Breve introduzione tecnica all'Intelligenza Artificiale

di Paolo Traverso

Abstract: *Artificial Intelligence: A Brief Technical Introduction* – This section is a brief introduction to Artificial Intelligence (AI). Our aim is to introduce some elementary concepts related to AI in a way that is hopefully understandable to people non-expert in the field. In this introduction, we emphasize that AI is much more than Machine Learning and Deep Learning, by introducing the two main different approaches to AI: Model Based AI and AI based on Machine Learning. We summarize the main pros and cons of both approaches. In spite of several AI success stories in the past by Model Based AI, there is no doubt that the current impact and high expectations raised by AI is due to the recent successes in data intensive (supervised) Machine Learning, and especially to Deep Learning. Deep learning has led to impressive gains on most key areas of AI, such as computer vision, natural language understanding, speech recognition, game playing, and robotics. In spite of the significant progress, we still need a lot of work in research and a paradigm shift in AI. The goal for the future will be to provide AI based solutions that can be of great help for our life and, at the same time, reliable, trustworthy solutions that can be used even in areas and systems at "high risk", as stated by the Proposal for a Regulation of the European Union on AI.

Keywords: Model Based AI, Machine and Deep Learning; Integrative AI; AI Regulation.

1. Premessa

Questa introduzione all'Intelligenza Artificiale (AI, nel resto della sezione, dall'inglese "*Artificial Intelligence*") è una introduzione per non esperti. Naturalmente quindi molte delle spiegazioni sono state volutamente semplificate e corrono il rischio di non essere complete e precise. Inoltre l'introduzione tende a sottolineare ed evidenziare certi aspetti che possono essere utili per una analisi di tipo giuridico piuttosto che altri aspetti tecnici.

L'introduzione mira anche a far chiarezza su certi "bias" riportati comunemente nella letteratura divulgativa relativa all'AI, ad esempio che l'AI sia una disciplina nata di recente, che sia solo una AI basata sull'apprendimento automatico (*Machine Learning* e in particolare *Deep Learning*), che l'AI risolva qualunque problema e sia la panacea di tutti i mali, che le tecniche alternative all'apprendimento automatico siano del tutto superate o inutili.

• L'articolo è stato scritto nell'ambito del Jean Monnet project *Trento AI Laboratory* (TrIAL)

DPCE online
ISSN: 2037-6677

2. Intelligenza Artificiale Tradizionale e Moderna?

Esiste una Intelligenza Artificiale Tradizionale e una Moderna?

No.

Sebbene numerose sono le idee e le teorie che possono considerarsi precorritrici della ricerca in Intelligenza Artificiale, fra le quali sicuramente significative sono alcune opere di Alan Turing[1], il termine Intelligenza Artificiale fu coniato per la prima volta nel 1956, durante un workshop al Dartmouth College negli Stati Uniti, dove i "padri dell'intelligenza artificiale" (John McCarthy, Marvin Minsky, Claude Shannon, Nathaniel Rochester) si diedero l'obiettivo di costruire una macchina, un computer in grado d ragionare, apprendere, agire in un modo simile a quello dell'essere umano.

Non è l'obiettivo di questa sezione enunciare una "storia dell'intelligenza artificiale", ma ricordare invece che l'AI non è una disciplina recente, nata negli anni 2000 (come crede erroneamente qualcuno), ma si basa su di una storia di studi, ricerche, sviluppi e applicazioni (alcune di queste di significativo successo) più lunga di mezzo secolo.

È ancor più importante evidenziare che anche le più moderne tecniche di Intelligenza Artificiale, che recentemente hanno avuto un enorme successo e sono riuscite a darci risultati che alcuni anni fa erano impensabili, trovano la base in teorie ben più datate. Ad esempio, l'idea di "Apprendimento Automatico", "*Machine Learning*", "Rete neurale", alla base del moderno "*Deep Learning*" (di cui parleremo nella sezione dedicata all'approccio basato sull'Apprendimento Automatico - il *Machine Learning AI*), trova la sua origine negli studi che risalgono agli anni 60 e in importanti evoluzioni e risultati della ricerca negli anni 80. Gli studi di quegli anni sono ancor oggi alla base della recente ricerca e delle applicazioni in AI. Queste teorie degli anni passati si sono naturalmente evolute in modo significativo, anche attraverso nuove scoperte e paradigmi innovativi, ma devono il loro successo anche e soprattutto all'enorme potere computazionale e all'enorme quantità di dati (provenienti da internet, dal Web, dai sensori) disponibili oggi. Questo è il caso anche di altre tecniche e approcci di AI utilizzati nel passato e anche al giorno d'oggi con successo in applicazioni industriali di grande rilievo, di cui parleremo nella sezione dedicata all'approccio basato sulla modellazione (Il *Model Based AI*).

Quindi in questa introduzione non useremo mai il termine "AI Moderna" e "AI Tradizionale", terminologia utilizzata da alcuni, siccome anche i più significativi approcci moderni che hanno contribuito fortemente alla notorietà dell'AI, come le moderne tecniche di Machine Learning, trovano la loro origine nel passato.

