà, che è rivolta al bene universale, rimane indeterminata rispetto ai beni particolari. Un individuo in grave condizione di obesità potrà decidere di mangiare meglio per dimagrire, preferendo una dieta bilanciata ad una ricca di grassi; è cioè in grado di preferire al bene particolare (cibo) un bene più elevato (la salute), che in ultimo si rifà al fine di preservare la propria vita. Un animale in grave obesità, invece, alla presenza di pietanze invitanti non si asterrà dal mangiarle, non essendo in grado di relativizzare queste in vista di un bene più grande. A fianco degli esseri privi di ragione, la lezione classica inserisce anche i bambini, i quali agiscono volontariamente, ma senza scelta; pensiamo ad esempio ad un bambino di 1 anno che per la prima volta mangia del cibo sui piatti e non imboccato: esso certo si muoverà volontariamente verso il cibo, ma in prima battuta non sceglierà quale preferire, mangiando da tutti i piatti e passando indifferentemente dall'uno all'altro.

Nell'articolo 4, "se la scelta abbia per oggetto soltanto le nostre azioni", Tommaso pone in rilievo la differenza, che abbiamo già approfondito, tra *actus humani* e *actus hominis*: nel *sed contra*, infatti, l'espressione "le nostre azioni" si tramuta in "le azioni umane". Affinché un atto sia propriamente umano (cioè nostro), è necessario che i "mezzi" che portano alla specificazione dell'atto siano frutto di una scelta consapevole (cioè passino attraverso l'*electio*). Negli atti che capita all'uomo di compiere, invece, non c'è propriamente la scelta dei mezzi.

Si afferma quindi che

> la scelta ha per oggetto i mezzi, come l'intenzione ha per oggetto il fine. Ora, il fine o è un'azione, o è una certa realtà. Ma anche in quest'ultimo caso è necessario che intervenga un'azione umana: o perché l'uomo produce la realtà che costituisce il fine, come il medico produce la guarigione (infatti si dice che il fine del medico è produrre la guarigione), oppure perché l'uomo in qualche maniera fa uso o gode della realtà che costituisce il fine: come il fine dell'avaro è il danaro, o il possesso del danaro. E lo stesso argomento vale per i mezzi ordinati al fine. Infatti è necessario che quanto è ordinato a un fine o sia un'azione, oppure una realtà con l'intervento di un'azione, la quale o dovrà produrre il mezzo ordinato al fine, o servirsi di esso. E in questo modo la scelta ha sempre per oggetto delle azioni umane.

[188] Cfr. I-IIae, q. 15, a. 1, *resp.*
[189] Cfr. I-IIae, q. 13, a. 2, *resp*.

Abbiamo riportato per esteso il *respondeo* poiché in questo luogo si nota in modo inequivocabile la differenza tra "ciò che porta al fine" e ciò che si intende normalmente come "mezzo". In ogni agire pratico che interessi l'uomo, "ciò che porta al fine" non è mai (o non si identifica mai con) un mezzo interscambiabile indifferentemente con altri, ma è piuttosto la scelta di una precisa azione (o sequenza di azioni) che, se scambiata con un'altra simile, non porterebbe ad un risultato analogo. Ciò che cade sotto l'ambito della scelta non è mai propriamente uno strumento, ma semmai il modo in cui lo strumento viene utilizzato lo strumento in vista del fine, oggetto della scelta è dunque una certa l'azione.

Se ad esempio ci troviamo a dover conficcare un chiodo in una trave di legno usando un martello, la scelta propriamente intesa non sarà se fare uso o meno del martello, quanto piuttosto su quale braccio utilizzare per maneggiarlo. Se siamo abituati ad usare la mano destra, certo non sceglieremo di utilizzare il martello con la sinistra. Il martello come "mezzo" è intercambiabile (potremmo decidere di usare al suo posto una pietra presente nei paraggi); l'azione in quanto tale non è intercambiabile: usare la sinistra quando invece siamo abituati alla destra porterà probabilmente ad un risultato non soddisfacente; in ogni caso si può raggiungere il fine (piantare il chiodo), attraverso un "mezzo" (martello o pietra), ma tramite delle azioni che non sono indifferentemente intercambiabili (avremo un lavoro ben fatto oppure uno scadente). È chiaro dunque che i "mezzi" a cui si riferisce Tommaso sono sempre e solo le azioni umane, e mai qualcosa di solamente strumentale.

L'elezione, riferendosi alle azioni, si limita alla scelta di sole azioni possibili. Come spiega Tommaso nell'articolo 5, noi scegliamo azioni in grado di farci raggiungere il fine; se tra queste ve ne fossero anche di impossibili, allora ci sarebbe negato il raggiungimento del fine e non ci muoveremmo nemmeno. Se non si dà scelta di cose impossibili, relativamente ai "mezzi", si può invece dare la volontà di oggetti impossibili: quello che Tommaso denota come volontà incompleta o *velleitas*.

Per ultimo c'è da chiedersi se la scelta dell'uomo relativa ai mezzi sia necessaria: molte volte infatti Tommaso paragona l'agire dell'uomo alla forma del sillogismo pratico; in questo caso, il fine sta ai mezzi come i principi stanno alle conclusioni[190]. Una volta posti i principi, cioè il fine, sembra che necessariamente ne discendano le conclusioni, cioè i mezzi. Il movimento verso i mezzi sarebbe dunque necessitato dalla scelta del fine, e l'uomo in tal caso non sarebbe libero.

[190] Cfr. I-IIae, q. 13, a. 5, *co*. 1.

L'argomentazione di Tommaso è chiarissima rispetto a questo punto, e merita di essere riportata

> la scelta dell'uomo non è necessaria. E questo perché non è mai necessario ciò che può non essere. Ora, si può dimostrare che è cosa indifferente scegliere o non scegliere partendo dalla duplice facoltà che l'uomo possiede. Infatti l'uomo ha la facoltà di volere o non volere, di agire o non agire; e ha anche la facoltà di volere e di compiere questa o quell'altra cosa.

