Intelligenza Artificiale / Machine Learning: tra Complessità e Sicurezza

Negli ultimi anni Intelligenza Artificiale e Machine Learning sono diventati nomi noti a tutti e usati molto spesso. Ormai moltissime applicazioni informatiche hanno componenti o sono per lo più basate su Intelligenza Artificiale e/o Machine Learning, ma la loro adozione comporta, come sempre, dei rischi anche di sicurezza. In questo articolo verrà proposta una breve e, come si vedrà, incompleta panoramica su alcuni aspetti di Intelligenza Artificiale e Machine Learning che sono o potrebbero diventare rilevanti per la sicurezza, informatica e non. Ma a questo scopo è molto utile fare un passo indietro e partire con un breve riassunto di cosa sono Intelligenza Artificiale e Machine Learning.

Intelligenza Artificiale

Con il nome "Intelligenza Artificiale", in breve AI (*Artificial Intelligence*), si intende una branca dell'informatica con lo scopo di sviluppare sistemi informatici con competenze simili a quelle dell'uomo quali quelle:

1. Dell'apprendimento,
2. Del ragionamento e
3. Della soluzione di problemi.

I sistemi AI hanno lo scopo di aiutare l'uomo in tantissimi campi, dal trovare informazioni (i *Personal Virtual Assistant* come *Cortana, Google Assistant, Siri,* ecc.), a dimostrare teoremi matematici, alla guida autonoma di automobili, alla realizzazione di Robot (o Automata) che ad esempio possono lavorare in ambienti nocivi per l'uomo, a diagnosticare

malattie, gestire il traffico, supportare le vendite, giocare a scacchi o Go, ecc. ecc., e di fare tutto questo molto velocemente, con grande precisione, pochissimi errori e grande affidabilità. AI può quindi interessare quasi tutti i settori della vita umana, dall'educazione alla sanità, ai trasporti, agli ambienti di lavoro, divertimento, commercio, economia, agricoltura, sicurezza ecc.

E' chiaro quindi che lo sviluppo di AI richiede, oltre all'informatica, conoscenze in molte altre discipline quali ad esempio:

- Matematica
- Biologia
- Psicologia
- Sociologia
- Neurologia
- Statistica.

La storia di AI è molto antica, ad esempio miti di uomini meccanici (Automata) sono già presenti presso gli antichi greci, ma si può far risalire agli anni '40 l'inizio dello sviluppo moderno di AI, in particolare con il contributo di Alan Turing nel 1950. AI ha vissuto altri due periodi di grande sviluppo: il primo negli anni '80 con la formulazione dei Sistemi Esperti: algoritmi specifici in grado di analizzare molte informazioni in un determinato ambito per dedurne le decisioni migliori a supporto ed imitazione dei migliori esperti e *decision maker*. Il secondo periodo è quello attuale, iniziato indicativamente verso la fine degli anni '90 e dovuto in gran parte sia all'aumento della potenza di calcolo degli elaboratori sia allo sviluppo delle Reti Neurali Artificiali (o *Artificial Neural Network,* ANN) utilizzate precedentemente quasi esclusivamente nella ricerca ad esempio della fisica dei sistemi elementari. Questo ha portato all'affermarsi del *Machine Learning* (in breve ML), degli "Agenti Intelligenti" e di tutto quello che oggi consideriamo sotto la sigla AI/ML.

La scienza AI odierna e del prossimo futuro viene indicata come “debole” o “ristretta”, in quanto un sistema AI attuale è in grado di svolgere solo uno o pochi compiti specifici in maniera tipicamente reattiva e con memoria limitata. Non esistono quindi sistemi AI in grado di svolgere funzioni veramente simili a quelle del cervello umano, in particolare di apprendere e ragionare come o meglio di un uomo. In altre parole, non siamo ancora neppure vicini alla realizzazione di veri Androidi (o Robot umanoidi / antropomorfi, o Replicanti) né di elaboratori senzienti (come HAL 9000 nel film “Odissea dello spazio 2001”) basati su sistemi AI con intelligenza e capacità uguali o superiori a quelle del cervello umano (detti sistemi AI Generali e Superiori rispettivamente). Non c’è quindi un pericolo immediato che le macchine prendano il controllo degli umani, come descritto in tantissimi racconti, libri e film di fantascienza; ma alla luce di quanto già succede oggi e che sarà descritto più avanti, è comunque una preoccupazione da tenere in conto sin da subito.

