# Intelligenza Artificiale: storia, progressi e sviluppi tra speranze e timori*

Luigi Portinale

## Abstract

l'Intelligenza Artificiale (IA) è una disciplina scientifica matura nata in seno all'informatica che ormai pervade i più svariati ambiti (scientifici e non scientifici), nonchè diversi aspetti della nostra vita quotidiana. Le fasi iniziali dello sviluppo dell'IA furono caratterizzate da grandi aspettative circa la possibilità di costruire facilmente programmi in gradi di esibire comportamenti "intelligenti" a livello umano. Nonostante le previsioni molto ottimistiche degli albori, la dura verità con cui i ricercatori si scontrarono presto fu di capire che comprendere e replicare i meccanismi dell'intelligenza umana è un compito molto arduo. Tale compito inoltre deve complementare metodologie molto sofisticate di progettazione di tali sistemi ed adeguate risorse computazionali. La storia dell'IA è stata infatti costellata di alti e bassi conosciuti in letteratura come le "stagioni dell'IA"; oggigiorno stiamo assistendo ad un interesse nell'IA forse mai visto prima. L'impatto delle metodiche di IA sta portando allo sviluppo di nuove applicazioni, producendo inoltre molteplici effetti in contesti di vario tipo, da quello socio-economico, a quello giuridico, a quello etico. In questo lavoro rivedremo brevemente lo sviluppo dei principali approcci che sono stati sviluppati all'interno dell'IA nei vari anni fino ai giorni nostri, caratterizzati dall'utilizzo pervasivo della metodica denominata apprendimento profondo o *deep learning*. Discuteremo le diverse tendenze presenti all'interno della disciplina, i punti di forza e debolezza, le limitazioni e l'impatto conseguente all'adozione di sistemi e dispositivi sempre più "intelligenti".

*Artificial Intelligence (AI) is a mature discipline stemming from computer science, pervading now every discipline (scientific and non-scientific) and several aspects of our everyday life. The initial steps of AI were essentially based on a great excitement about the possibility of building programs able to exhibit intelligent behavior at the human level. However, the initial optimistic predictions about such a possibility immediately faced the hard truth: understanding and replicating human intelligence is an extremely difficult problem and involves the availability of sophisticated methodologies, as well as the availability of suitable computational resources. In fact, the history of AI has evolved between ups and downs in the so-called "AI seasons", and today, we are experiencing a renewed interest for AI methods and applications never seen before. The impact of such methodologies pushes the current technologies towards several new applications, by producing complex side-effects on the everyday envi-*

* Su determinazione della direzione, il contributo è stato sottoposto a referaggio anonimo in conformità all'art. 15 del regolamento della Rivista

*ronment, in particular at the socio-economical, juridical and ethic level. In the present paper, we will review the mainstream approaches that have been developed in AI during the years, until the current rise of deep learning. We will then discuss the current trends, their strengths and their limitations and the impact on the socio-economical system related to the adoption of more and more sophisticated intelligent devices and tools.*

## Sommario



## Keywords

Intelligenza Artificiale - apprendimento automatico – deep learning - apprendimento profondo – machine learning

## 1. Introduzione

La disciplina dell'Intelligenza Artificiale (IA) nasce in seno all'informatica (o come si dovrebbe più opportunamente chiamarla scienza dell'informazione) finendo poi per andare ad influenzare e pervadere moltissime altre discipline sia scientifiche che non scientifiche quali ad esempio: le scienze biomediche, l'economia, la sociologia, le scienze cognitive, la giurisprudenza solo per citarne alcune. Sebbene lo sviluppo ed il progetto di sistemi intelligenti richiedano competenze in ambito informatico e matematico-statistico anche molto avanzate, la consapevolezza dell'impatto di tali sistemi e delle tecnologie collegate sta ormai diventando estremamente importante anche nel campo delle scienze umane, ed il diritto non fa certamente eccezione.

Un sottogruppo specifico dell'Intelligenza Artificiale (che sfrutta gran parte sia della matematica che dell'informatica) è il *machine learning* (ML) o apprendimento automatico. L'obiettivo di questa disciplina è di sviluppare sistemi artificiali (che possiamo chiamare agenti) imparando dalle capacità acquisite tramite l'esperienza. Nonostante spesso il ML sia completamente confuso con l'IA, dove essere chiaro che si tratti solamente di una sottocategoria con obiettivi specifici riguardanti l'apprendimento automatico da dati e situazioni. L'IA, al contrario, ha obiettivi molto più ampi che riguardano la costruzione di agenti intelligenti che, una volta apprese le informazioni necessarie (sia dai dati sia perché fornite esternamente in qualche altra forma), sfruttano quest'ultime per svolgere specifici compiti, i quali normalmente richiedono una qualche forma di intelligenza per essere portati a termine. Per esempio, i vecchi "sistemi esperti[1]" erano solitamente costruiti senza alcuna componente di apprendimento; la conoscenza base necessaria a svolgere il compito di riferimento (come ad esempio il suggerimento di una terapia, la scoperta di un deposito minerale, il design della configurazione di un computer, giusto per citare alcune delle reali applicazioni effettuate

[1] P. Jackson, *Introduction to Expert Systems*, Boston, 1998.