Abbiamo visto infatti come la volontà non sia necessariamente attratta dagli oggetti che considera[191], sia rispetto ai fini intermedi che rispetto ai "mezzi". Data l'intrinseca finitezza di entrambi, la volontà può scegliere di sospendere il suo atto, senza perciò cadere nella contraddizione: verso i beni finiti è possibile infatti "volere di non volere". L'unica volizione radicale, e quindi necessaria, è quella verso il fine ultimo, che come tale non può essere annullata né ignorata. Il *respondeo* continua affermando che

> ne abbiamo la riprova nella stessa struttura della ragione umana. Infatti la volontà può tendere verso quelle cose che la ragione apprende sotto l'aspetto di bene. Ora, la ragione può apprendere come bene non solo il volere e l'agire, ma anche il non volere e il non agire. Inoltre in tutti i beni particolari la ragione può osservare l'aspetto buono di una cosa, oppure le sue deficienze di bene, che si presentano come un male; e in base a ciò può apprendere ciascuno di tali beni come degno di scelta o di rifiuto. Ora soltanto il bene perfetto, cioè la beatitudine, non può essere appreso dalla ragione sotto l'aspetto di male, o come difettoso. Ed è per questo che l'uomo vuole per necessità la beatitudine, e non può volere l'infelicità o la miseria.

Si ripropone qui il tema dell'inestricabilità tra ragione e volontà, per cui se la volontà si porta verso i beni finiti che apprende sotto l'aspetto di bene, è proprio però della ragione proporre alla volontà tali beni. Ora, dato che alla ragione pertiene l'ultima decisione intorno all'agire pratico (l'*imperium*), essa può comandare anche il non-agire, poiché ciascuna cosa può essere considerata sia sotto l'aspetto di bene, che sotto la sua deficienza rispetto ad esso. È da sottolineare che solamente grazie all'apertura trascendentale di entrambe le facoltà (ragione e volontà) è possibile che la ragione non operi meccanicamente nel comandare un atto e con ciò non fare violenza alla volontà. Se quest'ultima fosse necessitata anche dai beni particolari, il rapporto della ragione con la volontà sarebbe assimilabile a uno di tipo "tirannico", che obbliga e sopprime le inclinazioni dei suoi sottoposti. Nella prospettiva di Tommaso, la ragione è in grado di ricomprendere le motivazioni proprie della volontà (ponendosi in "ascolto") e di riorientarle in modo efficace nell'ultima decisione attorno all'agire pratico.

[191] Cfr. I-IIae, q. 10, a. 2, *resp*.

> Ma la scelta non ha per oggetto il fine, bensì i mezzi, come si è già spiegato; e non riguarda il bene perfetto, cioè la beatitudine, ma gli altri beni particolari. Quindi l'uomo non sceglie per necessità, ma liberamente[192].

In queste ultime righe Tommaso conduce a conclusione quanto affermato sopra: vista l'indipendenza delle due facoltà verso i beni contingenti e particolari, e dato che la scelta non ha per oggetto il fine ma i "mezzi", cioè i beni particolari, l'uomo è in grado di scegliere in modo libero.

Successivamente all'*electio* ci sono i momenti dell'*imperium*, in cui l'intelletto pratico ordina lo svolgimento dell'azione prescelta, che da parte della volontà si esplica nell'*usus*, cioè l'atto vero e proprio. Ci fermiamo qui nella spiegazione particolareggiata degli atti perché per il nostro discorso attorno all'IA i precedenti sono quelli più importanti e centrali.

5.4.4.1 Indicazioni per un'etica dell'IA

Abbiamo visto come la determinazione dei fini, anche nel campo dell'IA pertenga sempre all'uomo in quanto tale: i valori o i risultati che ci si aspetta il *software* riesca a soddisfare durante il funzionamento sono infatti inseriti dai programmatori. Il *software* di IA può solo analizzare dei dati per apprendere in che modo soddisfare un obiettivo, secondo un procedimento che non è determinato a priori nel codice.

Anche se una prima parte di addestramento sui dati è effettuata in modo non-supervisionato, il controllo successivo è sempre effettuato da un esser umano, per fornire al modello un *feedback* circa i risultati che quest'ultimo ottiene e per verificare che la fase di apprendimento abbia funzionato correttamente.

Il tema dell'elezione o scelta è uno dei punti più critici da trattare quando bisogna fare un parallelo tra agenti morali umani e agenti morali artificiali. Se in qualche modo è possibile dirigere l'operato della macchina verso il soddisfacimento di alcune procedure normative inserite in essa dall'uomo, come abbiamo visto è invece molto più difficile ricostruire il processo secondo cui i *software* di IA determinano in che modo agire: in breve, l'opacità intrinseca degli algoritmi di *deep learning* è relativa sia all'uomo che allo stesso *software* di IA. Come nota giustamente Gianfranco Basti[193], anche per l'essere umano il processo di formazione della scelta è in sé opaco: infatti, quando prendiamo una decisione, tutti i termini che abbiamo qui sopra esposto non ci

[192] Cfr. I-IIae, q. 10, a. 2, *resp*.

[193] Cfr. G. Basti, *La sfida etica dell'intelligenza artificiale e il ruolo della filosofia*, "Aquinas", vol. 65, 2022, pp. 313-318.

sono partitamente chiari al momento, né passiamo consciamente dall'uno all'altro, secondo una logica ben ordinata. Pensiamo al riguardo anche al ruolo delle emozioni o alla lezione freudiana sull'influenza dell'inconscio, per cui i fattori che ci orientano verso un'azione o l'altra possono non esserci del tutto trasparenti.

Ciò che però l'uomo è in grado di fare è di tornare in modo riflessivo sulla determinazione del giudizio morale, in una sorta di «*auditing* morale a sé stesso»[194], in modo da valutare una seconda volta l'azione che sta scegliendo di compiere, prendendo ad esempio in considerazione altri vincoli/regole morali o verificando che l'azione scelta soddisfi effettivamente i vincoli/regole che erano stati determinati all'inizio del processo decisionale. Con ciò si è in grado di rendere conto della propria azione, cioè di renderla spiegabile a noi e anche agli altri. Attraverso questo passaggio, la decisione diventa un atto di responsabilità morale, in grado di soddisfare i criteri di "responsabilità" (*responsibility*) e "rendicontabilità" (*accountability*)[195]. Si può interpretare anche secondo questo punto di vista l'*electio* di Tommaso, in cui la volontà (legata sempre anche alla ragione) torna sulle azioni passibili di scelta, valutando quale sia da portare a compimento: una sorta di "controllo finale"; in ciò, la volontà può anche decidere di sospendere il suo atto e di non proseguire ulteriormente l'azione, se ad esempio siano sopraggiunte altre informazioni rilevanti o sia cambiato il contesto. Facciamo notare che all'interno della trattazione dell'Aquinate un riferimento importante va fatto anche alla virtù intellettuale della *prudentia*: un abito della ragione, che fornisce ad essa una predisposizione retta nei riguardi dei mezzi ordinati al fine[196].

