# *2. Caratteristiche, prospettive e problematicità dell'Intelligenza Artificiale*

di *Diego Latella, Gian Piero Siroli, Guglielmo Tamburrini*

## 2.1. Introduzione

La IA nasce ufficialmente nell'estate del 1956 quando John McCarthy, uno dei padri di quella che lui stesso avrebbe chiamato "Artificial Intelligence", all'epoca giovane assistente alla cattedra di matematica al Dartmouth College ad Hanover (Stati Uniti), organizzò il Dartmouth Summer Research Project on Artificial Intelligence, un workshop al quale parteciparono molti pionieri della IA, come Marvin Minsky, Herbert Simon e Allen Newell, e altri eminenti scienziati come Claude Shannon e Nathaniel Rochester[1]. L'obiettivo del workshop era quello di chiarire e sviluppare concetti e idee sulle cosiddette "thinking machines".

A tutt'oggi, però, non si è pervenuti a una definizione soddisfacente di IA; in questa sede utilizziamo quella – in un certo senso riduttiva – proposta dall'UNIDIR, secondo cui: «[l']intelligenza artificiale è il campo di studi dedicato a rendere intelligenti le macchine. L'intelligenza misura la capacità di un sistema di determinare la migliore linea d'azione per raggiungere i propri obiettivi in una vasta gamma di ambienti» (UNIDIR, 2018, p. 2, trad. it. a cura degli autori).

In pratica, molti ricercatori nel campo della IA, piuttosto che affrontare il problema di una definizione teorica completa della disciplina, preferiscono darne una caratterizzazione ottenuta indirettamente, attraverso gli sviluppi tecnologici della stessa. È così che l'avanzamento delle conoscenze di IA avviene realizzando delle macchine capaci di risolvere autonomamente specifici problemi la cui soluzione necessita di una certa dose di "intelligenza". Così, nel 1952, un computer è riuscito a vincere a tris, nel 1994 fu il turno della dama, quindi degli scacchi nel 1997, fino allo sviluppo, nel 2014, di

[1] Si veda, ad esempio (Franklin, 2014).



 ISBN 9788835152224

programmi capaci di giocare e vincere sugli umani a vari giochi quali Atari, Go (2016) e poker (2017).

Un effetto secondario di questa tendenza è che solo le macchine più all'avanguardia vengono considerate "intelligenti": ad esempio, negli anni '60 venivano considerate intelligenti le macchine capaci di giocare a scacchi, mentre oggi queste macchine vengono considerate dei semplici programmi informatici più o meno efficienti. Con il passare degli anni, la "intelligenza" delle macchine si è spostata venendo associata a compiti che noi riteniamo più complessi o creativi, come, per esempio, il gioco del Go. In altri termini, ciò che accade è che la capacità di risolvere un certo problema viene considerata espressione di intelligenza solo fino a quando non viene realizzata una macchina capace di trovare una risposta.

In ogni caso, le tecnologie della IA, e in particolare di quella branca della IA che va sotto il nome di *machine learning*, sono state applicate con notevole successo in vari domini, che vanno dal riconoscimento di immagini, della voce e del linguaggio naturale, alla traduzione automatica fra lingue (umane), all'aggregazione e analisi di immense quantità di dati (*data analytics* e *data science*) con fini predittivi o di diagnosi e al controllo di sistemi autonomi, come i veicoli a guida autonoma. Infine, sistemi di IA hanno sempre suscitato particolare interesse in ambienti militari (Din, 1986; Andride, 1987) e le applicazioni della IA in questo contesto vanno dai sistemi C3IR (Command, Control, Communication, Intelligence and Reconnaissance), a quelli di supporto alle decisioni, fino ad arrivare alle armi autonome[2].

Accanto agli aspetti puramente tecnologici citati fino adesso e che costituiscono argomento di primario interesse nel contesto di questo libro, la IA come disciplina squisitamente scientifica ha visto e vede tuttora sviluppi concettuali e teorici di altissimo livello. Tra questi si annoverano la creazione e lo studio di modelli teorici per la rappresentazione della conoscenza, sia per la pianificazione (*planning*) e l'apprendimento (sul quale ci soffermeremo più in dettaglio più avanti) sia per la *explainability* – intesa come la "capacità di fornire spiegazione" – dei procedimenti computazionali e logici che portano ai risultati generati da sistemi di IA, così come sistemi logico-deduttivi per il ragionamento automatico, che fondano le proprie radici, anche storiche, nella logica matematica e nella teoria della calcolabilità. Lo studio di questi ultimi ha permesso di sviluppare, ad esempio, dimostratori automatici di teoremi. Molte classi di logiche sviluppate o utilizzate nell'ambito della ricerca di base sulla IA, come le logiche epistemiche (van

[2] Si veda, fra gli altri: (USDOD, 2016), (Allen e Chan, 2017), (Boulanin e Verbruggen, 2017), (Cummings, 2017), (Dyndal *et al.*, 2017), (Amoroso *et al.*, 2018), (CRS, 2018), (UNIDIR, 2018), (Rossi, 2019).



 ISBN 9788835152224

Ditmarsch *et al.*, 2015), temporali (Emerson, 1990), o spaziali (Aiello *et al.*, 2007) sono specializzazioni della logica modale (van Benthem e Blackburn, 2006), un campo di proficua ricerca nell'ambito della logica matematica.

Infine, è necessario citare lo studio e lo sviluppo della cosiddetta *swarm intelligence*, un paradigma utilizzato anche in ambito IA, nel quale vengono sviluppati algoritmi ispirati dal comportamento di popolazioni di agenti biologici, come colonie di insetti, stormi di uccelli o banchi di pesci. In questo caso, ogni singolo agente segue regole comportamentali e di interazione estremamente semplici che, però, danno origine a interessanti proprietà emergenti della popolazione al punto tale che quest'ultima, nell'insieme, può risolvere in maniera estremamente efficace ed efficiente problemi di notevole complessità.[3] Un'interessante area di ricerca e sviluppo collegata anche alla *swarm intelligence* è quella della *swarm robotics*, nell'ambito della quale si studiano e sviluppano "sciami" di sistemi robotici, i quali hanno la capacità di coordinare le loro azioni per operare collettivamente per il raggiungimento di un obiettivo condiviso. Ogni individuo dello sciame è pensato e realizzato come entità autonoma, che reagisce ai vari stimoli in base a sue regole interne. Lavorando come un gruppo, lo sciame può eseguire compiti sia semplici che complessi, che un singolo robot non sarebbe in grado di svolgere (Ekelof e Persi Paoli, 2020).