Un’automobile a guida autonoma come esempio

Per comprendere la complessità degli attuali sistemi AI/ML è utile partire da un esempio molto semplice ed ormai noto a tutti: cercheremo di descrivere a livello molto alto l’approccio AI/ML alla realizzazione di un’automobile a guida autonoma (o equivalentemente un Robot adibito a qualche compito specifico). Per potersi muovere un’automobile a guida autonoma ha bisogno di descrivere l’ambiente in cui si trova, è necessario quindi che sia dotata di **sensori** che le permettono di “vedere” la strada, gli altri veicoli, i cartelli stradali, eventuali pedoni, biciclette, motoveicoli, animali, e quant’altro potrebbe trovarsi su o intorno al suo percorso, ma anche di “sentire” una sirena di un mezzo di soccorso, un clacson di un altro veicolo ecc. Dovrebbe essere ovvio che la tecnologia odierna permette di effettuare tutti questi tipi di rilevazioni ma con sensori diversi (ad esempio

anche solo a secondo delle condizioni climatiche: sole, nebbia, pioggia, neve ecc.), ognuno con un diverso algoritmo necessario per identificare gli elementi attesi quali la presenza di un pedone, la lettura di un cartello stradale, la posizione della corsia stradale ecc. Inoltre servono anche altri sensori per monitorare il veicolo stesso: acceleratore, freno, marcia, motore, sterzo e tutti i parametri del veicolo.

Si paragoni questa situazione con quella di un uomo alla guida: per percepire e descrivere l'ambiente bastano solo due sensori, gli occhi e le orecchie; e per monitorare il veicolo anche mani (sul volante) e piedi (sui pedali). Inoltre, l'unità di elaborazione umana è composta da un unico, grande algoritmo: il nostro cervello.

Ma per poter procedere, un'automobile a guida autonoma ha bisogno di **valutare** tutte le rilevazioni fornite dai sensori **e decidere** come procedere sulla strada. Anche in questo caso vi sono diversi ragionamenti e tipi di decisione da prendere, tra cui:

- Identificare il percorso da seguire per raggiungere la destinazione (come i navigatori satellitari che tutti ben conosciamo);
- Sulla base delle condizioni stradali rilevate, identificare la maniera migliore per percorre i successivi pochi metri di strada;
- Identificare eventuali situazioni anomale o di pericolo che richiedono azioni di emergenza.

Di nuovo, per svolgere tutte le valutazioni necessarie alla guida autonoma e prendere le relative decisioni sono utilizzati algoritmi di AI/ML distinti e dedicati ad ogni singolo ambito.

Infine, molte decisioni portano all'**implementazione di un'azione** tramite un **attuatore,**[1] ovvero un dispositivo che traduce una decisione di un sistema informatico in un'azione

fisica. Ovviamente anche gli attuatori necessitano logiche e algoritmi dedicati.

Riprendendo quanto indicato all'inizio, questo esempio ci descrive come un sistema AI/ML:

1. Apprenda tramite i sensori;
2. Ragioni tramite algoritmi dedicati valutando le informazioni e prendendo decisioni;
3. Risolva i problemi implementando delle azioni tramite gli attuatori.

Nella terminologia AI/ML corrente, questo approccio è descritto come quello di un "Agente Intelligente" (o solo "Agente") le cui caratteristiche principali sono:

1. Un Agente deve avere la capacità di percepire (osservare) l'ambiente;
2. Le osservazioni sono utilizzate dall'Agente per prendere delle decisioni;
3. Una decisione tipicamente porta l'Agente ad eseguire un'azione;
4. L'azione intrapresa da un Agente deve essere un'azione razionale.