[1] A. M. Turing, *Intelligent Machinery. A report by A. M. Turing for the National Physical Laboratory*, 1948, in bit.ly/3KPYwTH. C. R. Evans, A. D. J. Robertson (a cura di) *Cybernetics: Key Papers*, Baltimore Md. e Manchester, 1968; A. M. Turing, *Intelligent Machinery, A Heretical Theory**, in *Philosophia Mathematica*, 3, 1996, 256-260; C. S. Webster *Alan Turing's unorganized machines and artificial neural networks: his remarkable early work and future possibilities*, in *Evolutionary Intelligence*, 5, 2012, 35–43.

3. Intelligenza Artificiale: Definizione e Diversi Approcci

Oliviero Stock, uno dei più grandi e riconosciuti ricercatori in Intelligenza Artificiale, ha affermato in una recente intervista che se chiedessimo a 100 ricercatori di AI di definire l'AI, avremmo (forse addirittura più di) 100 definizioni diverse.[2] Questo credo derivi da quanto complicato sia definire il termine "Intelligenza", difficoltà amplificata dal fatto che l'AI col termine "Artificiale" si riferisce ad una "Intelligenza posseduta dalle Macchine, dal Computer". Si potrebbe anche discutere se può esistere una macchina intelligente. Ma in questa sezione non affronteremo questo problema filosofico.

Relativamente alla definizione di AI, ci rifaremo a quella data dal Gruppo di Esperti costituito dalla Commissione Europea, *"High-Level Expert Group on Artificial Intelligence"*,[3] il quale, probabilmente per evitare problemi di natura filosofica nella definizione di AI, ha preferito definire non tanto l'Intelligenza Artificiale di per sé, ma le capacità di un Sistema di Intelligenza Artificiale, il cui schema è rappresentato in Figura 1.

> Artificial Intelligence (AI) refers to systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals.

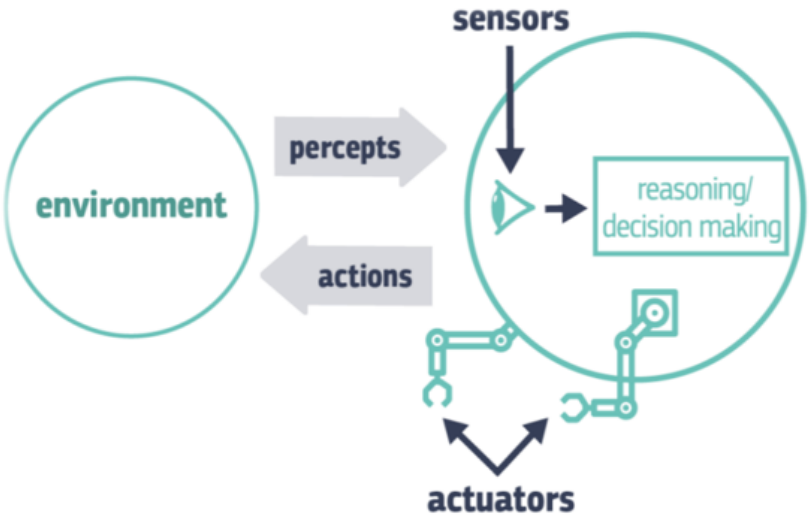

Figura 1: Rappresentazione Schematica di un Sistema di Intelligenza Artificiale

[2] Molto interessante la sua definizione: "Far fare alle macchine lo sforzo di capire noi, anziché viceversa", si veda il video alla pagina www.fbk.eu/en/initiatives/intelligenza-artificiale/.

[3] High level Group in Artificial Intelligence, European Commission, *A Definition of AI: Main Capabilities and Disciplines*, Bruxelles, dicembre 2019, in ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf. Per una definizione di Intelligenza Artificiale, si veda anche la Proposta di Regolamento Europeo sull'Intelligenza Artificiale (*"Proposal for a Regulation of the European Union on AI, report of the European Commission and European Parliament."*, art. 3), in cui un sistema di Intelligenza Artificiale viene definito come un sistema che, per un dato insieme di obiettivi definiti dall'uomo, è in grado di generare output (contenuti, previsioni, raccomandazioni, decisioni) che influenzano l'ambiente con cui interagisce.

DPCE online
ISSN: 2037-6677

Quindi un Sistema di AI deve essere in grado di percepire l'ambiente circostante, analizzarlo e comprenderlo, ragionare e prendere delle decisioni, con un certo livello di autonomia, e compiere delle azioni nell'ambiente.

Anche se molto di alto livello, questa definizione ci pone già degli interessanti interrogativi che hanno a che fare con l'aspetto giuridico e normativo dell'AI. Ad esempio, quale livello d'autonomia può essere concesso ad una macchina? E quando l'autonomia decisionale si unisce alla capacità di agire, ad esempio di un robot, questo ci porta subito alla mente situazioni pericolose per l'essere umano, spesso esaltate in parecchi film distopici. Ma anche il concetto di analizzare l'ambiente circostante e prendere delle decisioni può avere risvolti che hanno a che fare con i diritti dell'essere umano, si pensi ad esempio ad un sistema di Intelligenza Artificiale in grado di capire da una serie di dati se una persona è pericolosa per la società o meno.