## 2.2. *Machine Learning*

Fra le tecniche sviluppate nell'ambito della IA per scopi civili, quelle di *Machine Learning* (ML), su cui ci concentreremo nel resto del capitolo, hanno riscosso particolare successo negli ultimi anni. Questo successo è in parte dovuto alla incredibile quantità di dati disponibili su svariati aspetti del problema che, di volta in volta, si vuole affrontare e alla grande capacità di calcolo oggi realizzabile. La disponibilità di dati è, a sua volta, conseguenza della diffusione di sensori accessibili a basso costo, che possono monitorare praticamente qualunque aspetto fisico e sociale del pianeta e che vanno dai sensori ad hoc dedicati al monitoraggio di particolari fenomeni fisici, ai dispositivi elettronici di uso comune come gli smartphone, i tablet e, con l'avvento dell'*Internet of Things* (IoT)[4], tutti i più comuni elettrodomestici, i mezzi di trasporto e, più in generale, le infrastrutture (regionali, nazionali e urbane).

Il ML rappresenta un paradigma duale rispetto all'informatica tradizionale,

[3] Si veda, ad esempio, la rivista scientifica specializzata *Swarm Intelligence*, Dorigo, M. (Ed. in Chief), Springer, Berlin, D.

[4] Si veda, ad esempio (USGAO, 2017a); per le questioni di sicurezza sollevate dall'avvento dell'Internet of Things si suggerisce (Schneier, 2018).

ma anche alla IA "tradizionale" (a volte indicata dall'acronimo "GOFAI": *Good Old-Fashioned Artificial Intelligence*). Entrambe queste ultime, per definizione hanno come obiettivo la soluzione di problemi che, per loro stessa natura, risultano difficili per gli umani, ma sono di facile automazione e quindi relativamente semplici per le macchine, ad esempio perché ripetitivi e basati su regole di inferenza logica. Viceversa, il ML è una disciplina che cerca di affrontare e risolvere quei problemi che, invece, non sono facilmente descrivibili in termini di regole logiche, né necessariamente ripetitivi e che, spesso, risultano "semplici" per gli umani, come, ad esempio, riconoscere gli oggetti che compongono una scena della vita comune (Goodfellow *et al.*, 2016).

In altri termini, mentre le attività cognitive di alto livello, come ad esempio il ragionamento, possono essere, entro certi limiti, ricreate artificialmente utilizzando le tecniche tipiche della GOFAI, quelle di livello più basso, più assimilabili all'apprendimento umano dei primi anni di vita, sono più facilmente simulabili utilizzando tecniche di ML, a patto che si abbiano dati di alta qualità e potenza di calcolo sufficienti.

Il ML è una disciplina che vede integrati aspetti e risultati scientifici dell'informatica, della statistica e dell'algebra lineare, con intuizioni provenienti da altre discipline, come le neuroscienze. La caratteristica fondamentale delle tecniche di ML è che, a differenza delle altre tecniche tipiche dell'informatica, esse affrontano il problema di programmare i computer a "imparare" partendo dai dati e dall'esperienza (Buchanan e Miller, 2017).

Gli utilizzi principali del ML includono quelli di classificazione, quelli di strutturazione di dati originariamente non strutturati e di riconoscimento di pattern. Fra i principali domini di applicazione vanno ricordati: la *computer vision*, cioè la capacità di un computer di riconoscere e identificare specifici oggetti in una immagine; l'elaborazione del linguaggio naturale, come la traduzione automatica da una lingua ad un'altra; la IoT, dove i dispositivi di uso comune possono apprendere le abitudini e le limitazioni degli utilizzatori facilitandone quindi l'uso; il supporto al design e alla ricerca scientifica, dove sistemi artificiali di supporto intelligenti possono (aiutare a) ideare soluzioni innovative; i sistemi medicali, dove sottosistemi di IA – per esempio sistemi di riconoscimento di immagini – possono fornire un valido supporto sia in fase di diagnosi che in fase di definizione delle terapie (come nella segmentazione propedeutica alla radioterapia); i sistemi di trasporto, sia per quanto riguarda gli aspetti di controllo del traffico che in relazione allo sviluppo di veicoli a guida autonoma; i sistemi di supporto ai procedimenti giudiziari, come quelli di polizia predittiva; i sistemi militari, sia come (sottosistemi di) sistemi di supporto alle decisioni, sia come componenti di armi autonome.

Le fasi di una tipica procedura di ML sono quattro:



 ISBN 9788835152224

1. *programmazione*, cioè la definizione, progettazione e implementazione di una procedura software[5] con la quale si istruisce un computer a "imparare" a partire da un insieme di dati (*training data set*);
2. *allenamento* (training), che consiste nell'esecuzione del programma di cui al punto precedente su un *training data set* al fine di impostare e calibrare i (molti) parametri presenti nel programma; il *training data set* è costituito da dati che possono – ma non necessariamente devono – essere "etichettati", come vedremo più avanti;
3. *testing* del programma calibrato al punto precedente, e cioè esecuzione controllata dello stesso, su un altro insieme di dati – il *testing data set* – e relativa analisi dei risultati al fine di valutarne la qualità;
4. Infine, se la fase precedente viene completata con una valutazione positiva, l'*uso* del programma calibrato su nuovi dati, diversi da quelli di training e da quelli di testing.

Se la terza fase non si conclude positivamente, si dovrà invece procedere a ulteriori fasi di training ed eventualmente rivedere anche le scelte di progettazione o di acquisizione adottate nella fase di programmazione. Va notato subito che, a seconda del tipo di approccio al ML che si è scelto, è possibile che attività di training abbiano luogo anche durante l'uso del programma.