Il problema degli attuali sistemi di IA, come già evidenziato, risiede non solo nell'opacità dei suoi processi ma anche nella difficoltà per lo stesso sistema di rendere questi processi spiegabili a una interlocuzione esterna. Gianfranco Basti propone lo sviluppo di un meta-sistema di valutazione etico-legale[197] delle decisioni del *software*, in modo simile a quanto avviene per i sistemi di IA che si occupano di dimostrazioni matematiche, che verificano la correttezza di queste ultime tramite sistemi di meta-controllo della consistenza. Un sistema etico dotato di un programma di valutazione permetterebbe di controllare le decisioni del *software* prima che esse

[194] Cfr. G. Basti, *La sfida etica dell'intelligenza artificiale e il ruolo della filosofia*, p. 314.

[195] Cfr. *ibidem*.

[196] Cfr. Tommaso d'Aquino, *La Somma Teologica*, I-IIae, q. 57, a. 5, *resp.*, «Invece la buona predisposizione di un uomo rispetto ai mezzi richiede il diretto intervento di un abito della ragione: poiché deliberare e scegliere atti aventi per oggetto i mezzi appartiene alla ragione. È quindi necessario che nella ragione vi sia una virtù intellettuale che le conferisca una predisposizione retta nei riguardi dei mezzi ordinati al fine. E questa virtù è la prudenza. Quindi la prudenza è una virtù necessaria a ben vivere»

[197] Cfr. G. Basti, *La sfida etica dell'intelligenza artificiale e il ruolo della filosofia*, p. 316.

diano luogo a delle azioni sul mondo reale. Fino a che questa opacità nei processi decisionali non verrà risolta, resta a nostro avviso pericoloso affidarsi in modo cieco e non consapevole ad applicazioni di IA con risvolti etici, che al giorno d'oggi si riscontrano in un numero crescente di ambiti: finanziario (concessione prestiti, automatizzazione degli investimenti), sociale-occupazione (*screening* preventivo dei CV, raccolta informazioni per profilazione), giudiziario (imbastimento prove, determinazione quota cauzione, stima della recidiva), sistemi di guida automatizzata, e molti altri. Specialmente negli ambiti più problematici, ci sarà bisogno di un essere umano che valuti e certifichi i risultati calcolati dalla macchina; la scelta, come tale, dovrà (o dovrebbe, visto lo sdoganamento di tali tecnologie di IA) sempre ricadere su un essere umano o permettere l'ingresso di quest'ultimo all'interno del processo decisionale[198].

Se volessimo tradurre in un linguaggio più informatico – legato ai processi – ciò che abbiamo visto fino a questo punto con l'esposizione della teoria dell'azione di Tommaso, potremmo affermare che ogni passaggio presenta sempre un "doppio controllo" da parte delle facoltà della ragione e della volontà e non si risolve mai nella prevaricazione dell'una con l'altra; *verum et bonum convertuntur*, ciò che la ragione "vuole", oltre ad esser vero è anche buono, ciò su cui la volontà "ragiona", oltre ad esser buono è anche vero.

Se riprendiamo lo schema in dodici atti di pagg. 104-105, possiamo suddividere i vari momenti a coppie, come è già stato fatto, e tradurli in un ipotetico programma per *software* di IA, che si occupi di integrare la parte "razionale" (nel senso che questa ha nella teoria degli agenti razionali artificiali, cioè della massimizzazione di una funziona di utilità) con la parte più propriamente "etica":

1. Atti riguardanti il fine
    a. In questa sezione ciò che conta maggiormente è l'intervento umano: come abbiamo visto, ha poco senso parlare di "volizione" o "intenzione" del fine per una macchina. I "fini remoti" della macchina, cioè i valori etici principali da salvaguardare, devono essere inseriti necessariamente dai programmatori. Anche gli stessi "fini remoti" dell'agente "razionale" saranno inseriti a priori (es. un algoritmo di analisi investimenti avrà come obiettivo quello di individuare gli investimenti più fruttuosi; quello LLM di "apprendere" una lingua).

[198] Precisiamo che questa osservazione non vuole sembrare un'opposizione radicale alle tecnologie di IA o una loro piena bocciatura. Semplicemente riteniamo che escludere l'uomo dal processo decisionale in applicazioni delicate significhi dimenticare l'aspetto "strumentale" di tali applicazioni di IA.

2. Atti riguardanti i mezzi
    a. *Consilium*: il *software* di IA conduce una ricerca delle possibili azioni (i cosiddetti "mezzi") da intraprendere, non preferendo alcun modello etico rispetto all'altro e non ponendo in rilievo alcun valore a questo livello. Ad esempio, un sistema di investimento potrebbe ricercare tutti i possibili investimenti fruttiferi; un sistema d'arma automatizzato potrebbe individuare tutti i bersagli colpibili; un *software* giudiziario potrebbe stilare una classifica di membri della popolazione più a rischio di delinquenza; il *software* di un'auto potrebbe valutare se scansare un'auto e finire sul marciapiede o tentare una frenata d'emergenza.
    b. *Consensus*: il *software* di IA valuta le azioni precedenti sulla base del rispetto dei valori principali decisi dai programmatori. Il sistema di investimento potrebbe evitare di proporre la collocazione di capitali in aziende che sfruttano la manodopera (valore di "rispetto della dignità umana") o quante intrattengono rapporti con stati non-democratici; il sistema d'arma potrebbe stimare quali soluzioni permettano di salvaguardare vite innocenti e quali no; il *software* giudiziario potrebbe stimare quanto le sue predizioni possano contenere presupposti razziali/discriminatori; il *software* dell'automobile potrebbe valutare quali delle due azioni abbia più probabilità di riuscita o di salvaguardare il maggior numero di persone e oggetti.
    c. *Electio*: qui come abbiamo detto si concentra il nucleo della discussione. Per i sistemi d'arma, ad esempio, questo atto dovrebbe essere sempre supervisionato da un essere umano, lo stesso per tutti gli altri ambiti di applicazione. Nelle macchine a guida autonoma già circolanti e nei prototipi di quelle a venire, il paradigma che si osserva è quello che esclude completamente l'uomo dalla possibilità di agire sui comandi, cioè di fatto togliere il potere di scelta e di azione, non solo in situazioni normali, ma anche qualora si presenti una situazione di pericolo. I modelli non dovrebbero solo fornire quale azione è da preferire e perché in un tale contesto, ma anche una lista di tutte le altre azioni possibili e le loro spiegazioni; in questo modo, l'addetto umano avrebbe a disposizione l'intero contesto di "*consilium+consensus*" per capire se al modello possa essere sfuggito qualcosa o se abbia tratto delle conclusioni errate.