Nel seguito di questa sezione vogliamo approfondire un po' di più diverse tecniche di Intelligenza Artificiale, in modo da porre le basi per una analisi anche di tipo giuridico. A questo scopo, a grandi linee e ancora in modo molto approssimativo, si può dire che esistono due approcci diversi in Intelligenza Artificiale, che sottendono metodologie e anche tecnologie diverse:

- L'Intelligenza Artificiale basata sui Modelli (*Model Based AI)*, secondo la quale si definisce e realizza un modello formale, ad esempio matematico, di un certo fenomeno, si "porta sul computer" questo modello e si utilizzano poi strumenti dedicati ad analizzare il modello, ad esempio per verificarne caratteristiche e proprietà, oppure per generare soluzioni che permettano ad un sistema computerizzato di risolvere un problema oppure di raggiungere un obiettivo desiderato in modo automatico.

- L'intelligenza Artificiale basata sull'Apprendimento Automatico ("*Machine Learning AI*"), in cui il modello viene costruito a partire dai dati, tipicamente dando al sistema una serie di esempi dai quali il sistema impara.

Nelle due successive sezioni, approfondiremo un po' di più questi due diversi approcci.

4. L'Approccio basato sulla Modellazione: Il Model Based AI

Nell'approccio basato sui modelli (si veda Figura 2), un progettista definisce e costruisce un modello di un fenomeno. Questo modello deve avere la caratteristica di poter essere inserito in un computer in modo da essere utilizzato dal computer stesso per calcolare, dare risposte, fare analisi, o operare e agire nel mondo.

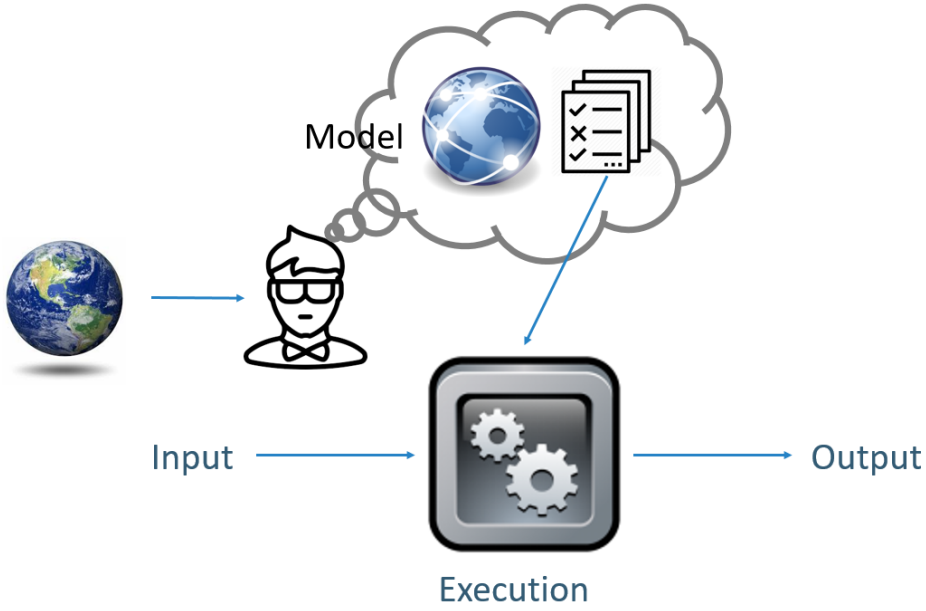

Figura 2: Model Based AI

Un modello formale può essere ad esempio una ontologia,[4] ovvero uno schema concettuale che, ad esempio, classifica diversi concetti. Un semplice esempio è una ontologia che ci dice che gli esseri viventi si distinguono in piante e animali, gli animali in mammiferi e pesci e così via, fino a definirne le loro proprietà. Esistono ontologie in diverse branche della scienza, ad esempio ontologie in biologia e medicina. Si pensi ad esempio a quanto possa essere utile per applicazioni mediche una ontologia delle malattie, delle cure, degli alimenti, delle loro proprietà. Una ontologia è formalmente rappresentabile nel linguaggio formale della logica, in particolare della logica descrittiva.[5]

Più in generale, un formalismo logico, può rappresentare la conoscenza e permettere inferenze e deduzioni. Ad esempio, la logica ci può permettere di rappresentare la conoscenza del fatto che tutte le persone sono mortali, che Socrate è un uomo e che un uomo è una persona. Meccanismi di inferenza logica possono permettere ad un sistema di AI di dedurre automaticamente che Socrate è mortale, come caso particolare della regola generale.