Si possono identificare vari tipi di ML, fra i quali il *supervised ML*, l'*unsupervised ML*, il *reinforcement learning* e il *deep learning*. Caratteristica principale del *supervised ML* è il fatto che il *training data set* è costituito da dati etichettati, ovvero ogni dato nel data set è accompagnato da un'"etichetta" che lo descrive e che, di norma, è creata manualmente. Facciamo un semplice esempio: uno degli usi del ML è quello del riconoscimento di testi scritti a mano (e la loro trascrizione in testo basato su un insieme di caratteri standard). Se, per semplicità, ci limitiamo al riconoscimento di cifre numeriche da 0 a 9 scritte a mano, possiamo immaginare il training data set come una grande tabella, ogni riga della quale è costituita da due campi: uno contiene un'immagine di una cifra manoscritta e l'altro l'etichetta, cioè la sua trascrizione; in figura 1 riportiamo una minima parte di una simile tabella, con due immagini diverse del numero 3 scritto a mano, ma entrambe associate al carattere "3":

[5] Naturalmente, spesso è possibile sostituire questa fase con una di semplice "acquisizione" del software, sia dal mercato che attraverso l'uso di prodotti *freeware*.



 ISBN 9788835152224

*Fig. 1 – Un frammento di* training data set *per il riconoscimento delle cifre scritte a mano*

| Immagine | Etichetta |
|---|---|
|  | 3 |
|  | 3 |

Il *training data set* viene usato dal programma di ML per impostare e, soprattutto, calibrare i parametri caratteristici del particolare modello computazionale di ML selezionato per l'utilizzo nell'applicazione d'interesse. Esistono svariati modelli computazionali per il *ML supervised*, che vanno, solo per citarne alcuni, dalla semplice regressione lineare, alle *Support Vector Machines (SVM)*, alle reti neurali (Goodfellow *et al.*, 2016). Caratteristica comune a tutti questi modelli computazionali è il fatto di essere fortemente parametrizzati. Il "training" consiste nell'esecuzione di procedure, tipicamente iterative, per fissare i valori di questi parametri e calibrare questi ultimi per affinamenti successivi. La calibratura dei parametri deve evidentemente essere guidata da un qualche criterio che il programma di ML deve seguire; questo criterio è rappresentato tipicamente da una funzione di costo da minimizzare. Quest'ultima, in buona sostanza, rappresenta l'errore che, a seguito di una certa impostazione dei parametri, la macchina può commettere nel classificare o riconoscere i dati e che, quindi, va minimizzato. In ultima analisi, quindi, l'operazione di training altro non è che la soluzione di un problema di minimizzazione di una certa funzione matematica.

Implicito in quanto appena detto è il fatto che si abbiano rappresentazioni numeriche degli oggetti sui quali si fa ML. Ad esempio, nel caso di riconoscimento di oggetti in immagini, per semplicità in bianco e nero, la rappresentazione digitale dell'immagine è una matrice numerica bidimensionale di pixel[6], ciascuno dei quali è portatore di un valore numerico che identifica l'intensità luminosa del pixel stesso; questa matrice, a sua volta, può essere rappresentata come un vettore numerico – cioè una sequenza di numeri – di dimensione pari al numero totale di pixel dell'immagine. Solo a titolo di esempio, in figura 2 riportiamo una piccola matrice di 16 pixel (4 x 4) (nei sistemi reali, si hanno matrici con migliaia o milioni di pixel!) e la sua rappresentazione come vettore, dove le righe della matrice sono semplicemente giustapposte, una a fianco all'altra. In realtà, il vettore contiene dei numeri,

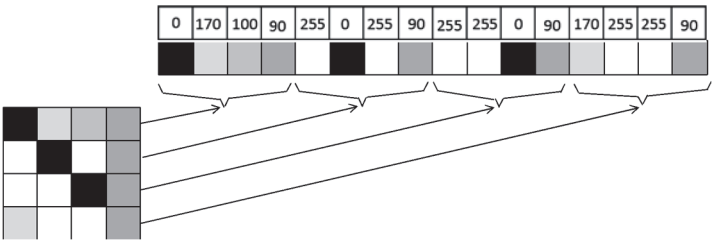
[6] Il pixel è l'elemento più piccolo indirizzabile e controllabile di una immagine rappresentata nello schermo di un computer.



 ISBN 9788835152224

che vanno da 0, corrispondente al nero (intensità minima), a 255, corrispondente al bianco (intensità massima), come rappresentato dal vettore numerico collocato nella parte superiore della figura.

*Fig. 2 – Esempio di rappresentazione numerica di immagini*




Quindi, un'immagine digitale è rappresentata da un vettore numerico. Esso rappresenta l'input del modello computazionale, il cui output può a sua volta essere un vettore, ad esempio con tanti elementi quanti sono gli oggetti diversi che si vogliono identificare, nel quale ogni elemento fornisce la probabilità che una data immagine in input contenga un certo specifico oggetto.

Il *ML unsupervised*, invece, viene usato tipicamente quando non si ha a disposizione dei training data set etichettati. In questo caso, sono gli algoritmi stessi che, ricevendo come input un (grande) insieme di dati non strutturato, cercano di identificare in esso dei pattern o una qualche struttura. Un esempio di questo tipo è il metodo "k-means clustering": dato uno spazio multidimensionale, l'algoritmo lo ripartisce in k sottoinsiemi, ciascuno rappresentato dal suo centroide[7], in modo tale che sia minimizzata la distanza fra ogni punto e il centroide del suo sottoinsieme di appartenenza (Buchanan e Miller 2017). La figura 3 mostra un esempio di applicazione dell'algoritmo di k-means, per k=3: in questo caso, vengono inizialmente scelti (di solito in modo casuale) 3 punti, rappresentati dai tre tondini in bianco, grigio chiaro e grigio scuro nella figura 3(a). L'algoritmo, procedendo in modo iterativo produce una partizione dell'insieme di tutti i punti secondo il criterio sopra indicato, mostrata in figura 3(b), dove ogni punto è colorato con lo stesso livello di grigio del centroide del suo insieme di appartenenza; come si può facilmente notare, ogni punto di un dato colore è più vicino al centroide dello stesso colore di quanto non lo sia ai centroidi di colore diverso.