con questo approccio) era solitamente costruita tramite un processo manuale chiamato 'ingegneria della conoscenza', dove un informatico doveva lavorare a stretto contatto con un esperto del settore, con l'obiettivo di trasferire parte delle conoscenze dell'esperto nel sistema utilizzando il formalismo adatto. Questo approccio mostrò subito i suoi limiti, evidenziando sempre di più il bisogno di un apprendimento automatico come promesso dagli approcci del ML (ritorneremo su questi aspetti più avanti).
D'altra parte, valutando l'impatto e l'influenza nei confronti delle altre discipline, possiamo notare che, oltre che nell'informatica e nella matematica, l'Intelligenza Artificiale gioca un ruolo di fondamentale importanza in molte altre discipline – ed è da esse in qualche modo anche influenzata –, alcune scientifiche (come nel caso della biologia e delle neuroscienze), altre legate alle scienze umanistiche (come nel caso della filosofia, sociologia e del diritto), altre al confine tra le due (come per le scienze cognitive e la psicologia). Vale la pena notare che l'impatto sopra menzionato è tuttavia bidirezionale: l'IA sta decisamente offrendo nuove possibilità e applicazioni interessanti in tutti i campi sopracitati, ma allo stesso tempo prende da questi qualche spunto e principio. Per esempio, la scienza cognitiva è stata d'ispirazione per il cosiddetto paradigma Cased-Based Reasoning (CBR), uno dei modelli di ragionamento più famosi nell'applicazione dell'IA al diritto. L'idea è di risolvere nuovi problemi, o interpretare nuove situazioni, sfruttando la soluzione o l'interpretazione di problemi già risolti (o interpretati) in passato. Questo metodo di risoluzione dei problemi, che evita di partire da zero ogni volta che si presenta un problema, è un tipico schema cognitivo degli esseri umani che può essere trasferito con successo agli agenti artificiali. Un'altra fonte di ispirazione viene dalle neuroscienze: il nostro cervello, che è lo "strumento" con il quale noi umani sviluppiamo schemi di ragionamento, è composto da miliardi di cellule interconnesse chiamate neuroni; ogni connessione è chiamata *sinapsi* e le informazioni necessarie per ragionare e svolgere attività viaggiano da un neurone all'altro attraverso queste sinapsi sotto forma di segnali elettrici. Anche se non sappiamo effettivamente come questo processo funzioni nel dettaglio, una versione ultra-semplificata del cervello, il cosiddetto modello di *rete neurale*, è una delle metodologie di maggior successo nell'Intelligenza Artificiale moderna e costituisce la base dell'approccio di *machine learning* chiamato *deep learning*. Ciò che è veramente notevole è che, nonostante un neurone artificiale non abbia nulla a che fare con un neurone naturale, la ricostruzione semplificata dell'intera architettura del cervello, come fatta in una rete neurale artificiale, sembra essere in grado di replicare alcuni compiti importanti come la classificazione degli oggetti, il riconoscimento delle immagini, l'interpretazione delle frasi, e via dicendo. Tutto si riduce essenzialmente ad una questione di "manipolazione di matrici", in altre parole una serie di somme e moltiplicazioni di numeri reali (ovvero entità che un computer può facilmente maneggiare in maniera molto più efficiente di quanto non lo possa fare un essere umano).
Il cammino che porta a ciò che vediamo oggi ebbe inizio nel 1956, quando un gruppo di ricercatori, tra cui John McCarthy e Marvi Minsky, organizzò il Darmouth Summer Research Project on Artificial Intelligence, il quale viene attualmente considerato come la nascita ufficiale della disciplina, e il momento in cui il termine "*Artificial Intelligence*" fu coniato. I primi passi dell'IA erano essenzialmente mossi da una grande eccitazione

per la possibilità di sviluppare programmi in grado di mostrare un comportamento intelligente al livello di quello umano. Tuttavia, le previsioni ottimistiche iniziali riguardo questa possibilità dovettero immediatamente confrontarsi con la dura realtà dei fatti: comprendere e replicare l'intelligenza umana è un problema estremamente difficile che coinvolge la definizione di metodologie sofisticate, così come la disponibilità di risorse computazionali adeguate. La storia dell'IA si è così sviluppata con una serie di alti e bassi nelle cosiddette "*AI seasons*", dove a giorni di primavera (con molte scoperte e progetti attivi) seguirono giorni d'inverno (dove le scoperte erano quasi nulle e la delusione per i risultati molto profonda).

Oggi, stiamo vivendo un'estate molto soleggiata, con un rinnovato interesse per i metodi di IA e applicazioni mai viste prima. Ciò è essenzialmente dovuto alla disponibilità di tre risorse principali: nuove metodologie per la costruzione e il perfezionamento dei modelli di IA (da ricerche basilari in informatica e nelle relative discipline), un'enorme disponibilità di dati di vario tipo (testi, immagini, suoni, dati strutturati, ecc.) e infine risorse computazionali su larga scala e ad alta prestazione. Lo sviluppo del cosiddetto approccio *deep learning*[2] ci permette di affrontare problemi molto difficili che richiedono capacità complesse di ragionamento come nel supporto alle decisioni, di comportamento reattivo come nei dispositivi autonomi, di comportamento interattivo come nei dispositivi di assistenza personale che comprendono il linguaggio naturale, e così via.

L'impatto di tali metodologie spinge le attuali tecnologie verso nuove applicazioni, producendo effetti collaterali complessi sull'ambiente di tutti i giorni, in particolare a livello socioeconomico, giuridico ed etico. Nel presente articolo, analizzeremo gli approcci tradizionali sviluppati nell'IA nel corso degli anni, fino all'attuale ascesa del *deep learning*. Discuteremo dunque poi delle attuali tendenze, dei loro punti di forza, delle loro limitazioni e dell'impatto sul sistema socioeconomico relativo all'adozione di strumenti e dispositivi intelligenti sempre più sofisticati

## 2. Le stagioni dell'Intelligenza Artificiale

Fin dagli albori, l'Intelligenza Artificiale ha fatto sorgere un quesito fondamentale: cosa ci si aspetta da un sistema artificiale, affinché possa essere definito "intelligente"? Di fatto, come notato da molti ricercatori, il problema con l'IA sta nel nome stesso: se da una parte siamo abbastanza sicuri del significato della parola "artificiale" (ovvero qualcosa che è stato costruito da un essere umano), dall'altra non abbiamo una definizione precisa del termine "intelligenza". Questo perché attribuiamo talmente tante sfaccettature all'intelligenza umana, che è diventato davvero difficile condensarle tutte in un'unica definizione.

Le principali scuole di pensiero ad aver preso piede sono state infatti due: la cosiddetta "IA debole" (*weak AI*) e la cosiddetta "IA forte" (*strong AI*). Secondo la *strong AI*, il computer non è soltanto uno strumento nello studio della mente, ma piuttosto, come affermato da John Searle, «il computer opportunamente programmato è davve-

[2] I. J. Godfellow - Y. Bengio - A. Courville, *Deep Learning*', Boston, 2016.

ro una mente, nel senso che si può letteralmente dire che i computer, con i programmi giusti, comprendono e hanno altri stati cognitivi»[3] . Per dimostrare l'impossibilità dei programmi informatici di raggiungere questo livello di cognizione (incluso l'avere coscienza), Searle immaginò un esperimento mentale chiamato "La Stanza Cinese" ("*Chinese Room*"). Si consideri un programma che prende come input una sequenza di caratteri cinesi e, attraverso una serie di regole specifiche e molto complesse, sia in grado di produrre altri caratteri cinesi come output. Il programma è talmente sofisticato che potrebbe superare il cosiddetto test di Turing (potrebbe essere confuso con un essere umano, se coinvolto in una conversazione con un altro essere umano)[4] . L'ipotesi sollevata da Searle è che il computer non capisca effettivamente ciò che sta facendo, ossia che non abbia conoscenza esplicita del suo compito. A sostegno di ciò, immagina se stesso dotato della versione inglese delle istruzioni del programma di cui sopra; se ricevesse una sequenza di caratteri cinesi, sarebbe in grado di produrre un'altra sequenza di caratteri cinesi seguendo le istruzioni e simulando una conversazione in cinese. Tuttavia, lui "non parla una singola parola di Cinese" e quindi non capisce il cinese; allo stesso modo, anche il programma non è in grado di comprendere effettivamente il cinese.