Un modello può rappresentare anche un ambiente dinamico, ad esempio le conseguenze di un piano, ossia una serie di azioni che può compiere un robot per muoversi e trasportare oggetti in un certo ambiente. Dato al robot un certo obiettivo (in pianificazione automatica chiamato "goal"), un modello dell'ambiente in cui il robot può operare, gli effetti delle sue azioni, un sistema di AI può generare automaticamente un piano, ovvero

[4] Stuart J. Russel, Peter Norvig, *Artificial Intelligence: A Modern Approach*, Harlow, 2020.
[5] F. van Harmelen, V. Lifschitz, B. Porter (a cura di), *Handbook of Knowledge Representation*, Amsterdam, 2008, 841-867.

DPCE online
ISSN: 2037-6677

una sequenza di azioni per raggiungere l'obiettivo. Queste tecniche sono relative agli studi in pianificazione automatica (*Automated Planning*).[6]

Dato un modello, esistono tecniche per verificare se il modello ha determinate proprietà. Ad esempio, si può rappresentare con un modello il funzionamento di un robot e verificare che non possa mai compiere un'azione non voluta in una certa situazione perché tale azione può essere pericolosa per le persone che operano in quell'ambiente. Tecniche come quelle denominate di *Model Cheching*[7] sono in grado di verificare se una proprietà è garantita da un modello. E se non è garantita, possono estrarre dal modello un "controesempio", ovvero un esempio che non rispetta la proprietà desiderata.

Il vantaggio di questi approcci è la possibilità di dare una spiegazione (*Explainability*) e l'affidabilità (*Trustworthiness*). Ad esempio, dato un modello, esistono tecniche formali per garantire che il modello soddisfi una certa proprietà. Viceversa, si può spiegare attraverso un contro esempio perché una caratteristica desiderata non è garantita. Una rappresentazione logica ci può dire qual è la causa di un certo fenomeno: ad esempio, Socrate è mortale perché è un uomo e tutti gli uomini sono mortali.

Lo svantaggio dell'approccio basato sulla modellazione sta nell'*effort* necessario nel costruire modelli di fenomeni complicati. Rappresentare tutta la conoscenza che abbiamo del mondo è praticamente impossibile. Nonostante questo svantaggio, le tecniche basate sulla modellazione sono state di enorme successo, si vedano ad esempio le applicazioni basate su tecniche di pianificazione automatica realizzate dalla NASA.[8]

5. L'approccio basato sull'Apprendimento: Machine Learning AI

Nell'approccio basato su apprendimento automatico il modello di un fenomeno viene ottenuto dai dati, ad esempio dai dati disponibili su web, dai numerosi sensori nelle nostre città, dai sensori indossabili (si pensi ai dati raccolti sulla nostra salute dai sensori presenti in un semplice smart watch).

Nell'approccio cosiddetto di "apprendimento supervisionato" (*supervised learning*), una specifica tecnica di apprendimento automatico (*machine learning*), i dati servono ad "addestrare" il modello prima di poterlo utilizzare. Questa è la fase chiamata di *"training"* (si veda Figura 3). I dati fungono da esempi. Per ogni esempio si dice alla macchina come deve comportarsi, quale risultato deve dare. Ad esempio, per far sì che un computer impari a riconoscere in una foto di un animale di che animale si

[6] M. Ghallab, D. Nau, P. Traverso, *Automated Planning: Theory and Practice*, Burlington, 2004; M. Ghallab, D. Nau, P. Traverso, *Automated Planning and Acting*, Cambridge, 2016.
[7] E. M. Clarke, T. A. Henzinger, H. Veith, R. Bloem, *Handbook of Model Checking*, Cham, 2018.
[8] N. Muscettola, P. P. Nayak, B. Pell, B. C. Williams, *Remote Agent: To Boldly Go Where No AI System Has Gone Before*, in *Artificial Intelligence*, 1-2, 1998, 5-47.

tratta, si danno al computer tante foto di diversi animali (ad esempio giraffe, cavalli, zebre) indicando al computer per ogni foto di che animale si tratta. Il computer riesce così a riconoscere le similitudini fra tutte le foto che rappresentano un tipo di animale e le differenze con quelle di un diverso animale, in modo che, se poi viene data al computer una nuova foto, mai vista prima, riconosce le similitudini con il tipo di animale giusto. I dati sono quindi il "carburante" del *machine learning*. La macchina viene addestrata, non programmata. Non c'è la fase di modellazione di un fenomeno da parte di un esperto. Per far questo si possono usare diverse tecniche. Una tecnica molto utilizzata è quella delle reti neurali (*neural metwork*s).