Questa tecnica consente quindi di aggregare i dati di un certo insieme attorno ai k centroidi e quindi può risultare utile in quelle applicazioni che

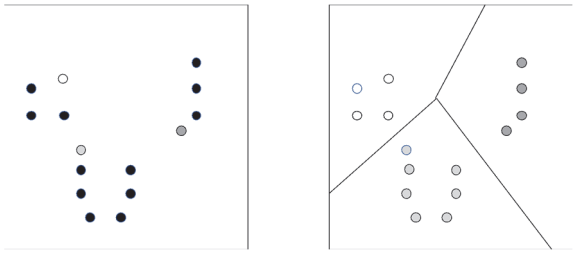
[7] Il centroide di un insieme di punti in uno spazio multidimensionale è il punto che rappresenta la *media* dei punti nell'insieme; in altri termini, ogni coordinata del centroide è la media dei valori della stessa coordinata dei punti dell'insieme.



 ISBN 9788835152224

richiedono classificazioni dei dati come, ad esempio, l'identificazione di cellule cancerogene in un campione di tessuto, il clustering di parole con significati simili, le analisi di mercato o addirittura l'identificazione di mine in un campo di battaglia (Landman *et al.*, 2019).

*Fig. 3 – Esempio di applicazione dell'algoritmo di k-means, con k=3*



(a) (b)

Infine, uno degli usi più comuni del *ML unsupervised* è quello di scoprire una qualche struttura in insiemi di dati per poter poi utilizzare la conoscenza acquisita sulla struttura nella progettazione di sistemi di *ML supervised*.

Nel caso del *reinforcement learning* un agente artificiale (un programma in esecuzione su uno o più computer) può "imparare" semplicemente interagendo con il suo ambiente. Tipicamente l'agente esegue un'azione, in maniera conforme a certe regole, ne "osserva" (tramite opportuni sensori) l'effetto sull'ambiente e quindi determina se l'azione è stata di aiuto per il raggiungimento degli obiettivi dell'agente. L'ambiente può anche essere costituito dall'agente stesso: è questo il metodo utilizzato da AlphaZero per imparare a giocare a Go e ad altri giochi (Buchanan e Miller, 2017; UNIDIR, 2018).

Infine, il *deep learning* può essere pensato, in termini generali, come un insieme di tecniche "architetturali" attraverso le quali vengono combinate varie tipologie di ML, sia *supervised* che *unsupervised*, per l'estrazione di caratteristiche rilevanti dei dati in input e per l'apprendimento attraverso livelli multipli di rappresentazione, che corrispondono a diversi livelli di astrazione e che quindi formano una gerarchia di concetti. Ad esempio, il *deep learning* può combinare un processo *unsupervised* per apprendere le caratteristiche dei dati sottostanti, come i bordi di un viso, e quindi fornire tali informazioni a un algoritmo di apprendimento *supervised* per riconoscere le caratteristiche e produrre il risultato finale, come identificare correttamente una persona in una foto (Buchanan e Miller, 2017).

## 2.3. Limiti e problematicità del *Machine Learning*

È innanzitutto opportuno premettere che i sistemi di IA, inclusi quindi quelli di ML, sono costituiti da programmi in esecuzione in computer progettati e costruiti con tecnologie sostanzialmente tradizionali[8]. Di conseguenza, i sistemi di IA sono vulnerabili a tutti gli attacchi cibernetici che sfruttano le vulnerabilità dei normali sistemi informatici, come è meglio descritto nel Capitolo 3 di questo libro.

Bisogna poi sottolineare che il comportamento di un sistema di ML (*supervised*) durante la fase di uso dipende fortemente dalla qualità dei dati e delle procedure di training. Da tale qualità dipende anche la possibilità di evitare che la macchina, durante l'uso, esibisca dei pregiudizi (*bias*).

La creazione di bias nella fase di training può dipendere dal fatto che il sistema di ML abbia eseguito un addestramento poco accurato, cioè su dati che non rappresentano bene la popolazione di oggetti sui quali la macchina viene usata; oppure dal fatto che il training sia stato anche molto accurato, ma svolto su dati a loro volta distorti (*biased*); oppure, semplicemente, dal fatto che il sistema non sia stato sufficientemente addestrato (Buchanan e Miller, 2017).

La presenza di *bias* è un problema particolarmente insidioso:

> Nel peggiore dei casi, l'apprendimento automatico può nascondere discriminazioni con l'imprimatur della scienza. Ad esempio [...] una facoltà di medicina britannica ha usato un algoritmo che scartava candidati qualificati di sesso femminile o provenienti da minoranze perché era stato addestrato sulle decisioni prese in precedenza da una commissione di valutazione non imparziale. Un'indagine di ProPublica ha rilevato che un algoritmo sviluppato dalla società Northpointe, Inc. per fornire un punteggio di valutazione del rischio [di supporto] per i giudici [nell'emissione di una] condanna era distorto rispetto alla razza e inaccurato. Gli afroamericani avevano molte più probabilità di essere etichettati come "ad alto rischio", ma [...] gli afroamericani che erano stati etichettati come "ad alto rischio" avevano [di fatto] molte meno probabilità di commettere un altro crimine rispetto ai bianchi "ad alto rischio" (Buchanan e Miller, 2017, p. 32, trad. it. a cura degli autori).

È quindi di estrema importanza poter disporre di grandi quantità di dati di alta qualità e di procedure di training altrettanto affidabili. Questo non sempre è possibile e, soprattutto, è particolarmente difficile per alcuni

[8] Va però detto che cominciano ad essere prodotti dei dispositivi hardware progettati e sviluppati appositamente per applicazioni di IA come il Tensor Processing Unit, un circuito integrato per reti neurali di Google (https://cloud.google.com/tpu/docs/tpus, accesso effettuato il 1 settembre 2022).

domini di applicazione; primi, fra questi, sono quelli collegati al campo di battaglia, soggetti, per definizione, alla *fog of war*.

È inoltre fondamentale sottolineare come dato un certo problema, per esempio, di classificazione, diversi algoritmi e modelli computazionali possono dare risultati abbastanza diversi; al riguardo, si rimanda a Buchanan e Miller (2017) e agli esempi mostrati nel sito scikit-learn[9] (Pedregosa *et al.*, 2011).

Se l'uso di tecniche di ML può dunque costituire un utile supporto per il decision-making (dando ad esempio delle indicazioni "di massima" su come classificare oggetti, scenari o situazioni di interesse), affidarsi a tali tecniche come unico strumento decisionale può risultare estremamente rischioso, per lo meno all'attuale stato dell'arte.