E' importante sottolineare un aspetto che emerge da questo semplice esempio: a differenza dell'uomo che utilizza il proprio cervello come unità di elaborazione, i 5 sensi come sensori ed il corpo come attuatore, un tipico Agente o sistema AI/ML utilizza una **grande numerosità e varietà di sensori, algoritmi e attuatori** con caratteristiche spesso molto diverse tra loro.[2]

Pertanto può essere difficile individuare quali approcci, modelli, algoritmi siano i più adatti, efficaci e performanti per sviluppare, implementare o anche solo adottare un sistema AI. Come esempio, si provi ad esplorare alcune delle decine di librerie esistenti di algoritmi ML [si veda Rif. 1]. In altre parole, l'ampiezza di AI è tale da costituire da sola un

aspetto di complessità da non sottovalutare in progetti che includono l'utilizzo di questi sistemi.

Sinora si è fatto cenno al concetto di Agente ma AI, ed in particolare ML, è costituito anche e principalmente da sistemi più semplici, puramente informatici e che non coinvolgono un'interazione con un ambiente esterno. Dal punto di vista IT è un programma con dati in input ed in output; per un Agente sono tutte quelle componenti che implementano la logica che analizza i dati ed elabora le decisioni. Nel rimanente di questo articolo si considererà solo la parte puramente informatica di AI.

Machine Learning

ML può essere definito come una sotto-branca di AI che si occupa di sviluppare sistemi informatici in grado di apprendere dai dati forniti in ingresso in modo tale da poter fornire delle elaborazioni utili ma senza essere programmati in modo esplicito.

Ripensando ai concetti base di programmazione, è ben noto che l'approccio più semplice per scrivere un programma è quello di implementare direttamente nel codice l'algoritmo che esegue l'elaborazione voluta (approccio "procedurale").

Un approccio a più alto livello adottato anche nei Sistemi Esperti, è quello di implementare nel codice una logica generica (chiamata anche Motore delle Regole o *Rules Engine*) in grado di analizzare insieme ai dati specifici in ingresso, una Base di Conoscenza (o *Knowledge Base*, KB) fornita dagli esperti, in modo da ottenere i risultati attesi.

ML compie un ulteriore passo di astrazione in quanto nel codice sono implementate delle strutture logiche molto semplici ed il modello deve imparare dai dati in ingresso anche le regole di elaborazione. L'esempio più immediato è quello di una Rete Neurale che non è altro che una rete con

dei nodi di ingresso in cui vengono inseriti i dati da elaborare, nodi interni in cui vengono eseguite le elaborazioni e nodi finali in cui appaiono i risultati (si veda ad esempio Fig. 1). Le elaborazioni svolte dai nodi sono basate su dei coefficienti numerici inizialmente ignoti e scelti in maniera pseudo-casuale. Il modo più semplice per fissare i coefficienti numerici è di fornire in ingresso dei dati per i quali si conosce il risultato dell'elaborazione, e di calcolare i coefficienti numerici in modo da ottenere il risultato corretto.



Fig. 1 Struttura di una rete neurale molto semplice (da Wikipedia)

Quindi all'inizio una rete neurale non ha insita alcuna logica e si comporta in modo casuale nella elaborazione dei dati in ingresso. Nella fase di **apprendimento (**ovvero determinazione dei coefficienti numerici), i dati forniti alla rete neurale introducono nel programma la logica di elaborazione. Facendo un paragone con i Sistemi Esperti, i dati in ingresso nella fase di apprendimento forniscono alla rete neurale sia il Motore delle Regole che la Base di Conoscenza, entrambi codificati nei coefficienti numerici dei nodi. In altre parole, un modello ML è basato principalmente su due componenti:

1. Algoritmi statistici e/o predittivi principalmente basati su algebra lineare;
2. Un insieme di dati (*data-set*) utilizzato per l'apprendimento da parte del modello.

L'interrelazione tra queste due componenti è la chiave del successo, o del fallimento, di un modello ML. Per gli algoritmi, la principale difficoltà è di identificare quali sono più adatti ad apprendere una determinata logica, ad esempio alcuni algoritmi risultano più efficaci ad identificare il contenuto di un'immagine (ad esempio un

Convolutional Neural Network, o CNN) mentre altri a selezionare il percorso ottimale[3] da percorrere (ad esempio un algoritmo ad albero). Invece il problema principale del data-set utilizzato per istruire il modello ML, è quello di dover contenere le informazioni necessarie per l'apprendimento in maniera equilibrata. Questa seconda difficoltà si sta' dimostrando sempre più complessa essendo la principale causa dei problemi dei modelli ML. Prova ne è che Gartner prevede che fino al 2022, l'85% dei progetti di intelligenza artificiale produrrà risultati errati a causa di errori nei data-set di apprendimento, errori negli algoritmi o errori da parte dei team responsabili della loro gestione [Rif. 2].