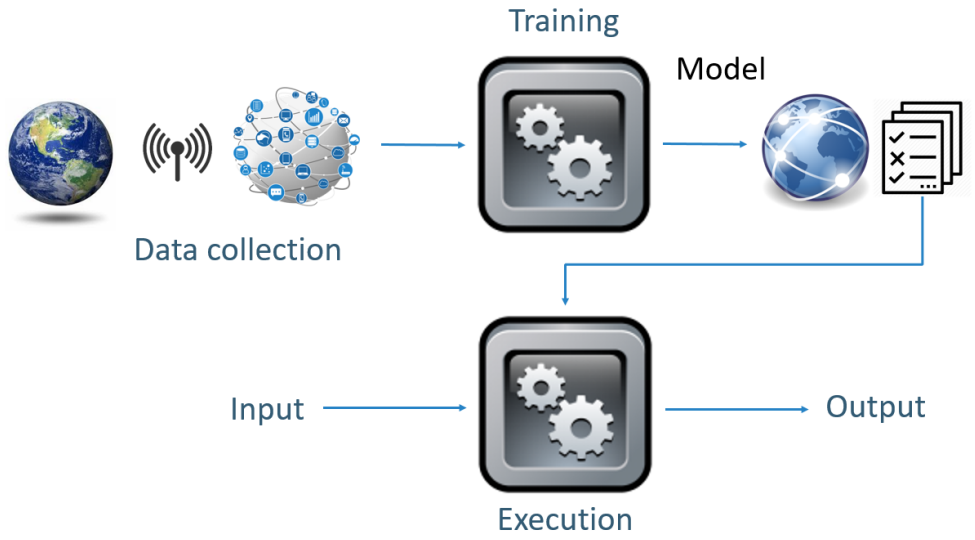

Figura 3: Machine Learning AI

L'ormai famoso "apprendimento profondo" ("*deep learning*")[9] è un tipo particolare di machine learning dove la rete neurale ha più strati. In questo caso "profondo" infatti vuol dire semplicemente "a più strati". La rete neurale riceve in ingresso i dati che forniamo al calcolatore (ad esempio la foto di un animale). Il primo strato converte i dati di ingresso (i pixel della foto) e li passa al secondo strato e così via, fino all'ultimo strato che determina l'output, il risultato che mi dà la rete neurale (ad esempio il fatto che si tratta di una foto di una giraffa). Ad ogni strato la rete neurale viene addestrata con dei valori particolari in modo che mi dia la risposta giusta. In Figura 4 è rappresentata graficamente una rete neurale a più strati. L'ingresso (i pixel della foto) sono i pallini rossi, le uscite – *output layer* – sono rappresentati dai pallini blu (la risposta che si tratta di una giraffa piuttosto di un altro animale), e gli strati intermedi – *hidden layer* – sono i pallini gialli che collegano uno strato a quello successivo. Possiamo immaginare i pallini gialli come le manopole di una radio, che regoliamo fino a che troviamo la

[9] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, Boston, 2016.

DPCE online
ISSN: 2037-6677

frequenza giusta. Regolandoli sempre meglio, la macchina imparerà e ci darà i risultati che le chiediamo.

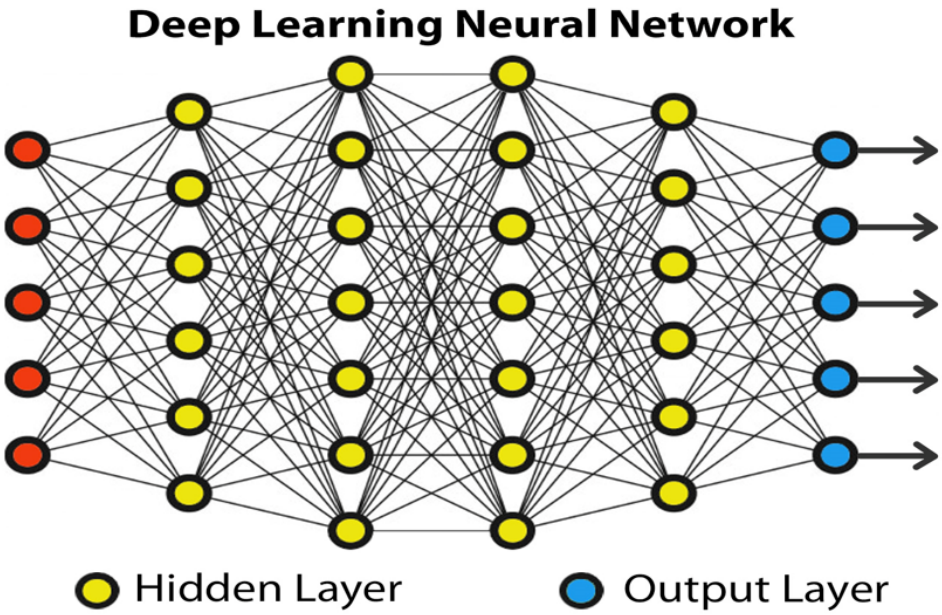

Figura 4: Una Rete Neurale a più strati per il deep learning

I legami fra i pallini di uno strato e l'altro rappresentano graficamente una funzione parametrica, ovvero una funzione con dei parametri che inizialmente non sono conosciuti. L'addestramento – la fase di training - serve appunto a determinare, a dare dei valori numerici a questi parametri. Il meccanismo tipicamente utilizzato durante il training è un meccanismo di propagazione all'indietro (*backpropagation*). Nell'output layer, il pallino blu che corrisponde alla risposta "giraffa" deve avere ad esempio un valore (vicino ad) uno quando ho dato in ingresso alla rete la foto di una giraffa, e gli altri pallini blu dovranno avere un valore (vicino a) zero. Durante l'addestramento, forzo il numero uno nel pallino blu corrispondente alla giraffa e zero negli altri. Il meccanismo di backpropagation permette di assegnare dei valori ai parametri delle funzioni che collegano lo strato dei pallini gialli allo strato dei pallini blu in modo che succeda proprio così, ovvero il risultato sia (vicino al valore) uno nel pallino blu della giraffa e (prossimo allo) zero negli altri. E così via fino a determinare il valore di tutti i parametri di tutte le funzioni parametriche che collegano uno strato a quello immediatamente precedente.