È chiaro che per i modelli di IA, anche per i valori morali si tratterà sempre di minimizzare una funzione, cioè di valutare quanto disti quantitativamente la previsione del *software* dal valore desiderato inserito da noi esseri umani. Come abbiamo avuto modo di capire, l'essere umano non agisce in modo statistico né tanto meno algoritmico, ma è frutto di un dinamismo tra facoltà che si innesta sull'apertura trascendentale di cui l'uomo è dotato. I *software* di IA, anche nella loro applicazione agli "agenti morali artificiali", avranno costitutivamente un orizzonte limitato, con metodi e procedure anch'essi limitatati, seppur non strettamente determinati a priori nelle loro azioni.

5.5 Considerazioni intorno alle conseguenze dell'azione

Dopo aver analizzato nel dettaglio il processo decisionale umano, Tommaso affronta nelle questioni 18-21 la componente propriamente "morale" dell'atto umano, connettendolo quindi con la determinazione di quali azioni possano essere definite buone e quali cattive e quali elementi debbano essere presi in considerazioni per fare questa valutazione. Nella I-IIae, q. 18[199] vengono specificate le dimensioni principali di cui tenere conto, cioè l'individuazione di un contenuto (*objectum*) dell'azione – con le debite circostanze al contorno – e la finalità specifica a cui esso viene orientato. Tale contenuto d'azione non deve essere confuso con l'azione fisica, ma è piuttosto la "forma" dell'agire umano, il modo specifico di essere in quanto azione umana. Inoltre esso non è rilevabile immediatamente dall'esterno, ma è individuabile solo attraverso la domanda "che cosa sto facendo?" o "che cosa stai facendo?"; difatti ciò che si sta veramente facendo è trasparente solamente alla persona che sta compiendo l'azione, mentre gli altri potranno solamente interpretare le sue azioni e desumerne (senza pretesa di correttezza) quale sia il suo contenuto d'azione.

Ad uno stesso atto fisico che è il medesimo, se osservato dall'esterno possono corrispondere diversi contenuti d'azione, a seconda della persona che compie l'azione: ad esempio, all'atto fisico di mangiare un piatto di lasagne potrebbero corrispondere diversi contenuti d'azione, quali "mi sto nutrendo", "mi sto ricordando di mia nonna" o "sto accontentando mio suocero". È chiaro che la bontà o meno di un atto umano dipenderà primariamente dalla qualità del contenuto d'azione che poi potrà venire meglio specificato dall'inclusione delle debite circostanze che accompagnano e dalle conseguenze che seguono tale atto.

[199] Cfr. I-IIae, q. 18, aa. 2, 3, 4.

È evidente che nel momento in cui compiano un'azione, la nostra intenzione è quella di raggiungere il fine in un certo modo, tenendo conto anche di certe conseguenze che siamo sicuri accadranno; denomineremo queste ultime "conseguenze intese". Se giochiamo a basket e passiamo la palla ad un nostro compagno, ci aspettiamo che con un passaggio ben eseguito quest'ultimo riesca a prendere la palla e a continuare l'azione. Supponiamo però che il compagno prenda male tale palla e si prende una storta alle dita, decretando la fine della sua partita. È chiaro che questa conseguenza non era intesa, e quindi non si può imputare a chi ha lanciato la palla di aver voluto far male intenzionalmente al compagno. Conseguenze di questo tipo vengono dette "conseguenze prevedibili", che possono accadere saltuariamente, ma non sono ricercate di per sé dal soggetto dell'azione.

Una situazione di questo tipo, in cui cioè ad un'azione corrispondono due differenti effetti (o conseguenze), di cui però uno è intenzionale mentre l'altro no, viene già descritta da Tommaso nella II-IIae, q. 64, all'articolo 7, intitolato «se sia permesso uccidere per difendersi». Lì si dice nel *Respondeo*:

> nulla impedisce che un atto abbia due effetti [*unius actus esse duos effectus*], di cui uno è intenzionale [*in intentione*] e l'altro involontario [*praeter intentionem*]. Gli atti morali però ricevono la specie da ciò che è intenzionale, non da ciò che è involontario [...]. Perciò dall'azione della difesa personale possono seguire due effetti [*duplex effectus*], il primo dei quali è la conservazione della propria vita, mentre l'altro è l'uccisione dell'attentatore.

L'interpretazione di questo passo da parte di interpreti successivi ha dato vita alla cosiddetta "dottrina del duplice effetto", spesso interpretata erroneamente in senso consequenzialistico. Un'interpretazione di questo tipo è chiaramente erronea e porta ad indebite conclusioni il discorso che Tommaso fa in questo articolo, che, non a caso, trova posto nella *Secunda Secundae*, dove l'Aquinate si occupa di un'etica speciale. Nell'esempio della partita di basket, se io passassi la palla al mio compagno lanciandola volontariamente in modo eccessivamente forte, la mia intenzione sarebbe quella di fargli del male e solo secondariamente passargli la palla.

Come abbiamo visto, la teoria degli agenti razionali artificiali predilige una teoria di tipo consequenzialistico, orientata più alla valutazione delle conseguenze dell'azione che al valore dell'azione in sé (cioè dell'*objectum*). Anche per l'etica di una agente morale artificiale, il paradigma utilitaristico-consequenzialistico è quello che meglio si presta ad una traduzione in termini computazionali.