La *strong AI* non è stata seriamente considerata e investigata in informatica, poiché è stata data una maggiore attenzione agli aspetti pratici riguardanti la costruzione di sistemi in grado di mostrare un comportamento che potesse essere considerato dai più come intelligente. Questa è esattamente la questione sollevata dall'ipotesi della *weak AI*: l'obiettivo è di sviluppare programmi (software) in grado di mostrare comportamenti intelligenti in compiti ristretti e ben specifici. Un programma in grado di diagnosticare una specifica malattia in un particolare campo della medicina può essere considerato intelligente (secondo la *weak AI*), anche se lo stesso programma non è in grado di comprendere una frase in qualunque linguaggio naturale, di riconoscere un qualsiasi oggetto fisico o di giocare ad un semplice gioco come tris.

Inoltre, una chiara caratteristica di ciò che consideriamo come comportamento intelligente è la capacità di imparare dalle esperienze. Ciò ha dato origine ad un importante sottocampo dell'Intelligenza Artificiale, ovvero il Machine Learning (ML). Tuttavia, la costruzione di sistemi intelligenti (deboli) e l'investigazione delle capacità di apprendimento hanno solitamente seguito percorsi separati, con diversi sistemi di IA tradizionale (che possiamo chiamare sistemi "*knowledge-based*") costruiti senza una componente di apprendimento, o con una componente di apprendimento non strettamente integrata nel sistema stesso ma utilizzata come modulo indipendente. La somiglianza che possiamo vedere al giorno d'oggi in molti documenti e articoli tra IA e ML non è dunque propriamente giustificata, sia perché IA è un termine più generico di ML, sia per i motivi storici sopracitati.

Per quanto riguarda i primi anni dell'IA, essi sono stati caratterizzati da quello che John McCarthy chiamò il periodo di "guidare la bici senza mani" (*Look Ma, no hands*). Questo periodo fu contraddistinto da un grande ottimismo e da aspettative molto alte in ciò che l'IA era in grado di realizzare. Alcune delle figure più illustri del momento

[3] J.R. Searle, *Minds, Brains and Programs*, in *Behavioral and Brain Sciences*, 3(3), 1980, 417 ss.

[4] A.M. Turing, *Computing Machinery and Intelligence*, in *Mind,* 59, 1950, 433 ss.

come Herbert Simon (premio Nobel per l'economia nel 1978 e premio Turing nel 1975) e Marvin Minsky (premio Turing nel 1969) si spinsero addirittura a fare le seguenti previsioni:

- «Le macchine saranno capaci, nell'arco di trent'anni, di fare qualsiasi lavoro un uomo può fare» (H. Simon, 1965)
- «Nel giro di una generazione […] il problema di creare un'intelligenza artificiale sarà sostanzialmente risolto» (M. Minsky, 1967)
- «Tra tre-otto anni avremo una macchina con l'intelligenza generale di un essere umano medio» (M. Minsky, 1970)

Nessuna di queste aspettative fu effettivamente soddisfatta, poiché i problemi legati al rendere l'IA un successo erano molto più complicati da trattare di quanto ci si aspettasse inizialmente.

L'IA iniziò un ciclo di "stagioni" che alternava delusioni e fallimento ad entusiasmo e successo. Il primo inverno dell'IA iniziò negli anni '70, quando divenne chiaro che sviluppare sistemi in grado di sfruttare la conoscenza umana fosse effettivamente molto difficile e quando i limiti degli approcci basati sulla logica formale (approcci predominanti al tempo) divennero a loro volta evidenti (in particolare in situazioni che implicavano l'utilizzo di un sapere incerto e il ragionamento in condizioni di incertezza). Inoltre, una scoperta (conosciuta come il paradosso di Moravec) ebbe un impatto non indifferente sul settore: contrariamente ad alcune supposizioni tradizionali, l'elaborazione sensoriale di base e la percezione richiedono significativamente più risorse computazionali rispetto ai processi di modellamento del ragionamento ad alto livello. In altre parole, è più semplice costruire un sistema in grado di giocare ad un gioco intelligente come gli scacchi a livello di un campione umano, piuttosto che costruire un sistema con le stesse capacità sensoriali di un bambino di due anni.

Il riconoscimento di tali limiti ebbe come conseguenza la riduzione della portata delle metodologie di IA e questa fu in realtà una buona scelta, dal momento che ha focalizzato maggiormente l'attenzione sull'ipotesi più pratica (e in qualche modo più semplice) della *weak AI*. Infatti, gli anni '80 divennero quella che fu chiamata la prima primavera dell'IA. Le ricerche nell'IA si focalizzarono su specifiche architetture denominate "sistemi esperti" (*expert systems*); si trattava di sistemi che mostravcano competenze a livello di esperti umani, ma solamente in aree molto ristrette (come, ad esempio, la diagnosi di classi limitate di malattie, il suggerimento di specifiche terapie antibiotiche, la scoperta di depositi minerali, la determinazione della configurazione ottimale di un computer per un cliente specifico, e così via). Era, se vogliano, la vittoria della *weak AI* sulla *strong AI*. Sistemi esperti come MYCIN , CADUCEUS , PROSPECTOR , R1 divennero i rappresentanti più importante dei cosiddetti rule-based systems, ovvero sistemi con un paradigma di ragionamento consistente nella formalizzazione delle competenze di risoluzione di problemi o compiti specifici in un set di regole if-then (se-allora); tali regole permettono al sistema di testare la verità di una specifica condizione (la parte *if*), e se questa condizione si verifica, il sistema può trarne una conclusione (la parte *then*). L'idea era quella di avere un modo meccanico e possibilmente efficiente di simulare una forma di ragionamento deduttivo (ovvero da una premessa, ad uno o più passi intermedi ed infine ad una conclusione finale).