Inoltre, nel deep learning, (alcune del)le funzioni corrispondenti ad ogni pallino giallo sono funzioni non lineari, altrimenti i vari strati sarebbero comprimibili e semplificabili in uno strato unico. Quindi, semplificando, possiamo definire una rete neurale utilizzata nel deep learning come una

funzione composta da funzioni parametriche non lineari.

Alcune considerazioni sono necessarie a questo punto. Naturalmente il riconoscimento di immagini non è l'unico task con cui si può utilizzare il metodo basato su apprendimento a partire dai dati. Questo metodo viene al giorno d'oggi utilizzato con impressionante successo per il riconoscimento di scene nei filmati, l'analisi dei testi scritti in linguaggio naturale, il riconoscimento del parlato e dei suoni. In tutti questi casi l'approccio basato sull'apprendimento automatico si è dimostrato nettamente superiore all'approccio basato sulla modellazione nel caso in cui siano a disposizione abbastanza dati da effettuare un addestramento adeguato.

Inoltre, esistono delle tecniche basate sull'apprendimento automatico che affrontano problemi diversi rispetto a quelli del riconoscimento e della classificazione. Ad esempio, tecniche di "*Generative Adversarial Networks* (*GAN*)"[10] sono in grado di addestrare reti neurali per generare piuttosto che classificare immagini, video, suoni, musica, opere d'arte come quadri venduti a centinaia di migliaia di dollari. Inoltre tecniche di apprendimento sono molto utilizzate in robotica, ad esempio per imparare cosa conviene fare per raggiungere un determinato scopo, imparare a valle delle proprie azioni. Queste tecniche sono dette di *Reinforcement Learning*.[11]

A questo punto, anche per l'approccio basato sull'apprendimento automatico, dovrebbero risultare chiari i vantaggi e gli svantaggi. Mentre nell'approccio model based è necessario con fatica costruire un modello spesso complicato, nell'approccio basato sull' apprendimento automatico il modello è ottenuto in modo semplice e naturale. Nel nostro semplice esempio, per addestrare una rete neurale basta "taggare" delle foto, ovvero dire per ogni foto di che animale si tratta. Inoltre mentre in certi casi è possibile (anche se faticoso) definire il modello di un fenomeno, ad esempio di una macchina, di un sistema produttivo, del sistema di controllo di un aereo, è invece impossibile in pratica definire un modello che mi distingua in modo efficace cosa appare in una foto (si pensi alle mille posizioni in cui può essere fotografata una giraffa).

D'altro canto, dovrebbero essere chiari anche gli svantaggi dell'approccio basato su machine learning. Nel caso di apprendimento supervisionato (quello al momento di maggior successo), devo avere tanti dati a disposizione per addestrare bene il modello. Se l'addestro con poche foto di una giraffa, basterà una foto un po' diversa, in una posizione ad esempio diversa per non riconoscerla. Ma soprattutto, ci sono casi in cui non ci sono a disposizione tanti esempi per addestrare bene il modello. Inoltre, mentre in certi casi può essere anche semplice "taggare" delle foto, in altri è un processo dispendioso e difficile che può richiedere l'intervento di un esperto.

Al di là di questi problemi di ordine generale, dedichiamo il paragrafo

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, *Generative adversarial networks*, in *Commun. ACM*, 11, 2020, 139-144.
[11] R. S. Sutton, A. G. Barto, *Reinforcement Learning - an Introduction*, Boston,1998.

DPCE online
ISSN: 2037-6677

successivo ad alcune caratteristiche dell'approccio basato sull'apprendimento automatico che possono avere implicazioni nel campo del diritto e delle normative.

6. Luci e Ombre dell'Approccio basato sull'Apprendimento

L'apprendimento automatico, specialmente il deep learning supervisionato, ha avuto recentemente un enorme successo. È indubbio infatti che il deep learning ha permesso di ottenere risultati che fino a pochi anni fa si ritenevano irraggiungibili con le tecniche basate sui modelli o altre tecniche di apprendimento. Questo successo, che ha contribuito fortemente a rendere sempre più famosa l'Intelligenza Artificiale anche ai non addetti ai lavori, è dovuto, oltre che a nuove teorie e tecniche sviluppate recentemente in questo settore, anche alla disponibilità di grandi quantità di dati per l'addestramento e al cresciuto potere computazionale, che permette di addestrare e far funzionare enormi reti neurali, con tantissimi parametri che possono quindi, ad esempio nel caso della classificazione, distinguere in modo molto fine e preciso tantissimi diversi casi. Infatti, gran parte del successo del machine learning è dovuto all'enorme volume di dati oggi continuamente disponibili grazie alla digitalizzazione delle informazioni e allo sviluppo di Internet. Il cosiddetto fenomeno dei "*big data*". Pensiamo a tutte le foto, i video, le immagini prodotte da telecamere e webcam disponibili in internet. Ma anche ai dati sulla nostra salute, alle radiografie, ai dati relativi alle pubbliche amministrazioni, il meteo, le rilevazioni dei satelliti, quelli prodotti dai sensori nelle nostre città.