Classicamente, come viene presentato anche da Tommaso in alcuni articoli della Prima e della Seconda Parte della *Somma*, le conseguenze possono essere di due tipi: quelle intese, cioè ricercate e che si accompagnano in modo stretto con lo svolgimento dell'azione, e quelle previste, che possono accompagnare l'azione, ma in modo contingente e non causale. Come abbiamo visto, nel momento in cui si ha l'intenzione del fine si ha anche l'elezione dei mezzi, per cui scegliendo un'azione da compiere teniamo conto anche delle possibili conseguenze che possono discenderne, nei limiti di quelle che possiamo prevedere. È chiaro che la bontà o meno di un'azione dipenderà da molti fattori: la scelta del fine, la determinazione dei mezzi adeguati al suo raggiungimento, le circostanze e anche le conseguenze. Tommaso dedica la questione 7 della I-IIae alla discussione sulle circostanze degli atti umani[200], mentre dedica una più ampia parte alla trattazione della bontà e della malizia degli atti umani alle questioni 18-21.

Ora, per una corretta considerazione della "teoria del duplice effetto" così come viene esposta da Tommaso, è necessario leggere la questione 20 della I-IIae, il cui articolo 5 si intitola: «se gli eventi successivi possano accrescere la bontà o malizia dell'atto esterno». qui leggiamo nel *Respondeo*:

> l'evento successivo può essere previsto [*praecogitatus*] o non previsto. Se è previsto, è chiaro che esso accresce la bontà o la malizia. Quando infatti uno, pur prevedendo che dalla propria azione possono seguire molti mali, non desiste, dimostra che la sua volontà è più pervertita. Se invece l'evento successivo non è previsto, allora bisogna distinguere. Se tale evento segue l'atto direttamente e nella maggior parte dei casi, allora l'evento aggiunge bontà o malizia all'azione: è infatti evidente che un atto è migliore nel suo genere per il fatto che da esso segue un maggiore numero di beni, ed è peggiore se da esso segue un maggior numero di mali. Al contrario, se l'evento consegue all'azione in casi sporadici, e solo indirettamente [*per accidens*], allora l'evento non aggiunge nulla alla bontà o alla malizia dell'atto: infatti non si può giudicare una cosa in base a quanto le appartiene indirettamente e accidentalmente, ma solo in base alle sue proprietà essenziali.

Nella prospettiva consequenzialistica[201], invece, la differenza tra conseguenze intese e previste viene eliminata, creando una singola categoria di conseguenze che devono essere tenute tutte in conto per la determinazione dell'azione effettiva. Così facendo è pertanto impossibile stabilire a priori se una data azione sia buona o cattiva: l'uccisione di un innocente potrebbe trovare una

[200] Cfr. I-II, q. 7, a. 3. Tommaso riprende Aristotele (dal III libro dell'*Etica*) e Cicerone (dalla *Retorica*), individuando le principali circostanze in: chi, che cosa, dove, con quali mezzi, perché, in che modo, quando e intorno a che cosa.

[201] Una filosofa che ha duramente criticato questa teoria è stata G.E.M. Anscombe. Nel 1956 si oppose pubblicamente alla decisione dell'Università di Oxford di consegnare al presidente americano Truman una laura *honoris causa* in diritto civile (lo stesso presidente che ordinò il bombardamento atomico del Giappone). Particolarmente conosciuta per il suo scritto *Modern Moral Philosophy*, si veda a proposito la sua raccolta di scritti in G.E.M. Anscombe, *Scritti di etica*, trad. it. a cura di S. Cremaschi, Morcelliana, Brescia 2022.

giustificazione in vista di un "bene più grande" (ad esempio quella di uccidere un terrorista). Se pensiamo alla proposta di algebrizzare l'etica tramite il calcolo dell'utilità complessiva (dato dalla somma algebrica delle singole conseguenze), è chiaro che in questo calcolo possono rientrare tutte le conseguenze che possiamo aspettarci: sia intese che previste.

Facendo così si elimina tutto l'aspetto "intenzionale" dell'azione, cioè la scelta non è più rivolta a quegli specifici "mezzi" (con le loro conseguenze intese) che ci conducono al fine. Piuttosto si cerca di "ri-orientare" l'intenzione mettendo in risalto alcune conseguenze rispetto ad altre, in modo da giungere ugualmente al fine, anche se per conseguirlo sono stati individuati dei "mezzi" non adatti o non tollerabili eticamente. Sganciare un missile su un mercato pieno di innocenti per colpire un terrorista non è tollerabile; dunque si dovrebbe desistere dal compiere l'azione il cui fine è quella di uccidere il terrorista, che ha come conseguenza prevista quella di colpire anche molte vittime innocenti (è chiaro che in questa circostanza, all'evento di "sganciare il missile" segue necessariamente l'evento "morte di innocenti" e quindi, come visto sopra, quest'ultimo non può essere considerato un effetto involontario dell'azione). Tuttavia si potrebbe obiettare che, se quest'ultimo non viene ucciso immediatamente, farebbe perdere le sue tracce, e le risorse impiegate per la sua ricerca sarebbero state sprecate, oppure potrebbe colpire a breve in un'altra città e fare molti più morti rispetto di quelli che si otterrebbero uccidendolo adesso. In questo modo, l'azione di sganciare il missile per uccidere il terrorista viene giustificata da tutte le possibili conseguenze negative che seguirebbero l'annullamento dell'ordine di colpire l'obiettivo. Come conclusione, si potrebbe affermare che "uccidere gli innocenti" era solamente una conseguenza prevista, ma in modo accidentale, e che non rientrasse nella specificazione dell'azione propriamente intesa: quella di "porre fine alla minaccia nel più breve tempo possibile".

Nella progettazione sia degli "agenti razionali artificiali" sia degli "agenti morali artificiali" dovrebbe essere tenuta in conto sia la differenza tra le conseguenze intese e previste, sia il modo in cui le conseguenze previste entrino a specificare la bontà o meno di un atto. Spesso infatti, data la complessa capacità computazionale dei sistemi di IA, si estende il calcolo delle possibili conseguenze anche a quelle più remote, così da riuscire a determinare l'azione migliore possibile. Se da una parte questa può essere una buona pratica, perché i computer sono in grado di elaborare in modo molto più veloce rispetto all'uomo grandi quantità di informazioni e quindi possono tenere in conto molte più variabili, dall'altra bisogna evitare di ampliare l'orizzonte più di quanto sia necessario, includendo conseguenze che poco hanno a che fare con l'azione o la cui probabilità di accadere è infinitesima. Una considerazione ampia delle conseguenze apre

anche una questione circa la responsabilità dei *software*: di più cose si tiene conto, maggiore sarà la responsabilità associata all'azione. Tuttavia, come precisato da Tommaso nell'articolo riportato poco sopra, la responsabilità non si può estendere a tutte le conseguenze.