Al contempo è importante sottolineare la grande potenzialità ed efficacia già raggiunta dai modelli AI/ML. Innanzitutto bisogna ricordare che questi si basano principalmente su modelli statistici e pertanto non danno quasi mai risposte esatte, come invece nel caso di un approccio "procedurale" ove se si verifica una condizione ne segue un'affermazione unica. Invece i modelli AI/ML tipicamente "apprendono" delle caratteristiche generali (dei *pattern*) di un insieme di dati e calcolano quanto un altro set di dati è simile a quello di riferimento. I modelli AI/ML forniscono quindi una valutazione statistica di similitudine (probabilità) tra le caratteristiche di interesse dei due set di dati. Questo processo permette ai modelli AI/ML di essere predittivi, ovvero di individuare anche in set di dati molto complessi, caratteristiche simili, non identiche, a quelle di interesse. In altre parole, questi modelli sono in grado di individuare dei pattern anche in set di dati mai analizzati precedentemente, ed essere quindi predittivi. La maggior parte dei modelli AI/ML è quindi basata sulla statistica, apprende dei pattern ed è in grado di essere predittivo.

Debolezze dei modelli AI/ML

I principali rischi connessi all'utilizzo di modelli ML

provengono principalmente da due problemi:

1. Il modello produce risultati errati anche in situazioni ove non ci si aspetterebbe nessun errore; oppure esistono modi molto semplici / facili per ingannare il modello;
2. Il modello produce, anche volutamente, risultati ingannevoli o risultati falsi difficilmente distinguibili da dati veri.

Vi sono molti casi noti del primo punto, come ad esempio tutti gli esperimenti in cui è stato possibile ingannare un modello ML di riconoscimento delle immagini sottoponendogli delle immagini leggermente deformate, od aggiungendo colori, sticker (anche a cartelli stradali) ecc. (si veda ad esempio Rif, [3]).[4] La problematica in realtà è molto profonda ed include anche quello che usualmente viene chiamato *bias*[5] del modello. Ad esempio alcuni modelli di riconoscimento facciale riconoscevano con grande precisione maschi bianchi ma con bassa precisione femmine nere, ed analoghi problemi sono stati riscontrati in tanti altri modelli ML come alcuni per il calcolo del rischio per il rimborso di un mutuo a seconda delle condizioni sociali, etniche, economiche, ecc. A prima vista sembrerebbe facile risolvere il problema, dovrebbe bastare ri-eseguire la fase di apprendimento del modello con un data-set più completo ed equilibrato. Ma in pratica questo spesso non riesce in quanto il miglioramento della precisione di una caratteristica del modello porta quasi sempre a peggiorare o almeno non migliorare qualche altra caratteristica o a peggiorare la sua precisione ed efficacia generale. Ad esempio in presenza contemporanea di un bias sul genere e uno sull'etnia delle persone, risulta quasi sempre difficile se non impossibile rimuoverli entrambi [Rif. 4].

Inoltre è stato dimostrato che i modelli attuali di ML [Rif. 5] esibiscono un problema forse ancora più profondo chiamato *underspecification*. In pratica è stato osservato che il processo di apprendimento di un modello ML è suscettibile a

minime, impercettibili variazioni con potenziali serie conseguenze. Come indicato precedentemente, il processo di apprendimento di un modello ML dipende (fortemente) dal data-set utilizzato, ma anche dai valori pseudo-casuali del modello iniziale, dalla procedura di apprendimento stessa e dalla procedura di test della correttezza del modello prodotto. Ripetendo la stessa procedura di apprendimento è possibile generare un modello ottimo come un modello con bias od errori: il problema è che non si ha modo di saperlo sino a quando l'errore o il bias si manifesta. Purtroppo questo tipo di fallimento dei modelli ML avviene spesso quando meno ce lo si aspetta, ad esempio nel caso di due immagini che per l'occhio umano sono quasi identiche ma che il modello ML considera completamente diverse.