Ma nonostante i successi ottenuti, soprattutto in compiti che fino a tempo fa sembravano poter essere affrontati solo dall'essere umano (come riconoscere oggetti nelle foto, comprendere cosa sta accadendo in un video, etc.), queste tecniche presentano alcuni problemi che possono avere una rilevanza dal punto di vista del diritto, della normativa, delle considerazioni giuridiche.

- L'effetto **Scatola Nera** (*Black Box*) e il problema della mancanza di spiegazione (*lack of explainability).* Una rete neurale non è tipicamente in grado di spiegare le sue decisioni. Rifacendoci al nostro semplice esempio del riconoscimento delle foto di animali, una rete neurale non sa spiegare ad esempio che nella foto c'è una giraffa perché la giraffa ha il collo lungo, o c'è una zebra perché l'animale ha le strisce bianche e nere. Per la rete neurale, la differenza fra la foto di una giraffa e quella di un altro animale è semplicemente un insieme di valori diversi dei tanti parametri delle funzioni parametriche della rete neurale (i valori dei puntini gialli in Figura 4).

- I **Pregiudizi** (***Bias***) causati dai dati di addestramento. Il set di dati per l'addestramento può essere parziale rispetto alla realtà e indurre a conclusioni parziali, non neutrali e oggettive. Se ho un insieme di dati dove tutti gli studenti di una città sono bravi mentre quelli di un'altra

città sono meno performanti, la rete neurale tenderà a concludere che uno studente della prima città è tipicamente un genio e al contrario uno studente della seconda città è meno bravo. Famoso è il caso del sistema utilizzato negli Stati Uniti per predire la pericolosità di una persona che aveva avuto problemi con la giustizia. Chissà perché le persone di colore nero erano giudicate tipicamente più pericolose delle persone di colore bianco.

- La **mancanza del principio di causalità.** Il machine learning è in grado di riconoscere le relazioni fra i dati, ma non la causalità, ovvero qual è la causa e quale l'effetto. Una rete neurale, anche se ben addestrata, è in grado di capire che tipicamente col sole la temperatura aumenta, ma non sa dirti se è il sole che causa l'aumento di temperatura o la temperatura alta che causa la presenza del sole.

- **La possibilità di errore.** Anche questa tecnologia, come tutte le tecnologie, non è infallibile. Per di più, un intervento malizioso può facilmente trarre in inganno una rete neurale. Basti pensare all'esempio in cui una rete neurale è addestrata per riconoscere segnali stradali. È possibile modificare la foto di un segnale di stop in pochi pixel, anche in modo impercettibile all'occhio umano, per indurre la rete neurale in errore e classificare la foto come quella di un segnale di limite di velocità. Questo è dovuto al fatto che le reti neurali, come abbiamo visto, classificano in base a diversi valori numerici assegnati ai parametri delle funzioni parametriche. In certi casi i valori fra una classe e l'altra sono vicini e bastano piccole modifiche per far cambiare la decisione.

- La possibile **Violazione della Privacy.** Molti dei dati utilizzati al giorno d'oggi da grandi player che operano sul mercato, dalle famose multinazionali, raccolgono dati con un semplice consenso non propriamente informato da parte delle persone. Pensiamo solo a tutte le volte che interagiamo sui social, che cerchiamo qualcosa sui motori di ricerca, che acquistiamo un prodotto on line, a tutti i nostri clic e i nostri *"like"*, ai nostri smartphone interconnessi. Questi dati offrono un materiale enorme sui nostri gusti, i nostri interessi, i nostri acquisti, ma anche le nostre emozioni. Profilare in questo modo le persone senza un loro vero consenso informato non credo che sia rispettoso dei diritti umani.