Per quelle applicazioni di IA che si basano (o baseranno) su un approccio di tipo consequenzialista, si rende quindi necessario operare una duplice correzione, per evitare che i *software* possano determinare azioni giustificate da una errata considerazione delle conseguenze. In primo luogo devono essere distinte le conseguenze intese da quelle invece solamente previste, e in secondo luogo dev'essere operata una distinzione in seno a quelle previste, per cui quelle che si accompagnano quasi sempre come conseguenza di un'azione devono essere tenute in maggior conto rispetto alle altre. Per tenere conto di quest'ultima differenza, si potrebbe inserire un fattore correttivo ξ, che tenga conto del grado di probabilità delle conseguenze previste a seguito di un'azione e aumenti l'importanza di quelle che sono più probabili ad accadere.

6 Conclusioni

L'approfondimento degli algoritmi che regolano il funzionamento dell'IA e la panoramica attorno all'importanza e agli sviluppi di un'etica dell'IA ci hanno permesso di mettere a fuoco alcuni temi fondamentali sul rapporto tra intelligenza umana e artificiale, con particolare attenzione alla loro traduzione nella dimensione dell'agire concreto sul mondo da parte di agenti razionali artificiali.

Il termine IA rischia ancora oggi di destare qualche preoccupazione, soprattutto in riferimento ad una certa cultura fantascientifica sul tema, in cui si prospetta la sottomissione della specie umana ad una IA in grado di superare l'uomo in ogni ambito: razionale e pratico. Come abbiamo visto, dal punto di vista ingegneristico questo è ancora molto lontano dal poter esser anche solo progettato, in quanto le applicazioni di IA oggi a nostra disposizione hanno abilità che si applicano entro specifici domini: visione artificiale, riconoscimento del linguaggio, e così via.

Dal punto di vista filosofico, la possibilità che l'intelligenza possa arrivare ad essere una copia fedele di quella umana, non solo dal punto di vista dei processi biologici, ma anche di quelle caratteristiche che sembrano essere appannaggio esclusivo della mente umana è ancora dibattuta. Come abbiamo cercato di mostrare, la differenza tra le due "intelligenze" non è solamente di tipo quantitativo, ma propriamente qualitativo: l'orizzonte delle macchine è tipo algoritmico, e dunque intrinsecamene limitato; mentre quello umano è di tipo trascendentale, e dunque infinito. Con ciò si respinge una certa visione riduzionistica della mente al cervello, per cui quest'ultima non sarebbe altro che l'espressione dei molteplici collegamenti elettro-chimici presenti nei neuroni.

L'analisi delle fondamenta teoriche degli "agenti razionali" artificiali, cioè della teoria della probabilità e dell'utilità, ci ha permesso di notare come l'uomo possa essere definito a volte "irrazionale" dal punto di vista della teoria dell'utilità, poiché può compiere scelte che vanno contro le predizioni di tale teoria. È presente quindi uno scarto tra una razionalità "matematico-probabilistica" e la razionalità propriamente umana, di cui la teoria può descrivere solamente una parte. Definire "irrazionale" una scelta umana perché non massimizza l'utilità attesa significa perdere di vista l'orizzonte in cui l'uomo vive e invertire il rapporto esistente tra macchina ed essere umano. La scelta umana è propriamente razionale perché capace di considerare insieme molti aspetti dell'uomo, come i sentimenti e le passioni (il termine *λόγος* deriva da *léghein*, che include anche il significato di "raccogliere insieme"); mentre gli agenti artificiali possono considerare solamente alcuni aspetti, che possano essere tradotti in un linguaggio algoritmico.

Il capitolo sull'etica dell'IA ci ha mostrato come le teorie etiche più utilizzate siano quelle utilitaristiche, che meglio si prestano ad una traduzione algoritmica e a una implementazione nelle macchine. Le teorie etiche riconosciute per lo studio di un'etica per le macchine si restringono a soli tre approcci principali (deontologico, utilitaristico, doveri *prima facie*), ignorando per gran parte il contributo di quella filosofia che va da Aristotele a Kant.

Un punto di particolare importanza riguarda il rapporto che intercorre tra le teorie etiche e la loro possibile implementazione tramite algoritmi in *software* di IA. Le teorie, infatti, devono poter essere tradotte in modelli procedurali per essere utilizzate all'interno dei modelli di intelligenza artificiale. Come abbiamo visto, alcune teorie si possono prestare facilmente a questo tipo di traduzione, non andando ad inficiare in maniera significativa la teoria nel suo complesso; un esempio di ciò è l'utilitarismo, in particolare nella sua declinazione consequenzialistica. Molto più complesso è elaborare un modello procedurale che rispecchi in maniera fedele l'impostazione di Aristotele, di Tommaso o di Kant. Ciò è dovuto dal fatto che il dinamismo dell'agire etico descritto da queste teorie deve venire ricondotto ad una serie di "istruzioni" da implementare nello strumento di IA. È inevitabile quindi che nel passaggio dalla teoria al modello procedurale qualcosa venga perso o vengano compiute delle scelte su che cosa implementare e che cosa no. Il rischio è quello di favorire a priori certe teorie a discapito di altre, per via della maggiore facilità di proceduralizzazione; ma vi è anche il pericolo di snaturare le teorie quando esse vengono ridotte a modelli procedurali, perché necessariamente si opera una scelta su quali elementi della teoria tenere e di quali invece sia possibile fare a meno, perdendo così l'organicità della stessa.

Attraverso l'analisi dell'azione umana proposta da Tommaso, siamo stati in grado di ricostruire il processo decisionale umano, che avviene in maniera completamene differente rispetto a quello utilizzato da una qualsiasi applicazione di IA (o di macchina in generale). Quest'ultima infatti è basata ancora una volta su un approccio di tipo statistico-matematico, in cui è necessario minimizzare l'errore di una funzione rispetto ad un obiettivo dato. L'uomo è invece radicalmente libero, anche rispetto a quali valori scegliere e seguire, e nel dinamismo della scelta sono sempre implicate le facoltà della volontà e della ragione. Lo scegliere, quindi, non può essere separato dalla dimensione valoriale, e quindi in ultima misura morale. Inoltre, l'uomo è capace di spiegare le motivazioni delle sue azioni, mentre i *software* di IA presentano un'intrinseca opacità dei processi decisionali.