Nello stesso periodo, in Giappone, fu iniziato il cosiddetto progetto dei "Computer di Quinta Generazione", con l'obiettivo di utilizzare la programmazione logica basata su PROLOG come strumento principale per costruire i sistemi di IA. Il PROLOG è un linguaggio di programmazione nato in Europa[5] che mira ad implementare il tipo di ragionamento *rule-based* sopracitato, ma con una semantica più formale basata su uno specifico (e limitato) tipo di ragionamento logico: la risoluzione con le clausole di Horn. A quel tempo il principale linguaggio di programmazione utilizzato per sviluppare sistemi *knowledge-based* era il LISP, un linguaggio proposto da John McCarthy che divenne presto lo standard (specialmente in USA) nell'industria dei sistemi esperti. Alcune aziende produttrici di computer iniziarono addirittura a costruire e commercializzare architetture di computer *ad hoc* chiamate Lisp Machines. Erano di fatto computer "*general purpose*", ma progettati per eseguire efficientemente il linguaggio LISP come loro principale software e linguaggio di programmazione, solitamente attraverso uno specifico supporto hardware. Il LISP è stato uno dei primi linguaggi funzionali (influenzato dal lambda calcolo di Alonzo Church) e le sue istruzioni primitive per la manipolazione di liste (il nome è l'acronimo di LISt Processor) lo rese estremamente adatto alle operazioni di manipolazione simbolica richieste dai sistemi di IA di quel tempo. Il progetto dei Computer di Quinta Generazione era tra le altre cose la risposta dei Giapponesi alle operazioni di business condotte dalle aziende americane. Nonostante il fallimento dell'originale progetto giapponese, PROLOG è ancora adottato in molti sistemi di IA, specialmente con le estensioni fornite dalle versioni moderne, che permettono di svolgere in maniera efficiente compiti di ottimizzazione combinatoria come la programmazione di orari, l'allocazione di risorse, la progettazione e la pianificazione.

Infine, verso la fine del decennio, la riscoperta dell'algoritmo di "*backpropagation*"[6] per l'apprendimento dei parametri delle reti neurali fu la chiave per la rinascita del cosiddetto connessionismo o approccio sub-symbolico all'IA, ovvero l'uso di modelli basati sull'elaborazione di informazioni non simboliche, e su una metafora delle connessioni tra neuroni nel cervello umano.

Tuttavia, l'approccio dei sistemi esperti mostrò velocemente i suoi limiti provocando un altro inverno dell'IA durato fino alla metà anni '90. Il problema principale ad essere ancora irrisolto era "l'ostacolo dell'acquisizione della conoscenza". La fonte principale per far funzionare un sistema esperto è infatti la cosiddetta "base di conoscenza", un deposito di conoscenza relativo a uno specifico ambito utilizzato dal sistema per portare a termine il proprio compito. Il contenuto della base di conoscenza doveva essere acquisito manualmente, tramite un'attività collaborativa tra un esperto del settore /applicativo ed un ingegnere della conoscenza (lo scienziato esperto di metodi di IA, ma tipicamente digiuno delle competenze di dominio applicativo). Questa fase si dimostrò la più difficile nello sviluppo di un sistema esperto ed insieme alla mancanza di metodi e strumenti pratici di *machine learning* per l'apprendimento automatico del-

[5] A. Colmerauer - P. Roussel, *The birth of Prolog*, in *ACM Special Interest Group on Programming Languages (SIGPLAN) Notices*, 28 (3), 1993, 37 ss.

[6] D. E. Rumelhart - G. E. Hinton - R. J. Williams, *Learning representations by back-propagating errors*, in *Nature,* 323, 1986, 533 ss.

la conoscenza necessaria, le difficoltà nel processo di acquisizione della conoscenza segnarono la fine dell'era dei sistemi esperti. Anche i modelli connessionistici, che potrebbero sfruttare in modo più diretto i dati grezzi per l'apprendimento dei parametri di sistema, furono rapidamente messi alla prova dalla complessità dei problemi del mondo reale, e le loro limitate capacità di apprendimento dell'epoca non erano in grado di offrire soluzioni pratiche. Il mercato delle "LISP machines" iniziò a collassare e molte aziende produttrici di computer che avevano fatto investimenti rilevanti nel business dei sistemi esperti dovettero cambiare direzione, ritornando ad applicazioni più tradizionali.

Nuovi interessi nelle metodologie dell'IA (e una conseguente nuova primavera dell'IA) emersero negli anni '90, quando i metodi probabilistici divennero uno strumento ampiamente usato. Un'importante pietra miliare fu la pubblicazione di un famoso libro di Judea Pearl[7] in cui si prospettava l'uso del calcolo delle probabilità per trattare sia la modellazione che l'inferenza nei sistemi intelligenti. L'uso delle probabilità era stata rigettato da alcuni dei padri dell'IA (John McCarthy e Patrick Hayes) in quanto "epistemologicamente e computazionalmente inadeguato" per l'Intelligenza Artificiale. Pearl difese la sua visione dimostrando che una interpretazione coerente della teoria probabilistica era possibile attraverso un formalismo basato su grafi e chiamato Probabilistic Graphical Models, il cui principale rappresentate sono le Reti Bayesiane (Bayesian Networks). L'utilizzo di tali formalismi permette una modellazione della conoscenza incerta compatta ed efficiente, e l'uso di algoritmi di inferenza specializzati per rispondere ad ogni tipo di quesito probabilistico; inoltre, la capacità di apprendere sia la struttura che i parametri del modello grafico, resero possibile superare molte limitazioni dei sistemi logic-based e aprirono la strada ad applicazioni più reali. Un'evidenza di ciò è il fatto che all'inizio del nuovo millennio, un'azienda come la Microsoft assunse alcuni dei più rinomati ricercatori nel campo delle Reti Bayesiane (in particolare David Heckerman, Jack Breese e Eric Horvitz) per aprire una nuova divisione di ricerca sull'Intelligenza Artificiale e il Machine Learning. Questo team studiò e realizzò alcune delle applicazioni di maggior successo del periodo comprendente il primo filtro di spam al mondo basato sul *machine learning*, the *Answer Wizard* (che divenne il *back-end* per Clippy, il piccolo avatar che aiutava gli utenti durante le loro attività sulla suite di Microsoft Office), il Windows Printer Troubleshooters e la prima piattaforma di *machine learning* di Microsoft, ora rappresentata da Azure. Questi stessi ricercatori rilasciarono anche uno dei primi strumenti grafici per la costruzione e il ragionamento con le Reti Bayesiane, lo strumento MSBN (MicroSoft Bayesian Network). I *Probabilistic Graphical Models* e le metodologie delle reti Bayeisane iniziarono ad essere incorporati in molti sistemi da svariate aziende, nonostante il crollo delle precedenti aspettative sviluppatesi nell'era dei sistemi esperti avesse messo fine alla pubblicizzazione di certi sistemi in quanto AI-based. Ciononostante, il rapido successo e l'impatto positivo delle tecniche probabilistiche furono la causa di una rinascita dell'interesse per i sistemi intelligenti e per l'IA in generale. Allo stesso tempo il *machine learning* divenne sempre più in grado di affrontare problemi reali grazie l'introduzione

[7] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Burlington, 1988.