È inoltre importante sottolineare come ogni particolare modello computazionale ha i suoi punti di forza e i suoi punti deboli; di conseguenza i professionisti del ML devono spesso provare più di un modello/algoritmo per determinare quale, fra essi, risponda meglio alle esigenze poste dal problema da risolvere (Buchanan e Miller, 2017). La scelta di uno specifico modello/algoritmo di ML è tutta nelle mani degli esperti di ML e la sua adeguatezza dipende dalle loro specifiche competenze nell'ambito sia del ML che dominio di applicazione.

Anche in questo caso, quindi, risulta difficile pensare che sistemi di IA, costruiti secondo le conoscenze e le tecnologie all'attuale stato dell'arte, possano essere utilizzati per l'esecuzione di funzioni militari critiche, come decisioni autonome – cioè senza alcun significativo controllo umano – relative alla vita e la morte di avversari. Questo non solo per ragioni di natura etica e di diritto umanitario internazionale, come discusso nei Capitoli 5 e 9 del libro, ma anche sulla base di considerazioni tecniche.

Infine, nei sistemi di *ML supervised* (poiché, come si è visto, il training di un modello consiste nella calibrazione dei suoi parametri effettuata in maniera automatica, con iterazioni guidate da un'enorme quantità di dati nella fase di uso della macchina) risulta praticamente impossibile alla mente umana – incluse quelle degli stessi programmatori della macchina – comprendere perché, dato un certo input, essa produca uno specifico risultato: queste tecniche, quindi, producono sistemi di IA che vanno utilizzati come delle vere e proprie black box.

Questa caratteristica è particolarmente importante ed è anche molto critica per via del fatto che, purtroppo, i sistemi di ML a volte, ed in maniera non facilmente predicibile, producono risultati errati e totalmente inaspettati.

[9] https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html (accesso effettuato il 26 gennaio 2022).

Per esempio, sono stati effettuati degli esperimenti nell'area del riconoscimento automatico delle immagini in cui si è visto che leggere perturbazioni dei valori dei pixel dell'immagine di input, del tutto irrilevanti per gli umani, inducono la macchina a clamorosi errori, come quello di classificare l'immagine di uno scuolabus come se fosse quella di uno struzzo (Klarreich, 2016; Edwards, 2019; Open AI, 2019); analogamente, il sistema di riconoscimento può fallire facilmente se si agisce direttamente sugli oggetti da riconoscere nello spazio reale, del quale viene sottoposta al sistema un'immagine, per esempio una foto (Thys *et al.*, 2019).

È evidente che comportamenti e situazioni come quelli ai quali si è accennato sopra non possono essere tollerati per (sotto-)sistemi impiegati per lo svolgimento di funzioni critiche, come quelle di selezione e abbattimento del bersaglio da parte di future armi autonome.

Per applicazioni meno critiche, il problema del superamento del limite della *black box* costituisce oggi una vera e propria sub-disciplina della IA che va sotto il nome di "spiegabilità" (*explainability*) ed è attualmente oggetto di studio nel contesto di vari programmi di ricerca, come il programma *Explainable Artificial Intelligence* (XAI), della Defense Advanced Research Programs Agency (DARPA) statunitense (Gunning e Aha, 2019), e i progetti *Local Interpretable Model-Agnostic Explanations* (Ribeiro *et al.*, 2016) e Human-Centered Artificial Intelligence (Human AI, 2019). Trattandosi di ricerche ancora in corso è prematuro valutarne l'effettiva applicabilità sul campo dei loro risultati.

Va sottolineato che i sistemi di IA, e in particolare di ML, stanno dimostrando capacità e performance estremamente interessanti quando applicati a domini molto specifici e per l'esecuzione di compiti ben delimitati. Questo tipo di IA viene normalmente classificata come IA "stretta" (*narrow AI*). Ad esempio, nel caso delle reti neurali, si assiste al cosiddetto *catastrophic forgetting*: quando una rete cerca di apprendere nuove funzionalità o nuovi compiti, tipicamente dimentica quelli imparati in precedenza. Lo stato dell'arte in queste discipline non consente, al momento, di avere a disposizione sistemi di "IA generale", cioè sistemi realmente intelligenti, capaci, fra l'altro, di trasferire in altri domini conoscenze apprese in un certo dominio applicativo. A maggior ragione, una "super IA", cioè sistemi che esibiscano un'intelligenza superiore a quella naturale/umana in tutti i domini della conoscenza non è ad oggi realizzabile e si ritiene non lo sarà neppure nel prossimo futuro.

Concludiamo questa sezione ritornando brevemente sul problema delle vulnerabilità dei sistemi di IA e ML ad attacchi informatici specifici per questa classe di sistemi – genericamente denominati *adversarial attacks* – e rimandando al Capitolo 3 per una discussione più generale sui rischi di sicurezza informatica e cyberwar.



 ISBN 9788835152224

Un primo tipo di attacchi specifici per i sistemi di ML è quello dei cosiddetti *exploratory attacks*, tipici di avversari (anche umani) che si evolvono nel tempo, scoprendo, e quindi sfruttando, vari punti deboli dei programmi di ML. Ad esempio, un sistema automatico antispam potrebbe fare uso di un vocabolario per identificare gli spam; uno spammer può quindi imparare a storpiare leggermente le parole, inserendo dei semplici "errori" ortografici e, in questo modo, evitare di essere scoperto. L'essenza, quindi, di questi attacchi risiede nell'indurre la macchina a considerare legittimi input che invece non lo sono (Buchanan e Miller, 2017).