L'errore che non deve essere fatto con le applicazioni di IA è quello di considerarle "neutre" dal punto di vista morale. Esse infatti si configurano come strumenti e come tali devono

raggiungere l'obiettivo che gli viene prefissato dall'esterno. Se lo scegliere tipicamente umano non può fare a meno della dimensione morale, è chiaro che questa dimensione si rifletterà anche nella definizione degli "agenti morali artificiali" e nella sistematizzazione dell'etica dell'IA, in cui i valori vengono determinati dai programmatori o vengono acquisiti tramite tecniche di apprendimento automatico.

La scomposizione che Tommaso opera del processo di scelta può a nostro parere essere utile per un approfondimento che vada a emularlo all'interno dei sistemi di IA. In modo simile a quanto succede per il binomio ragione-volontà, un sistema di controllo etico dei processi dei *software* di IA non dovrebbe solamente valutare se l'output sia compatibile o meno con certi valori, ma il controllo dovrebbe applicarsi anche ai processi che portano all'output (i "mezzi"). In un'ottica di lungo periodo, un "sistema etico" potrebbe affiancare un meta-sistema di controllo dei processi, in modo da agire anche sui pesi e sui *bias* dei singoli neuroni artificiali: se una certa connessione tra neuroni porta ad un disallineamento valoriale (ad es. mancato rispetto della vita umana), il sistema etico potrebbe decidere di indebolire tale connessione e il *software* arriverebbe dunque a "minimizzare" una diversa funzione, che tenga conto anche della codifica valoriale.

7 Bibliografia

Testi sull'IA

Adams S. *et alii*, *Mapping the landscape of human-level artificial general intelligence*, "AI Magazine", vol. 33, 2012, pp. 25-42.

Aiello L.C., Dapor M., *Intelligenza artificiale: i primi 50 anni*, "Mondo Digitale", vol. 2, 2004, pp. 3-20.

Allen C., Varner G., Zinser J., *Prolegomena to any future artificial moral agent*, "Journal of Experimental & Theoretical Artificial Intelligence", vol. 12, 2000.

Buchanan B.G., Feigenbaum E.A., Lederberg J., *Heuristic DENDRAL: A program for generating explanatory hypotheses in organic chemistry*, "Machine Intelligence", 4, 1968. (PDF) Heuristic DENDRAL: A program for generating explanatory hypotheses in organic chemistry (researchgate.net)

Burattini E., R. Cordeschi (a cura di), *Intelligenza artificiale*, Carrocci editori, Roma 2001.

Codenotti B., Leoncini M., *La rivoluzione silenziosa*, Codice edizioni, Torino 2020.

Cordeschi R., *L'intelligenza artificiale*, Storia del pensiero filosofico e scientifico, vol. 8, Garzanti, Milano 1996. Reperibile online come parte di capitolo.

David E., *Would you kill the fat man?*, Princeton University Press, Oxford 2014.

Davis M., traduzione di G. Rigamonti, *Il calcolatore universale*, Adelphi, Milano 2003.

Feigenbaum E.A. *et alii.*, *DENDRAL: a case study of the first expert system for scientific hypothesis formation*, "Artificial Intelligence", vol. 61, 1993, pp. 209-261. Reperibile su dendral-history.pdf (mit.edu)

Floridi L., Chiriatti M., *GPT-3: Its nature, scope, limits, and consequences,* "Minds and machines", vol. 30, 2020, pp. 681-694.

Floridi L., *The 4th revolution: how the infosphere is reshaping human reality*, Oxford University Press, New York 2014.

Floridi L., *The logic od information: a theory of philosophy as conceptual design*, Oxford University Press, Oxford 2019.

Goertzel B., *Artificial General Intelligence: Concept, State of the Art, and Future Prospects*, "Journal of Artificial Intelligence", vol. 5, 2014, pp. 1-46.

Ielmini D., *Intelligenza artificiale: l'approccio neuromorfico*, "Mondo Digitale", vol. 79, 2018.

Ilkou E., Koutraki M., *Symbolic vs Sub-symbolic AI Methods: friends or enemies?*, Conference on Information and Knowledge Management, Galway 2020.

Loreti M., *Teoria degli errori e fondamenti di statistica. Introduzione alla fisica sperimentale*, Zanichelli 1998.

McCarthy J., Minsky M.L, Rochester N., Shannon C.E, *A proposal for the Darmounth Summer Research Project on Artificial Intelligence*, "AI Magazine", vol. 27, 1955.

McCulloch W.S. e Pitts W., *A logical calculus of the ideas immanent in nervous activity*, "Bulletin of Mathematical Biology", Vol. 52, pp. 99-115, 1990. Reperibile: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiX-MTZw6aIAxWXgf0HHaMmKQUQFnoECBEQAQ&url=https%3A%2F%2Fwww.cs.cmu.edu%2F~.%2Fepxing%2FClass%2F10715%2Freading%2FMcCulloch.and.Pitts.pdf&usg=AOvVaw0yhn00mFm05rk66caMGT8O&opi=89978449

Minsky M., Papert S.A., *Perceptrons: an introduction to computational geometry*, The MIT Press, Cambridge 1988, https://doi.org/10.7551/mitpress/11301.001.0001

Mitchell T.M., *Machine Learning*, McGraw-Hill, New York 1997.

Nilsson N.J., *Logic and artificial intelligence*, "Artificial intelligence", vol. 47, 1991, pp. 31-56.

Pagani P., *La forma dell'élenchos*, «Divus Thomas», vol. 118, 2015, pp. 152-191.

Pearl J., *Probabilistic reasoning in intelligent systems*, Morgan Kaufmann Publisher Inc., San Francisco 1988.

Portinale L., *Intelligenza artificiale: storia, progressi e sviluppi tra speranze e timori*, "Media Laws", vol. 3, 2021, pp. 13-28.

Rosenblatt F., *The perceptron, a perceiving and recognizing automaton*, Project PARA, Cornell Aeronautical Laboratory, Ithaca 1957.