di diversi metodi statistici, come le *Support Vector Machines* (SVM)[8] e gli approcci di *Ensemble Learning*[9]. Gli approcci di tipo *ensemble* implementano la cosiddetta idea della "saggezza della folla": se un dato "processo" è in grado di eseguire una data previsione con una particolare performance, allora considerare diversi processi che eseguono la stessa previsione può in linea di principio aumentare la performance complessiva nel compito in oggetto. Il "processo" può essere differenziato utilizzando algoritmi o dati diversi. Per esempio, è possibile eseguire una serie di algoritmi diversi – i quali producono modelli di previsione differenti – sulla stessa serie di dati, e poi inviare la previsione fornita dalla maggioranza dei modelli, possibilmente pesando il risultato in base alla rispettiva sicurezza della previsione, in modo che i modelli che prevedono un risultato con un'alta sicurezza abbiano un peso maggiore nel produrre una risposta finale. Al contrario, un approccio ensamble diverso potrebbe essere quello di utilizzare più volte lo stesso algoritmo o modello, ma con una serie diversa di dati. Questo tipo di approccio è strettamente legato ad alcune metodologie computazionali statistiche, come ad esempio il rinomato metodo *bootstrap*, il quale costituisce un modo efficace per migliorare la performance finale di un modello di previsione. La performance è solitamente misurata attraverso la metrica di *accuratezza*, che consiste nella percentuale delle previsioni corrette rapportata all'intera serie di previsioni fornite. Ottenere opinioni diverse sulla stessa serie di dati oppure la stessa opinione su un insieme di dati simili è il metodo utilizzato dall'approccio ensamble per incrementare la precisione finale.

A questa nuova stagione "primaverile" venne dato il merito per aver coinvolto nell'Intelligenza Artificiale e nel *machine learning* ricercatori e professionisti provenienti da diversi settori. Essa si è poi evoluta circa dieci anni fa in una vera e propria estate che diede inizio ad un'era completamente nuova, dominata dallo sfruttamento dei cosiddetti "big data" e testimoniata da un numero crescente di risultati di successo in aree molto diverse.

Nella prossima sezione ne verranno trattati alcuni dettagli.

## 3. L'era dell'apprendimento profondo

Nel 2010, la rivista *The Economist* uscì con un titolo significativo in copertina: *The data deluge* (Il diluvio di dati). La copertina riportava un uomo con un ombrello capovolto sotto un diluvio di dati. L'ombrello capovolto raccoglieva parte di questi dati e una pianta veniva annaffiata con la pioggia raccolta. L'era dei big data stava iniziando e l'IA divenne effettivamente un modo per sfruttare un'enorme quantità di dati resa disponibile sia dai dispositivi elettronici che dalle attività umane.

La disponibilità di questi big data è tuttavia solamente una delle fonti dell'attuale successo dei metodi di IA; altri aspetti fondamentali sono stati lo sviluppo di nuovi specifici formalismi e la disponibilità di risorse computazionali (come le *multi-core* CPU o le

[8] VC. Cortes - V. Vapnik, *Support-vector Networks*, in *Machine Learning,* 20, 1995, 273 ss.

[9] D. Opitz - R. Maclin, *Popular ensemble methods: an empirical study*, in *Journal of Artificial Intelligence Research,* 11, 1999, 169 ss.

GPU) che sono oggi molto più performanti di quanto non lo fossero qualche anno fa. Per quanto riguarda le questioni della modellizzazione, le vecchie reti neurali sono state estese a nuovi modelli chiamati *deep neural networks*, dando così origine ad una nuova serie di approcci chiamata *deep learning*. Possiamo dire che le metodologie di *deep learning* rappresentino uno specifico sottoinsieme del *machine learning*, in cui i modelli basati sulle architetture delle reti neurali sono estesi in modo tale da poter trattare dozzine se non centinaia di strati diversi di neuroni. Prima dell'era del *deep learning*, qualsiasi sforzo di apprendere i parametri delle reti con più di qualche strato nascosto era proibitivo e i tentativi di farlo erano destinati a fallire. Oggi, l'introduzione di metodi specifici per gestire alcuni problemi di apprendimento (come i problemi della scomparsa o dell'esplosione del gradiente che fanno in modo che l'algoritmo di apprendimento smetta di apprendere dai dati), insieme alla disponibilità di risorse computazionali molto performanti (in particolare, le nuove unità di elaborazione grafica o GPU, originariamente pensate per la lavorazione delle immagini, e oggi utilizzate anche per la computazione generale) permettono la costruzione e l'apprendimento di modelli molto profondi.

Inoltre, il *deep learning* permette di affrontare un'altra importante questione relativa ad ogni approccio di ML: l'estrazione di caratteristiche. Prima che un modello di ML possa essere utilizzato, le principali caratteristiche del problema da risolvere devono essere estratte dai dati a disposizione. La serie di attributi pertinenti al problema viene solitamente estratta manualmente dai dati prima di costruire il modello di ML; ciò significa che uno specifico compito riguardante l'ingegneria della feature potrebbe essere dato in carico all'analista. Al contrario, gli strati nascosti di una rete profonda sono in grado di estrarre automaticamente dai dati grezzi le caratteristiche rilevanti. Un esempio ci è fornito dal modello CNN (*Convolutional Neural Network*) utilizzato per l'interpretazione delle immagini. Data una serie di pixel rappresentanti un'immagine, ogni strato è in grado di estrarre specifiche caratteristiche dell'immagine come gli angoli e parti molto più specifiche della figura, fino a che la rete non è in grado di riconoscere quale sia l'immagine di input. Non c'è nessun bisogno di dividere l'immagine originale in parti più piccole, poiché viene già fatto dalla rete stessa. Un concetto correlato con questo è quello del feature embedding; ciò significa trovare una rappresentazione numerica adeguata degli oggetti (immagini, frasi od ogni altro tipo di segnale) con i quali si ha a che fare. Dal momento che alla fine ogni oggetto verrà rappresentato come un vettore di numeri, alcune operazioni numeriche, con le quali i computer hanno molta dimestichezza, possono essere effettuate per portare a termine compiti molto complessi come l'individuazione degli oggetti, la classificazione e la segmentazione delle immagini, l'interpretazione delle frasi nel linguaggio naturale, forme complesse di recupero dei dati, e così via. Per esempio, il successo dei moderni chatbot (agenti artificiali intelligenti in grado di intrattenere una conversazione con gli umani in un linguaggio naturale) è essenzialmente dovuto al fatto che, grazie al processo di *word embedding*[10], vengono a formarsi gruppi di parole simili e relazioni tra parole; queste relazioni vengono poi sfruttate per attribuite un ruolo specifico alle parole nella frase e determinare

[10] D. Jurafsky - H. M. James, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, chapter 6, Hoboken, 2000.

il significato della frase stessa.