La ricerca in AI sta affrontando questi problemi con diversi approcci, fra i quali ricordiamo, ad esempio, la ricerca in tecniche di apprendimento non supervisionato o semi-supervisionato, ovvero senza la necessità di "taggare" una enorme quantità di dati, il cosiddetto *Transfer Learning*, ovvero la capacità di utilizzare un sistema addestrato per un determinato compito anche per compiti diversi senza la necessità di ri-addestrarlo completamente, le tecniche di *Zero-shot (One-shot) Learning*, ovvero la

DPCE online
ISSN: 2037-6677

capacità di apprendere da pochi dati.[12]

Una menzione a parte va fatta per la ricerca in *"Integrative AI"*,[13] ovvero nel tentativo di unire e far convergere i due approcci diversi che abbiamo brevemente illustrato in questa sezione, il model based e il machine learning AI. L'idea alla base dell'Integrative AI è che i modelli non vengano utilizzati per descrivere in modo completo un fenomeno, ma soltanto alcune caratteristiche che servano a influenzare le decisioni ad esempio di una rete neurale. Ad esempio, rifacendoci al semplice caso del riconoscimento delle foto degli animali, non si vuole descrivere le proprietà di un elefante per riconoscerlo in una foto, ma il fatto che l'elefante non vola e quindi correggere quando la rete neurale tenderebbe a riconoscere un elefante … "sospeso per aria". Ma l'integrative AI può anche essere utilizzata per contraddire il modello, ad esempio quando la rete neurale continua a riconoscere un elefante che vola perché in realtà si tratta del famoso cartone animato dove Dumbo è invece in grado di volare … con l'aiuto di una piuma.

7. Conclusioni

In conclusione, l'intelligenza artificiale ha enormi potenzialità, ma come tutte le tecnologie con grandi potenzialità va compresa, gestita e naturalmente anche normata.

Dobbiamo essere coscienti e evitare che si avverino i rischi dei possibili utilizzi dell'AI, rischi menzionati brevemente nella sezione precedente, quali i rischi di pregiudizi nelle decisioni e nei giudizi (i famosi "bias"), rischi di affidarsi a macchine che fanno errori, i rischi di una intelligenza artificiale che ci profila per capire i nostri gusti, influenzarci (addirittura nel voto come Cambridge Analitica insegna), o il rischio che venga meno il diritto delle persone alla riservatezza, alla privacy e, più in generale, i rischi addirittura di andare contro diritti vitali per l'umanità, quali i diritti sanciti dalla costituzione.

Le norme e le regolamentazioni possono aiutarci in questo. Ma dobbiamo evitare anche il rischio opposto. La regolamentazione non deve impedire di utilizzare l'AI dove può fare del bene, dove può essere veramente utile per l'umanità. Non dobbiamo astenerci dall'utilizzo dell'AI in campi delicati ma importanti come la medicina, ad esempio. In questa direzione penso che vada il Regolamento Europeo sull'Intelligenza Artificiale,[14] anche se, come evidenziato molto chiaramente nel lavoro di Casonato e

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, *Generative adversarial networks*, cit., 139-144.

[13] Si veda ad esempio L. Serafini, *Learning and Reasoning with Logic Tensor Networks: the Framework and an Application*, in M. Homola, V. Ryzhikov, R. Schmidt (a cura di), *Proceedings of the 34th International Workshop on Description Logics*, 2021, in ceur-ws.org/Vol-2954/; S. Patra, J Mason, M. Ghallab, D. S. Nau, P. Traverso, *Deliberative acting, planning and learning with hierarchical operational models*, in *Artificial Intelligence*, 299, 2021, 103523.

[14] Proposal for a Regulation of the European Union on AI, report of the European Commission and European Parliament.

Marchetti,[15] c'è ancora tanto lavoro da fare.

Ma c'è tanto lavoro da fare anche per chi fa ricerca e sviluppo del campo dell'intelligenza artificiale. I ricercatori stessi devono guardare alle regole non come ad un vincolo che impedisce di sviluppare e utilizzare tecniche innovative di AI. Al contrario, queste vanno viste come una grande opportunità per sviluppare una intelligenza artificiale migliore, il più efficacie possibile ma al tempo stesso il più affidabile possibile e rispettosa dei diritti umani. Credo che anche da parte dei ricercatori, degli sviluppatori e degli utilizzatori dell'AI l'approccio vada, per così dire, ribaltato, per sfruttare appieno le potenzialità dell'Intelligenza Artificiale: il Regolamento Europeo non va visto come un vincolo, ma come strumento che spinga verso una Intelligenza Artificiale migliore.

Credo che combattere e evitare che i rischi di un "utilizzo sbagliato" dell'AI diventino realtà dipenda fortemente da noi tutti, ricercatori, imprenditori, giuristi, da chi lavora nelle istituzioni, dai *policy maker*, dai decisori istituzionali. Dobbiamo tutti lavorare assieme per evitare questi rischi e far sì che la futura AI sia veramente un AI per il bene delle persone[16].

Paolo Traverso
Fondazione Bruno Kessler
traverso@fbk.eu

[15] C. Casonato, B. Marchetti, *Prime Osservazioni sulla Proposta di regolamento della Unione Europea in Materia di Intelligenza Artifciale*, in *BioLaw Journal – Rivista BioDiritto*, 3, 2021, 1-23.

[16] Si veda ad esempio B. Braunschweig, M. Ghallab, *Reflections on Artificial Intelligence for Humanity. Lecture Notes*, Cham, 2021; A. Dengel, O. Etzioni, N. DeCario, H. H. Hoos, L. Fei-Fei, J. Tsujii, P. Traverso, *Next Big Challenges in Core AI Technology. Reflections on Artificial Intelligence for Humanity*, Cham, 2021, 90-115.