Rosenblatt F., *The perceptron: a probabilistic model for information storage and organization in the brain*, "Psychological Review", vol. 65, 1958, Rosenblatt1958.pdf (upenn.edu)

Rovighi S.V., *Elementi di filosofia 1*, Editrice Morcelliana, Brescia 2022.

Russell S.J., Norvig P., (a cura di) F. Amigoni, *Intelligenza artificiale, un approccio moderno*, vol.1, Pearson Italia, Milano 2021.

Russell S.J., Norvig P., *Artificial Intelligence, A Modern Approach*, Third edition, Pearson Education Inc., New Jersey 2010.

Samuel A.L., *Some studies in machine learning using the game of checkers*, "IBM Journal", Vol.3, Luglio 1959.

Searle J., *Minds, Brains and Programs*, "The behavioural and brain sciences", vol. 3, 1980, pp.417-457.

Turing A.M., *Computing Machinery and Intelligence,* "Mind", vol. 59, 1950, pp. 433–460. *JSTOR*, http://www.jstor.org/stable/2251299

von Neumann J., O. Morgenstern, *Theory of games and economic behavior*, Princeton University Press, Princeton 1953.

Testi sull'etica dell'IA

Allen C., Varner G., Zinser J., *Prolegomena to any future artificial moral agent*, "Journal of Experimental & Theoretical Artificial Intelligence, vol. 12, 2000.

Allen C., Wallach W., Smith I., *Why machine ethics?*, "IEEE Intelligent System", vol. 21, 2006, pp. 12-17.

Anderson M., Anderson S. L., *Machine ethics, creating an ethical intelligent agent*, "AI Magazine", vol. 28, 2007.

Anderson M., Anderson S.L., Armen C., *An approach to computing ethics*, "IEEE Intelligent system", vol. 21, 2006.

Anderson M., Anderson S.L., *Machine ethics*, Cambridge University press, Cambridge 2011.

Cracco L., *Machine ethics: stato dell'arte e osservazioni critiche*, "Filosofia", vol. 65, 2020, pp. 75-89. DOI: https://doi.org/10.13135/2704-8195/5234

Floridi L., Cabitza F., *Intelligenza artificiale: l'uso delle nuove macchine*, Bompiani, Milano, 2021.

Foot P., *The problem of abortion and the doctrine of double effect*, "Oxford Review", vol. 5, 1967.

Fossa F., *Etica funzionale. Considerazioni filosofiche sulla teoria dell'agente morale artificiale*, "Filosofia", vol. 65, 2020, pp. 91-106. DOI: https://doi.org/10.13135/2704-8195/5235

Moor J.H., *Four kinds of ethical robots*, "Philosophy Now", vol. 72, 2009.

Moor J.H., *The nature, importance and difficulty of machine ethics*, "IEEE Intelligent system", vol. 21, 2006.

Quinn M., *Bentham on mensuration: calculation and moral reasoning*, "Utilitas", vol. 26, 2014, pp. 61-104.

Wallach W., Allen C., *Framing robot arms control*, "Ethics and Information Technology", vol. 15, 2013, pp. 125-135.

Wallach W., Allen C., Smit I., *Machine morality: bottom-up and top-down approaches for modelling human moral faculties*, "AI & Society", vol. 22, 2008, pp. 565-582.

Testi Classici

Aristotele, *Etica Nicomachea*, Laterza, Bari 2018.

Kant I., B. Constant, S. M. Carmignani (a cura di), *Il diritto di mentire*, Passigli Editori, Firenze 2008.

Kant I., F. Gonnelli (traduzione), *Fondazione della metafisica dei costumi*, Laterza, Bari 2018.

Tommaso d'Aquino, (a cura di) domenicani italiani, *La Somma Teologica*, PDUL Edizioni Studio Domenicano, Bologna 1985.

Tommaso d'Aquino, *La Somma teologica*, Edizioni Studio Domenicano, Bologna 2014.

Testi sull'etica di Tommaso

Anscombe G.E.M., *Scritti di etica*, trad. it. a cura di S. Cremaschi, Morcelliana, Brescia 2022.

Basti G., *La sfida etica dell'intelligenza artificiale e il ruolo della filosofia*, "Aquinas", vol. 65, 2022.

Campodonico A., Vaccarezza M.S., *La pretesa del bene, teoria dell'azione ed etica in Tommaso d'Aquino*, Orthotes Editrice, Napoli 2012.

Finnis J., (traduzione di) A. Cevolini, *Contenuto e intenzione nei giudizi morali secondo Tommaso d'Aquino*, "Divus Thomas", vol. 104, n. 2, 2001, pp. 21-23. Reperibile su https://www.jstor.org/stable/45074911

Floridi L., *The ethics of artificial intelligence: principles, challenges, and opportunities*, Oxford university press, Oxford 2023.

Kluxen W., C. Vigna (a cura di), *L'etica filosofica di Tommaso d'Aquino*, Vita e Pensiero, Milano 2005.

Vaccarezza M.S., *Le ragioni del contingente, la saggezza pratica tra Aristotele e Tommaso d'Aquino*, Orthotes Editrice, Napoli 2012.

Sitografia

3Blue1Brown - But what is a GPT? Visual intro to Transformers | Deep learning, chapter 5

3Blue1Brown - But what is a Neural Network?

Artificial Intelligence (Stanford Encyclopedia of Philosophy)

Che cos'è l'intelligenza artificiale? | Tematiche | Parlamento europeo (europa.eu)

Cos'è l'Intelligenza Artificiale (AI)?| IBM

Cybernetics. Transactions of the Eighth Conference March 15-16, 1951, Macy Foundation, New York, reperibile su AITopics | Presentation of a Maze-Solving Machine

Ethics guidelines for trustworthy AI | Shaping Europe's digital future (europa.eu)

GDPR UE 2016: La protezione dei dati nell'UE - Commissione europea (europa.eu)

Le frontiere dell'umano: i dilemmi dell'algoretica. Intervista a Paolo Benanti - Pandora Rivista

Lighthill Report, Cambridge University, luglio 1972, accessibile su: lighthill.dvi (ed.ac.uk)

Stanford Encyclopedia of Philosophy all'indirizzo Ethics of Artificial Intelligence and Robotics (Stanford Encyclopedia of Philosophy)

Stanford's robotics legacy | Stanford Report

What Is Artificial Intelligence (AI)? | IBM