Il grande interesse nell'approccio del *deep learning* è infatti giustificato dagli ottimi risultati ottenuti in vari ambiti di applicazione, dove però molto spesso, ciò che è importante è solo il risultato finale. Tuttavia, ci sono situazioni in cui è necessaria anche una spiegazione dettagliata della risposta finale. È il caso del supporto alle decisioni dove, per far sì che il suggerimento del sistema venga ragionevolmente accettato, deve essere fornita anche una spiegazione del perché quella sia la risposta.

*Explainable AI* (o XAI) è una nuova *buzzword* che cerca di affrontare le questioni relative alla costruzione di una spiegazione ragionevole per la risposta data da un sistema. Come nel caso del *deep learning* (dove i modelli base di rete neurale erano già disponibili molti anni prima), anche per XAI le tematiche relative non sono cose totalmente nuove. Infatti, fornire spiegazioni appropriate era uno dei compiti principali richiesti ad un sistema esperto; poiché tali sistemi erano solitamente basati su una serie di regole da applicare ai dati a disposizione, la catena di regole attivata durante il processo di ragionamento era alle volte considerata una spiegazione per la conclusione alla quale si era arrivati. Tuttavia, estrarre una serie di regole da una rete neurale non è facile e non è nemmeno del tutto chiaro come affrontare il problema della spiegazione in questo tipo di scenario. La combinazione di approcci simbolici e sub-simbolici sta dunque diventando di grande interesse nell'IA.

Un'ulteriore possibile insidia degli approcci di *deep learning* è la possibilità di essere "ingannati" da particolari tecniche di *adversarial machine learning*[11]. Ad esempio, dato un classificatore di immagini con un'elevata precisione nell'individuare gli oggetti in un'immagine, una minuscola perturbazione nei pixel dell'immagine può produrre previsioni completamente differenti, anche se all'occhio umano l'immagine appare immutata. Un famoso esperimento dimostrò che una rete neurale in grado di riconoscere con grande precisione l'immagine di un panda, l'avrebbe quasi sicuramente identificato come un gibbone, dopo che l'immagine di input era stata "danneggiata" da alcuni piccoli cambiamenti nei pixel, anche se l'immagine sembrava essere completamente immutata all'occhio di qualsiasi osservatore umano. L'immagine "danneggiata" è quello che viene chiamato *adversarial example*. Quest'ultimo sfrutta il modo in cui gli algoritmi di *machine learning* (soprattutto quelli basati sul paradigma del *deep learning*) lavorano, al fine di ingannare il comportamento dello stesso o di altri algoritmi di IA. Negli ultimi anni, l'*adversarial machine learning* è diventata un'area attiva di ricerca dal momento che il ruolo dell'IA continua a crescere in molte delle applicazioni che utilizziamo oggi. Come possiamo facilmente immaginare, questa è una questione delicata da considerare quando si lavora con e si progettano sistemi basati sul *deep learning*, poiché non è difficile ipotizzare un uso malizioso (o addirittura criminale) di queste tecniche in diversi contesti (e.g., in un'applicazione militare o di difesa o in uno scenario legale o addirittura in un contesto oggi molto comune come quello delle macchine che si guidano da sole e che interpretano un segnale di stop come qualcos'altro).

[11] I. Goodfellow - P. McDaniel Patrick - N. Papernot, *Making machine learning robust against adversarial inputs*, in *Communications of the ACM*, 61 (7), 2018, 56 ss.

## 4. Intelligenza Artificiale e diritto

Poiché abbiamo visto che l'IA moderna e il ML stanno avendo grande impatto in molti settori e discipline, non è sorprendente che anche il diritto abbia a che fare con l'IA. Secondo Harry Surden[12], il collegamento tra l'IA e il diritto comporta «l'applicazione di tecniche informatiche e matematiche per rendere il diritto più comprensibile, gestibile, utile, accessibile e prevedibile». Ciò non si discosta molto da quel che ci possiamo aspettare dall'IA applicata ad altre discipline, e infatti le applicazioni dell'IA al diritto hanno una storia abbastanza significativa, dal momento che tali requisiti sono stati spesso proposti come input per i sistemi di IA.

Tuttavia, l'IA ha spesso successo in quegli ambiti e in quelle mansioni che richiedono schemi specifici, oppure regole e risposte ben precise e definite. Ciò è particolarmente vero per l'IA *data driven* (ad esempio: *deep learning*); i modelli possono estrarre informazioni utili dai dati a disposizione, se in essi sono presenti degli schemi (*pattern*), anche considerando il fatto che i dati che non vengono osservati durante la fase di addestramento possono essere inseriti nel sistema quando viene richiesto di fornire delle risposte. L'IA data driven ha successo solo parzialmente, invece, in quegli ambiti che potremmo definite orientati al giudizio, astratti e che coinvolgono persuasione e argomentazione. In questo caso un sistema intelligente, dove i dati sono integrati con informazioni aggiuntive sotto forma di reti semantiche, ontologie e grafi di conoscenza (*knowledge graph*), può giocare un ruolo significativo. Osservando dunque le principali applicazioni che si possono concepire per l'IA nel diritto, è abbastanza chiaro che questa integrazione sia di fondamentale importanza.

Un esempio di una simile integrazione è fornito dal paradigma *Case-Based Reasoning* (CBR)[13], dove i metodi di apprendimento cosiddetto *lazy* vengono integrati con fonti di informazioni specifiche. Una metodologia di apprendimento lazy impara a risolvere casi specifici archiviando tutti i casi passati già risolti dal sistema insieme alla soluzione corrispondente. Quando un nuovo caso deve essere risolto, il sistema recupera dalla sua memoria la serie di casi più simili a quello attuale e utilizza le soluzioni recuperate come base per la soluzione dell'attuale caso da risolvere[14]. Questo tipo di *precedent-based reasoning* ricorda lo schema seguito nei sistemi giudiziari che si affidano ai precedenti per ottenere una sentenza; infatti, proprio negli Stati Uniti molte ricerche si sono concentrate sull'applicazione della metodologia CBR al diritto[15]. Tuttavia, un sistema CBR completo non può basarsi solamente su dati passati e su precedenti, ma ha anche bisogno di informazioni specifiche per poter costruire una soluzione al caso di riferimento partendo da quelle vecchie. Le ontologie, così come le regole e i knowledge graph,

[12] H. Surden, *Artificial Intelligence and Law: An Overview*, in *Georgia State University Law Review*, 35, 2019, in *ssrn.com*.

[13] M.M. Richter - R.O. Weber, *Case Based Reasoning: a textbook*, New York, 2013.