La seconda problematica è in realtà costituita da due aspetti che saranno approfonditi più avanti. Un esempio del primo aspetto, che fa parte degli *Adversarial Attacks*, è l'introduzione di dati "malevoli" (anche abilmente camuffati) nel data-set di apprendimento che porta quindi alla produzione di risultati ingannevoli. La difficoltà di individuazione e gestione di questo attacco è dovuta al fatto che è praticamente impossibile dedurre le logiche apprese da un modello ML se non dai risultati che produce e quindi solo dopo il suo utilizzo. Il secondo aspetto è di utilizzare modelli ML per produrre documenti digitali (testi, immagini, video, suoni) falsi ma difficilmente distinguibili dagli originali sia per gli uomini sia per gli algoritmi ML stessi (*Deep Fake*).

Altre debolezze dei modelli ML (si veda ad esempio Rif. [6] per un breve riassunto) sono:

- Eseguire di nuovo il processo di apprendimento su di un modello già preparato per aggiungere ulteriori casi o correggere errori, può introdurre nuovi errori sui dati preesistenti; ad esempio il modello può "dimenticare" pattern che conosceva e quindi non riconoscere più

oggetti, situazioni, set di dati ecc. che precedentemente erano riconosciuti;
- Per costruzione, la maggior parte degli algoritmi ML genera nella fase di apprendimento le logiche che poi utilizza, non è strano quindi che spesso sia difficile se non impossibile capire e spiegare la logica di un certo risultato dell'elaborazione; questo diventa particolarmente significativo quando il risultato non corrisponde a quanto ci si aspetta intuitivamente e si vorrebbe quindi capirne il perché;
- Tipicamente i modelli ML producono come risultato dei valori di probabilità, ovvero di incertezza del risultato stesso, e spesso è necessario decidere quando un risultato è da considerarsi positivo; ad esempio, in un'applicazione di riconoscimento di volti, bisogna decidere se un match al 80% è sufficiente per essere positivo, e questo è difficile da definire a priori visto quanto descritto al punto precedente;
- Infine, anche se sembra un paradosso, è molto difficile insegnare la matematica ad un modello ML, e non è chiaro perché un calcolatore da tasca è molto più efficiente e corretto dei più avanzati modelli ML nell'eseguire calcoli numerici e risolvere problemi matematici.

Dalle debolezze dei modelli AI/ML ai rischi di sicurezza

I modelli AI/ML sono oggi pervasivi ed utilizzati in moltissimi ambiti per cui le loro debolezze ed errori possono creare dei rischi di sicurezza in ambiti e con conseguenze molto diverse (si vedano ad esempio Rif. [7,8,9]). Qui di seguito ne indichiamo alcuni di sicuro interesse.

Sicurezza informatica e Cyber-sicurezza

Le capacità di individuare pattern su grandi quantità di dati e di essere predittivi sono molto utili ed ormai ampiamente

adottate nelle soluzioni e nei servizi di sicurezza informatica, dagli anti-virus/malware (oggi estesi ai sistemi EPP/EDR/XDR ecc.) al monitoraggio della rete e ai SIEM/SOC. Esempi sono l'individuazione di vulnerabilità e attacchi *Zero-day*, l'analisi di codice malevole, l'identificazione di un attacco nelle fasi della *kill-chain*, l'identificazione di campagne di *spam/phishing/spear-phishing* ed in generale di attacchi basati sul *social-engineering*. I modelli AI/ML possono anche supportare il disegno di strategie ed il coordinamento di azioni di difesa, ed eseguire analisi per l'identificazione dei responsabili di un attacco.

Al contempo le stesse tecniche possono essere utilizzate da un attaccante per rendere i propri attacchi più difficilmente individuabili, più efficaci e con tempi di evoluzione molto più rapidi. E' il ciclo (ben noto) di ogni tecnologia che può essere utilizzata sia in difesa che in attacco (si pensi ad esempio alle armi da fuoco) ed è estremamente importante non rimanere indietro.

Nell'utilizzo dei modelli AI/ML per la sicurezza informatica dobbiamo necessariamente tenere conto delle loro debolezze e, come difensore, che un attaccante possa sfruttarle ad esempio per fare in modo che il proprio attacco non venga rilevato. Quindi da una parte è necessario dotarsi di strumenti basati su AI/ML per rimanere al passo con le tecniche di attacco, dall'altro è anche necessario non basarsi esclusivamente su questi, ad esempio sia continuando ad utilizzare approcci più tradizionali sia mantenendo nei processi l'intervento diretto dell'uomo.