Un'altra classe di attacchi molto insidiosi è quella dei cosiddetti *causative attacks*. In questo caso, l'avversario cerca di creare delle debolezze nel sistema che sfrutterà in un secondo momento. Un tipico attacco di questo tipo è quello del cosiddetto "avvelenamento" del training data set. Questo altro non è che un meccanismo per creare ad arte una distorsione nel sistema talmente forte che, in definitiva, lo porta a imparare "le cose sbagliate" e quindi poi commettere errori, anche importanti, nella classificazione effettuata durante il suo uso (Buchanan e Miller, 2017). Ad esempio, avversari del nostro sistema antispam possono "avvelenare" i dati di training con una gran quantità di messaggi con contenuto pedo-pornografico e istruirlo in maniera da fare considerare questi messaggi legittimi; in questo modo, se l'antispam impara che la presenza di riferimenti alla pedopornografia è indizio del fatto che un messaggio non è uno spam, non sarà in grado di bloccare messaggi di quel tipo. Gli attacchi di questo tipo sono particolarmente diffusi specie contro quei sistemi di ML che fanno re-training, cioè che continuano a (ri)calibrare i loro parametri anche sulla base dei dati che ricevono durante l'uso. Naturalmente, tutti i tipi di classificatori che basano la loro funzionalità sul training possono essere soggetti ad attacchi di tipo *poisoning*.

## 2.4. Osservazioni conclusive

In questo capitolo abbiamo fornito una breve introduzione alla IA e alle tecniche di ML, ed esposto i limiti e le problematicità di queste ultime.

È attualmente in atto una vera e propria corsa agli armamenti basati sulla IA, i big-data e le armi "cyber". A questa si aggiunge la progressiva digitalizzazione non solo del campo di battaglia, ma anche di tutte le infrastrutture militari, incluso il "complesso militare-nucleare", dall'infrastruttura di progettazione e approvvigionamento dei componenti dei sistemi d'arma nucleare, alle armi stesse, alla logistica, fino al sistema NC3–Nuclear Command Control and Communications (Lin, 2021). Si crea così un pericoloso nesso fra il cyber-space e il complesso nucleare militare, specie se si consi-



 ISBN 9788835152224

dera, come nota Lin (2021), che le iniziative di "modernizzazione" del complesso militare-nucleare statunitense, lanciate dal presidente Obama e ancora in corso, potrebbero prevedere, come pare, l'integrazione dei sistemi C3 convenzionali con il NC3. Va sottolineato che i sistemi militari attuali non sono per nulla esenti da vulnerabilità informatiche e non c'è motivo per sperare che lo siano quelli futuri (Latella, 2021; USDOD, 2013; USGAO, 2017b; 2018; 2019; 2021)[10]. D'altra parte, il cyber-space è ormai considerato un vero e proprio dominio e gli attacchi cyber, quando rivolti a infrastrutture militari o civili "critiche", vengono considerati dei veri e propri attacchi militari, ai quali eventualmente rispondere sia con altrettanti attacchi, o addirittura contemplando una risposta nucleare (USDOD, 2018; Futter, 2018; 2020; Marrone e Sabatino, 2021).

Questi elementi contribuiscono a incrementare l'incertezza, reale o percepita, durante la gestione di una crisi o di un conflitto, con il rischio concreto di escalation verso il conflitto nucleare e/o di guerra (nucleare) per errore. Il contributo che l'uso delle tecnologie della IA, anche per i limiti e le problematicità che abbiamo visto in questo capitolo, apporterà all'incremento di incertezza e confusione non sarà trascurabile (Geist e Lohn, 2018). Ad esempio, è ragionevole tenere presente che «è possibile immaginare un attacco [...] di poisoning dei dati che potrebbe portare un sensore [di un sistema di ML] a classificare un amico come nemico o a non rilevare la presenza di un nemico» (Allen e Chan, 2017, trad. it. a cura degli autori). Tutti questi elementi rendono ancora più delicata la gestione del rischio di escalation (Futter, 2019; Boulanin *et al.*, 2020; Turell *et al.*, 2020; Kubiak *et al.*, 2021).

Eppure nel rapporto della Commissione per la Sicurezza Nazionale sulla IA degli Stati Uniti – nominata dal Congresso e dall'Esecutivo e presieduta da E. Schmidt, già AD di Google e presidente esecutivo della stessa e di Alphabet Inc .– si legge: «Gli Stati Uniti devono prepararsi a difendersi dalle minacce [basate su IA] adottando velocemente e responsabilmente la IA per scopi di sicurezza nazionale e difesa» (NSCAI, 2021, p. 9) e «[d]ifendersi da avversari che dispongono di capacità IA senza utilizzare la IA è un invito al disastro» (NSCAI, 2021, p. 23). «Il dipartimento [della Difesa] deve agire adesso per integrare la IA nelle funzioni critiche, nei sistemi esistenti, nelle esercitazioni e war-games in modo da divenire una forza "AI-ready" entro il 2025» (NSCAI, 2021, p. 77, trad. it. a cura degli autori).

Queste autorevoli affermazioni fanno eco a posizioni radicali, controverse e a nostro avviso pericolose, perché attribuite a personalità di alto

[10] Sebbene, in generale, si faccia spesso riferimento solo alla situazione negli Stati Uniti, per la quale esiste ed è accessibile abbondante documentazione, non c'è motivo per considerare quella di altri paesi più avanzata o rassicurante.

livello, come quelle di William Roper, all'epoca capo dello Strategic Capabilities Office del Pentagono:

> [I] [...] dati lavorano per te. Tu accumuli più dati possibile e li addestri a insegnare e addestrare sistemi autonomi [...]. [L]o scopo del primo o del secondo giorno [di battaglia] non sarà [più] quello di uscire e distruggere aerei nemici o altri sistemi. Esso è [invece] quello di uscire, accumulare dati, fare ricognizione di dati, così che i nostri sistemi di apprendimento diventino più intelligenti di [quelli del nemico] (Tucker, 2017 in Buchanan e Miller, 2017, pp. 21-22, trad. it. a cura degli autori).

Parnas (2017) sottolinea come sia opportuno e necessario sviluppare attività di ricerca nel campo della IA, con particolare riferimento alla *explainability* e alla sicurezza riguardo agli attacchi tradizionali e *adversarial*. Inoltre, è importante che, prima di utilizzare questi sistemi fuori dai laboratori di ricerca, vengano valutate accuratamente e in maniera rigorosa tutte le implicazioni di natura etica e sociale e, nel caso dell'utilizzo per scopi militari e di sicurezza internazionale, anche quelle del diritto internazionale[11].