[14] A. Aamodt - E. Plaza, *Case-Based Reasoning Foundational Issues, Methodological Variations, and System. Approaches*, in *AI Communications*, 1994, 39 ss.

[15] K.D. Ashley, *Case-Based Reasoning and its Implications for Legal Expert Systems*, in *Artificial Intelligence and Law,* 1, 1992, 113 ss.

possono essere adottati con successo a questo fine[16]. Nonostante ciò, la realizzazione di questa parte importante di un sistema CBR (chiamato step di revisione o adattamento) può risultare nel dover implementare un vero e prorio sistema *knowledge-based* (in altri termini, un mini sistema esperto). In ambito legale ciò ha spesso implicato che meccanismi alternanti ragionamenti *case-based* e *rule-based* fosse una possibile soluzione[17]. Di fatto, come riportato da Quattrocolo[18], è impossibile trasformare tutte le norme e gli atti in regole matematiche e computazionali, per cui il modello case-based sembrerebbe essere l'opzione migliore; tuttavia, dato che la differenza basilare nel valore del precedente in un sistema *common law* e in uno *civil law* è fondamentale, l'appropriata combinazione di casi con regole e altre forme strutturate di conoscenza è la chiave per l'introduzione delle metodologie di IA in questo ambito.

Un ulteriore aspetto relativo all'uso dell'IA nel diritto riguarda il problema di esercitare il giudizio in condizioni di incertezza. Come precedentemente affermato, una delle ragioni del fallimento dei sistemi puramente logic-based nel primo periodo dell'IA era l'inappropriatezza di tali formalismi nel trattare in modo appropriato questioni di incertezza. Ciò portò alla cosiddetta rivoluzione probabilistica, guidata dall'introduzione delle reti Bayesiane e dal risorgere dell'interesse nel soggettivismo o nelle forme Bayesiane di ragionamento in situazioni di incertezza[19]. Ciò significa che, quando si ragiona in condizioni incerte e con l'obiettivo di fornire un suggerimento o una decisione, un sistema intelligente deve saper sfruttare ogni informazione rilevante in suo possesso oltre che i dati a disposizione; l'uso di informazioni precedenti deve essere integrato nel processo di ragionamento, esattamente come viene fatto nella statistica Bayesiana. La famosa "fallacia dell'accusatore" (*prosecutor fallacy*)[20] è un esempio in cui un sistema di IA correttamente progettato è capace di fornire risposte giuste contrariamente alle erronee conclusioni a cui molti umani (anche potenzialmente esperti) potrebbero arrivare. Supponiamo che una corrispondenza positiva di DNA sia stata ritrovata per un dato sospettato (chiamiamolo Fred) sulla scena del crimine. L'evidenza scientifica suggerisce che la probabilità di avere quel tipo di DNA per un soggetto casuale sia molto bassa, diciamo una persona su mille. Dal momento che Fred ha una corrispondenza positiva, la tesi del pubblico ministero è che Fred sia colpevole, poiché c'è una probabilità molto piccola (1 su 1000) che la corrispondenza sia positiva per caso. Questa argomentazione è un tipo di errore comune che molte persone farebbero; in particolare, l'origine del problema sta nel fatto che la stima di probabilità di innocenza uguale allo

[16] A. Wyner, *An Ontology in OWL for Legal Case-Based Reasoning*, in *Artificial Intelligence and Law*, 16, 2008, 361 ss.

[17] E.L. Rissland - D. B. Skalak, *Combining case-based and rule-based reasoning: a heuristic approach*, Proceedings of the 11th international Joint Conference on Artificial intelligence (IJCAI 89), vol. 1, Detroit, 1989, 524 ss.

[18] S. Quattrocolo, *Artificial Intelligence, Computational Modeling and Criminal Proceedings: a framework for a European legal discussion,* Dordecht, 2020.

[19] Vale la pena notare che, nonostante la parola "bayesiane" nel loro nome, le reti bayesiane possono anche essere interpretate come modelli frequentisti, poiché la loro definizione non si basa su alcuna interpretazione specifica del concetto di probabilità.

[20] W.C. Thompson - E. L. Shumann, *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy*, in *Law and Human Behavior*, 2(3), 1987, 167 ss.

0,1% (1 su 1000) si riferisce in realtà alla probabilità di ottenere una corrispondenza positiva, nel caso in cui Fred venisse considerato innocente. Questo perché, supponendo che Fred sia innocente, allora c'è la probabilità dello 0,1% che Fred dia una corrispondenza positiva di DNA (come affermato pocanzi questa è la probabilità di ottenere una corrispondenza positiva per caso). L'effettiva stima che il pubblico ministero dovrebbe considerare è invece la probabilità che Fred sia innocente, supponendo che la corrispondenza di DNA sia positiva. Questa è facilmente calcolabile utilizzando la famosa formula di Bayes fornendo anche la probabilità a priori (stimata prima di raccogliere qualsiasi prova sul DNA) che Fred sia innocente (o diversamente, colpevole). Per esempio, se non abbiamo uno specifico motivo di ritenere Fred colpevole, se il crimine è stato commesso in una comunità di 10.000 persone, possiamo facilmente calcolare la probabilità che Fred sia innocente, supponendo che la corrispondenza di DNA sia positiva, come circa del 91%. Ciò demolirebbe completamente il quadro accusatorio basato su una chiara fallacia nel processo di ragionamento. Un agente di IA che implementa correttamente questo tipo di ragionamento Bayesiano in condizioni di incertezza non sarebbe esposto a simili fallacie. Inoltre, l'effetto delle informazioni a priori è molto spesso omesso o sopravvalutato nel comune ragionare delle persone. Se Fred è un buon cittadino senza precedenti – e questa è la situazione della maggior parte delle persone nella comunità di riferimento – allora associare una distribuzione a priori uniforme per ottenere la probabilità che Fred sia colpevole può essere ragionevole; tuttavia, se conosciamo una storia di precedenti per crimini simili commessi da Fred, allora la probabilità a priori dovrebbe accuratamente riflettere questa situazione. La difficoltà di ottenere i giusti "*priors"* in uno specifico problema è ancora considerato un campo di ricerca attivo, ma alcuni metodi possono essere utilizzati con successo in molte situazioni pratiche, portando alla progettazione di sistemi di IA in grado di evitare errori simili a quelli appena discussi. È inoltre anche importante sottolineare che questo non si tratta di un problema astratto o "accademico", poiché ci sono stati casi reali in cui questo tipo di fallacia è purtroppo stata riportata, in particolare il caso di Sally Clark[21] in una corte del Regno Unito, e il caso di Lucia De Berk[22] in una corte olandese. Nel primo caso, in cui Sally Clark fu accusata dell'omicidio dei suoi due figli neonati, alcune semplici considerazioni dei principi alla base dei Modelli Grafico Probabilistici (e delle reti Bayesiane in particolare) avrebbero potuto evitare molte conclusioni erronee basate su una serie di presupposti sbagliati, come la mancata considerazione di una causa comune nell'improvvisa morte dei neonati.