Fake News e Deep Fake

Il problema è ben noto a tutti, ed anche in questo caso il ruolo dei modelli ML è duplice. Da una parte i modelli ML possono essere utilizzati per individuare *Fake News* o comunque informazioni e materiale non lecito o consentito (come sui *Social Network*), ma ben vediamo che alle volte sono ben poco

efficaci o sbagliano nell'individuazione portando al blocco di materiale lecito (si veda come esempio Rif. [10]).

Ma i modelli ML possono essere utilizzati anche per generare informazioni, documenti, immagini, suoni falsi (*Deep Fake*). L'idea è semplice: si consideri ad esempio un'immagine o una composizione musicale e la si modifichi sino a quando sia all'occhio / orecchio umano sia per gli algoritmi ML questa non risulta praticamente indistinguibile da un originale. Un approccio pratico per creare un falso è quello di utilizzare due modelli ML, uno che genera dei dati ad esempio in un formato di brano musicale ed un secondo che ha appreso a riconoscere i brani musicali di Mozart (questo è un esempio di *Generative Adversarial Network*, GAN), e di metterli in loop in modo che ad ogni iterazione il primo impari a generare un falso seguendo le indicazioni del secondo. Rimane ancora un problema non del tutto risolto quello di trovare il modo di identificare i *Deep Fake* così generati.

Frodi e propaganda

Fake News e *Deep Fake* possono essere utilizzati come componenti di frodi o di campagne di disinformazione orchestrate od anche automatizzate da modelli ML, sfruttando ad esempio le capacità di questi di creare e gestire *spam, phishing* e attacchi di *social-engineering*. Come nell'ambito della sicurezza informatica, modelli ML possono essere utilizzati in difesa per identificare frodi (ed ormai sono utilizzati comunemente per individuare frodi nelle transazioni economiche) e campagne di disinformazione, con simili lati positivi e negativi.

Sicurezza delle informazioni nei data-set

Un aspetto interessante riguarda la sicurezza delle informazioni contenute nei data-set utilizzati per l'apprendimento di un modello ML. Si consideri un modello ML utilizzato per identificare informazioni segrete o molto sensibili (in ambito militare o di sicurezza nazionale,

sanitario, economico, dei diritti di autore ecc.) ad esempio in qualità o a supporto di un servizio di *Data Loss Prevention* (DLP). Ora nei data-set utilizzati per l'apprendimento di questo modello ML devono essere presenti sufficienti informazioni segrete o molto sensibili in modo da poterle differenziare da quelle non segrete o sensibili. Il rischio di sicurezza è che studiando il funzionamento del modello, le sue debolezze o utilizzandolo come un oracolo, sia possibile estrarre dal modello stesso delle informazioni sensibili presenti nel data-set di apprendimento (processo chiamato *Model Inversion*). Al momento questo è ancora un ambito di studio, ma questo rischio di sicurezza deve essere valutato tra le considerazioni da fare sulla sicurezza delle informazioni contenute nei data-set utilizzati per l'apprendimento.

Abuso di sistemi autonomi

Un rischio che deriva da eventuali debolezze di un modello ML è quello di abuso di un sistema autonomo, sia questo un drone, un'automobile o un'applicazione IT che valuta la richiesta di un mutuo casa o che esegue una diagnosi di una malattia ecc. Ad esempio un attaccante a conoscenza di una debolezza di un modello ML di un'automobile a guida autonoma, potrebbe sfruttare questa debolezza per trasformare l'automobile in un'arma all'insaputa del proprietario e dei suoi passeggeri. Il tipico approccio per la mitigazione dei rischi di questo scenario consiste nel ridurre il livello di autonomia del sistema, facendo si che l'uomo possa sempre intervenire in situazioni critiche e per approvare le principali decisioni del sistema.