Si sente spesso affermare, anche in ambienti scientifici, che di fronte a un problema del quale non si conosce la soluzione, piuttosto che affrontarlo con il metodo scientifico, partendo dall'osservazione e cercando di individuare cause ed effetti per sviluppare una teoria, sia più opportuno affidarsi alla IA, perché con sufficiente potenza di calcolo e sufficienti dati essa lo risolverà. Noi invece pensiamo che questo approccio sia sbagliato e pericoloso e che la costruzione di sistemi informatici affidabili debba comunque essere guidata da solidi principi di ingegneria dei sistemi e *trustworthy computing*, come sottolineato da Cerf (2019) e Neumann (2019).

Concludiamo, quindi, affermando che condividiamo appieno i punti di vista di eminenti personalità della comunità degli informatici, quali David L. Parnas, William G. Cerf e Peter G. Neumann, sulla ricerca in IA e, più in generale, in informatica e sui rischi dell'uso di queste tecnologie, specie in campi di applicazione critici, dove si richiede il massimo possibile di affidabilità e, più in generale, di *dependability* e *trustworthyness* (Avizienis *et al.*, 2004). In particolare, riteniamo che sia necessaria un'approfondita discussione sull'utilizzo della IA in campo militare, che affronti non solo i potenziali vantaggi della militarizzazione della IA, ma anche e soprattutto i possibili rischi e la relativa governance (Stanley Center, UNODA, STIMSON, 2019; Boulanin *et al.*, 2020; Work, 2021). Allo stesso tempo, riteniamo che lo studio della IA sia sicuramente interessante, in particolare in campo civile, specie se usata in specifici domini di applicazione, per compiti ben definiti e con utili applicazioni.

[11] Al riguardo, si rimanda ai capitoli 5, 8 e 9 e a Tamburrini (2020) e Fossa *et al.* (2021).



 ISBN 9788835152224

## Riferimenti bibliografici


Aiello M., Pratt-Hartmann I., van Benthem J., eds. (2007), *Handbook of Spatial Logics*, Springer, Dordrecht, NL.

Allen G. and Chan, T. (2017), *Artificial Intelligence and National Security. STUDY 2017*, Harvard Kennedy School, Belfer Center for Science and International Affairs, Cambridge (MA), USA.

Amoroso D., Sauer F., Sharkey N., Suchman L., Tamburrini, G. (2018), *Autonomy in Weapon Systems. The Military Application of Artificial Intelligence as a Litmus Test for Germany's New Foreign and Security Policy*, The Heinrich Böll Foundation, Berlin, D.

Andride S.A. (1987), *Artificial Intelligence and National Defense. Applications to $C^3I$ and Beyond*, AFCEA International Press, Washington, USA.

Avizienis A., Laprie J-C., Randell B., Landwehr C. (2004), "Basic Concepts and Taxonomy of Dependable and Secure Computing", *IEEE Transactions on Dependable and Secure Computing.* 1(1):11-33.

van Benthem J. and Blackburn P. (2006), *Modal Logic: A Semantic Perspective*, In: Blackburn P., van Benthem J., Wolter F., eds., *Handbook of Modal Logic,* Springer, Dordrecht, NL.

Boulanin V. and Verbruggen M. (2017), *Mapping the Development of Autonomy in Weapon Systems*, Stockholm International Peace Research Institute (SIPRI) Stockholm, SE.

Boulanin V., Saalman L., Topychkanov P., Su F., Carlsson M.P., Richards L. (2020), *Artificial Intelligence, Strategic Stability and Nuclear Risk*. Stockholm International Peace Research Institute (SIPRI) Stockholm, SE.

Boulanin V., Goussac N., Bruun L., Richards L. (2020), *Responsible Military Use of Artificial Intelligence. Can the European Union Lead the Way in Developing Best Practice?* Stockholm International Peace Research Institute (SIPRI) Stockholm, SE.

Buchanan B. and Miller, T. (2017), *Machine Learning for Policymakers. What It Is and Why It Matters. Paper 2017*, Harvard Kennedy School, Belfer Center for Science and Int. Affairs, Cambridge (MA), USA.

Cerf V.G. (2019), "AI Is Not an Excuse!", *Communications of the ACM*, 62(10): 7.

CRS (2018), *U.S. Ground Forces Robotics and Autonomous Systems (RAS) and Artificial Intelligence (AI): Considerations for Congress*, CRS Report, Congressional Research Service, Washington, USA.

Cummings M.L. (2017), *Artificial Intelligence and the Future of Warfare*, Chatham House. The Royal Institute of International Affairs, London, UK.

Din A.M., eds. (1986), *Arms and Artificial Intelligence. Weapon and Arms Control Applications of Artificial Intelligence*, Stockholm International Peace Research Institute (SIPRI), Stockholm, SE.

Dyndal G.L., Berntsen T.A., Redse-Johasen S. (2017), "Autonomous military drones: no longer science fiction", *NATO Review Magazine,* 28 luglio 2017.

van Ditmarsch H., Halpern J.Y., van der Hoek W., Kooi B., eds. (2015), *Handbook of Epistemic Logic. London*, College Publications, Rickmansworth, UK.