Infine, un'altra questione importante che ha a che fare con l'IA puramente data driven è il problema della equità (*fairness)*; un dato algoritmo è considerato equo, o dotato di equità, se i suoi risultati non dipendono da specifiche variabili, specialmente da quelle considerate sensibili, come i tratti degli individui che non dovrebbero essere correlati al risultato (i.e., il genere, l'etnia, l'orientamento sessuale, la disabilità, le preferenze politiche, ecc.). I metodi di *machine learning* che si basano solamente sui dati possono essere in principio influenzati in modo improprio da tali dati, ovvero soggetti a qual-

[21] C. J. Bacon, *The Case of Sally Clark*, in *Journal of the Royal Society of Medicine*, 96(3), 2003, 105 ss.

[22] R.D. Gill - P. Groeneboom - P. de Jong, *Elementary statistics on trial (the case of Lucia de Berk)* , in *Arxiv*, 2019.

siasi errore presente nei dati stessi. Ciò è particolarmente evidente nelle applicazioni dell'IA al diritto. Per esempio, nelle corti statunitensi è abbastanza comune l'uso di algoritmi intelligenti di valutazione del rischio progettati per considerare i dettagli del profilo dell'imputato e restituire un punteggio di recidiva che stimi la probabilità che lui o lei possa reiterare un dato crimine; se l'algoritmo ha appreso il prpprio modello sulla base di dati "tendenziosi", allora questi errori si ripresenteranno nei punteggi che l'algoritmo calcolerà (per esempio incrementando la probabilità di recidiva per specifiche comunità etniche, come riportato dall'organizzazione giornalistica ProPublica[23]).

## 5. Conclusioni

Il presente articolo ha presentato un excursus storico sull'Intelligenza Artificiale e il *machine learning*, cercando di analizzare idee, aspettative, possibilità e limiti delle relative metodologie. Il punto di vista è quello dell'informatica, la scienza di cui l'IA e il ML fanno parte; si è inoltre anche brevemente discusso dell'impatto sociale dei sistemi intelligenti, focalizzando l'attenzione sulle applicazioni dell'IA al diritto. Oggi stiamo vivendo una nuova rivoluzione industriale guidata dall'IA e l'impatto di questa sulla vita quotidiana inizia ad essere percepito. Nuovi profili professionali con competenze interdisciplinari saranno necessari sia per concepire nuove applicazioni sia per fronteggiare questa importante rivoluzione. Per ultimo, ma non meno importante, l'etica e i valori sociali devono essere seriamente presi in considerazione in questo quadro per poter evitare problemi riguardanti la sicurezza, la trasparenza, la spiegabilità e la protezione dei valori umani quali i diritti, le differenze culturali e l'equità giudiziaria. A questo fine, il manifesto di Asilomar per i principi dell'IA dell'istituto Future for Life[24] promuove esattemente dei principi che dovrebbero essere soddisfatti per un' IA che porti dei benefici e non dei danni alle comunit1a umane. Tali principi sono suddivisi in tre aree differenti: ricerca, etica e valori, e problemi a lungo termine. Le questioni di ricerca sono legate agli obiettivi dell'IA (creare un' IA benefica e non un'IA dannosa e senza controllo), al modo in cui i fondi e gli investimenti dovrebbero essere ottenuti, all'attuale relazione tra la scienza (IA) e i decisori politici, all'incentivazione di una nuova cultura di collaborazione e fiducia, e all'evitare il ricorso a scorciatoie per gli standard di sicurezza. Ma l'enfasi del manifesto è principalmente sui principi di etica e dei valori umani che dovrebbero essere promossi e fermamente considerati nel progettare e sviluppare sistemi intelligenti che impattano la nostra vita quotidiana. La sicurezza (un sistema di IA non deve essere pericoloso), la trasparenza del fallimento (bisogna conoscere i motivi di un insuccesso), la trasparenza giudiziaria (ogni processo decisionale automatico deve essere ispezionabile da un'autorità umana competente), la responsabilità dei progettisti e degli sviluppatori, l'allineamento dei valori con quelli umani quali la dignità, la libertà, i diritti e la diversità culturale, la privacy personale, l'autonomia (le applicazioni di IA non devono limitare l'autonomia personale sia reale che anche solamente percepita), la condivisione di benefici e prosperità economica tra

[23] J. Angwin - J. Larson - S. Mattu - L. Kirchner, *Machine Bias*, in *ProPublica,* 23 May 2016.

[24] At *futureoflife.org/ai-principles*

persone, il controllo umano di ogni attività o applicazione dell' IA, la non sovversione (il potere conferito dal controllo di sistemi di IA super avanzati deve rispettare e migliorare, piuttosto che sovvertire, i processi sociali e civici da cui dipende la salute della società), ed infine il mancato ricorso ad armi letali autonome attraverso il cosiddetto *AI arms race*.

Per completare il quadro di analisi, i problemi a lungo termine riguardanti la valutazione e l'attenuazione dei rischi, l'attenta gestione e organizzazione dell'impatto delle applicazioni di IA (specialmente quelle che propongono una vera rivoluzione nelle relazioni, lavori e attività umani), l'evitare forti assunzioni riguardanti i limiti estremi delle future capacità dell'IA (il cosiddetto principio di *capability caution*), la presenza di rigorose misure di sicurezza e controllo relative alle possibili capacità di automiglioramento dei sistemi di IA ed infine il *common good principle*: le applicazioni e i sistemi di IA dovrebbe essere sviluppati solamente al servizio di ideali etici ampiamenti condivisi, e a beneficio di tutta l'umanità invece che di un solo stato o di una sola organizzazione. Prendendo in considerazione ciò che l'IA è stata, cosa attualmente rappresenta e tutti questi principi, ci si augura di progettare e costruire un futuro per l'Intelligenza Artificiale che possa liberarsi di quelle paure spesso associate agli scenari fantascientifici raffigurati da una società in cui gli umani perdono il controllo delle loro attività, o addirittura delle loro vite. In questo scenario degli agenti intelligenti, che non devono necessariamente essere dei robot, ma anche semplici programmi senza interazione fisica con l'ambiente esterno, collaboreranno con gli umani, aiutandoli a risolvere vecchi e nuovi problemi in un modo più efficace e flessibile.