Ambito militare

L'ambito militare è uno di quelli in cui l'adozione di modelli ML può avere effetti positivi e negativi estremamente profondi. L'adozione è possibile sia in ambito difensivo che offensivo e si estende dagli aspetti strategici, alla raccolta

e analisi delle informazioni, alla pianificazione delle azioni ed alla realizzazione autonoma di azioni tramite droni o direttamente in *cyberspace*, il tutto supportato da velocità di decisione e azione molto superiore a quelle umane. Ma, ovviamente, le conseguenze di errori computi dai modelli di *Machine Learning* in questo ambito potrebbero essere molto gravi, sino ad arrivare agli scenari apocalittici descritti in libri e film di fantascienza. L'adozione dei modelli ML in questo ambito deve essere quindi sempre accompagnata da opportune contromisure che tengano conto dei rischi inerenti all'utilizzo di una tecnologia ancora così nuova ma al contempo così potente.

In conclusione, ricollegandoci a quanto scritto all'inizio, è vero che la scienza AI odierna e del prossimo futuro è ancora "debole" o "ristretta", ma le sue capacità e potenzialità unite alle sue attuali debolezze e alla nostra limitata comprensione, richiedono una continua valutazione e rivalutazione dei rischi di adozione e utilizzo, e che l'uomo sia sempre presente per valutare e prendere decisioni nei processi critici in cui AI/ML è coinvolto.

Riferimenti Bibliografici

Rif. 1: Ad esempio: scikit-learn https://scikit-learn.org/stable/index.html, TensorFlow https://www.tensorflow.org/, PyTorch https://pytorch.org/, mlpack https://mlpack.org/, ecc.

Rif. 2: "Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence", febbraio 2018, https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence

Rif. 3: "Psychedelic toasters fool image recognition tech", BBC News, gennaio 2018, https://www.bbc.com/news/technology-42554735 ; "Adversarial

Patch”. Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer, maggio 2018, https://arxiv.org/abs/1712.09665

Rif. 4: “Engineering Bias Out of AI”, IEEE Spectrum, aprile 2021, https://spectrum.ieee.org/engineering-bias-out-of-ai

Rif. 5: “The way we train AI is fundamentally flawed”, MIT Technology Review, novembre 2020, https://www.technologyreview.com/2020/11/18/1012234/training-machine-learning-broken-real-world-heath-nlp-computer-vision/

Rif. 6: “7 Revealing Ways AIs Fail”, IEEE Spectrum, settembre 2021, https://spectrum.ieee.org/ai-failures

Rif. 7: “AI’s 6 Worst-Case Scenarios”, IEEE Spectrum, gennaio 2022, https://spectrum.ieee.org/ai-worst-case-scenarios

Rif. 8: “A National Security Research Agenda for Cybersecurity and Artificial Intelligence”, CSET, maggio 2020, https://cset.georgetown.edu/publication/a-national-security-research-agenda-for-cybersecurity-and-artificial-intelligence/

Rif. 9: “AI-enabled future crime”, M. Caldwell, J. T. A. Andrews, T. Tanay & L. D. Griffin, Crime Science, agosto 2020, https://crimesciencejournal.biomedcentral.com/articles/10.1186/s40163-020-00123-8

Rif. 10: “The innocuous photos banned by Facebook: Social media giant apologises after it blocks art gallery’s images of COWS, the England cricket team and a high-rise office block because they are judged ‘too sexy’”, MailOnline News, febbraio 2021, https://www.dailymail.co.uk/news/article-9249087/Facebook-apologises-blocks-art-gallerys-images-COWS-sexy.html

Note

[1] In Inglese sono utilizzati due termini: *Actuator* se l’azione ha effetto sul sistema AI stesso ad esempio generando

un suo spostamento nello spazio, e *Effector* se l’azione ha effetto sull’ambiente esterno al sistema AI.

[2] Anche se esistono degli Agenti molto semplici, quale ad esempio un termostato “smart”.

[3] Anche chiamato “Problema del commesso viaggiatore”.

[4] Questo tipo di errore può anche essere sfruttato per un *Adversarial Attack*.

[5] Si preferisce mantenere il termine tecnico inglese “bias” piuttosto che utilizzare la parola “pregiudizio”.

Articolo a cura di **Andrea Pasquinucci**

https://www.ictsecuritymagazine.com/articoli/intelligenza-artificiale-machine-learning-tra-complessita-e-sicurezza/