 ISBN 9788835152224

Edwards C. (2019), "Hidden Messages Fool AI", *Communications of the ACM*, 62(1): 13-14.
Ekelof M. and Persi Paoli G. (2020), *Swarm Robotics. Technical and Operational Overview of the Next Generation of Autonomous* Systems, United Nations Institute for Disarmament Research (UNIDIR), Geneva, CH.
Emerson E.A. (1990), *Temporal and Modal Logic*, In: van Leeuwen J., *Handbook of Theoretical Computer Science*, The MIT Press, Boston, USA.
Fossa F., Schiaffonati V., Tamburrini G. (2021), *Automi e persone. Introduzione all'etica dell'intelligenza artificiale e della robotica*, Carocci, Roma.
Franklin S. (2014), *History, motivation and core themes*, In: Frankish K. e Ramsey W. M. eds., *The Cambridge Handbook of Artificial Intelligence*, Cambridge University Press, Cambridge, UK.
Futter A. (2018), *Hacking the Bomb: Cyber Threats and Nuclear Weapons*. Georgetown University Press, Washington D.C., USA.
Futter A. (2019), *Managing the Cyber-Nuclear Nexus*, Policy Brief, European Leadership Network, Londra, UK.
Futter A. (2020), *What does cyber arms control look like? Four principles for managing cyber risks*, Global Security Policy Brief, European Leadership Network, Londra, UK.
Geist E. and Lohn A. (2018), *Security 2040. How might Artificial Intelligence affect the risk of nuclear war*, RAND Corporation.
Goodfellow I., Bengio Y.Y., Courville A. (2016), *Deep Learning*, The MIT Press, Cambridge (MA), USA.
Gunning D., Aha D.W. (2019), "DARPA Explainable Artificial Intelligence Program", *AI Magazine*, 40(2): 44-58.
HumanE AI (2019), *Human-Centered Artificial Intelligence*. EU Horizon2020 funded project https://www.humane-ai.eu (accesso effettuato il 26 gennaio 2022).
Klarreich E. (2016), "Learning Securely", *Communications of the ACM*, 59(11):12-14.
Kubiak K., Misrha S., Stacey G. eds. (2021), *Nuclear weapons decision-making under technological complexity.* Pilot Workshop Report. Global Security, European Leadership Network, Londra.
Landman N., Pang H., Williams C. (2019), "K-Means Clustering", *Brilliant.org,* testo disponibile al sito: https://brilliant.org/wiki/k-means-clustering/ (accesso effettuato il 26 gennaio 2022).
Latella D. (2021), "Sicurezza informatica, armi nucleari e stabilità strategica", *IRIAD Review*, n. 3, marzo-aprile, Istituto di Ricerche Internazionali Archivio Disarmo – IRIAD.
Lin H. (2021), *Cyber Threats and Nuclear Weapons*, Stanford University Press, Stanford (CA), USA.
Marrone A. and Sabatino E. (2021), *Cyber Defense in NATO Countries: Comparing Models*, IAI Papers 21, 5, Istituto Affari Internazionali, Roma
Neumann P.G. (2019), "How Might We Increase System Trustworthyness?", *Communications of the ACM*, 62(10): 23-25.

NSCAI (2021), National Security Commission on Artificial Intelligence. *Final Report.*
OpenAI (2019), *Attacking Machine Learning with Adversarial Examples*, testo disponibile al sito: https://openai.com/blog/adversarial-example-research/ (accesso effettuato il 26 gennaio 2022).
Parnas D.L. (2017). "The Real Risks of Artificial Intelligence", *Communications of the ACM*, 60(10): 27-31.
Pedregosa F., Varoquaux G.**,** Gramfort A.**,** Michel V., Thirion B.**,** Grisel O.**,** Blondel M.**,** Prettenhofer P.**,** Weiss R.**,** Dubourg V.**,** Vanderplas J.**,** Passos A.**,** Cournapeau D.**,** Brucher M.**,** Perrot M.**,** Duchesnay, E. (2011), "Machine Learning in Python", *Journal of Machine Learning Reearch*, 12: 2825-2830.
Ribeiro M.T. and Singh S., Guestrin C. (2016), *Why Should I Trust You? Explaining the Predictions of Any Classifier*. In: Krishnapuram B., Shah M., Smola A.J., Aggarwal C.C., Shen D., Rastogi R., eds., *Proceedings of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Francisco, CA, USA, Aug. 13-17, 2016*. USA: Association for Computing Machinery (ACM).
Rossi J.C. (2019), "Un'opera dell'uomo: le macchine autonome letali", *IRIAD Review*, 5:2-22.
Schneier B. (2018), *Click here to kill everybody. Security and survival in a hyperconnected world*, W.W. Norton & Company Inc., New York, N.Y., USA.
Stanley Center, UNODA, STIMSON (2019), *The Militarization of Artificial Intelligence*, Stanley Center for Peace and Security, United Nations Office for Disarmament Affairs (UNODA), Stimson Center, UN, New York, August 2019.
Tamburrini G. (2020), *L'etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*. Carocci, Roma.
Thys S., van Ranst W., Goedeme T. (2019), "Fooling automated surveillance cameras: adversarial patches to attack person detection", https://arxiv.org/abs/1904.08653 (accesso effettuato il 26 gennaio 2022).
Turell J. and F., Boulanin V. (2020), *Cyber-incident Management. Identifying and Dealing with the Risk of Escalation*, Stockholm International Peace Research Institute (SIPRI), Stockholm, SE.
UNIDIR (2018), *The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence – a primer for CCW delegates*, UNIDIR Resources N. 8, United Nations Institute for Disarmament Research (UNIDIR), Geneva.
USDOD (2013), *Task Force Report: Resilient Military Systems and the Advanced Cyber Threat*, US Department of Defense, Defense Science Board, Washington, USA.
USDOD (2016), *Report of the Defense Science Board Summer Study on Autonomy*, US Department of Defense, Defense Science Board, Washington, USA.
USDOD (2018), *Nuclear Posture review*, US Department of Defense, Washington, USA.
USGAO (2017a), *Internet of Things. Status and implications of an increasingly connected world,* Report to Congressional Requesters, GAO-17-75, US Government Accountability Office – GAO, Washington DC, USA.

USGAO (2017b), *Internet of Things. Enhanced Assessments and Guidance Are Needed to Address Security Risks in DOD.* Report to Congressional Committees, GAO-17-668, US Government Accountability Office – GAO, Washington DC, USA,

USGAO (2018), *Weapon Systems Cybersecurity. DOD Just Beginning to Grapple with Scale of Vulnerabilities*, Report to the Committee on Armed Services, U.S. Senate, GAO-19-128, US Government Accountability Office – GAO, Washington DC, USA.

USGAO (2019), *Future Warfare. Army Is Preparing for Cyber and Electronic Warfare Threats, but Needs to Fully Assess the Staffing, Equipping, and Training of New Organizations,* Report to Congressional Committees, GAO-19-570, US Government Accountability Office – GAO, Washington DC, USA.

USGAO (2021), *Weapon Systems Cybersecurity. Guidance Would Help DOD Programs Better Communicate Requirements to Contractors*, Report to Congressional Committees, GAO-21-179, US Government Accountability Office – GAO, Washington DC, USA.

Work R.O. (2021), *Principles for the Combat Employment of Weapon Systems with Autonomous Functionalities*, Center for New American Security, Washington DC, USA.



 ISBN 9788835152224