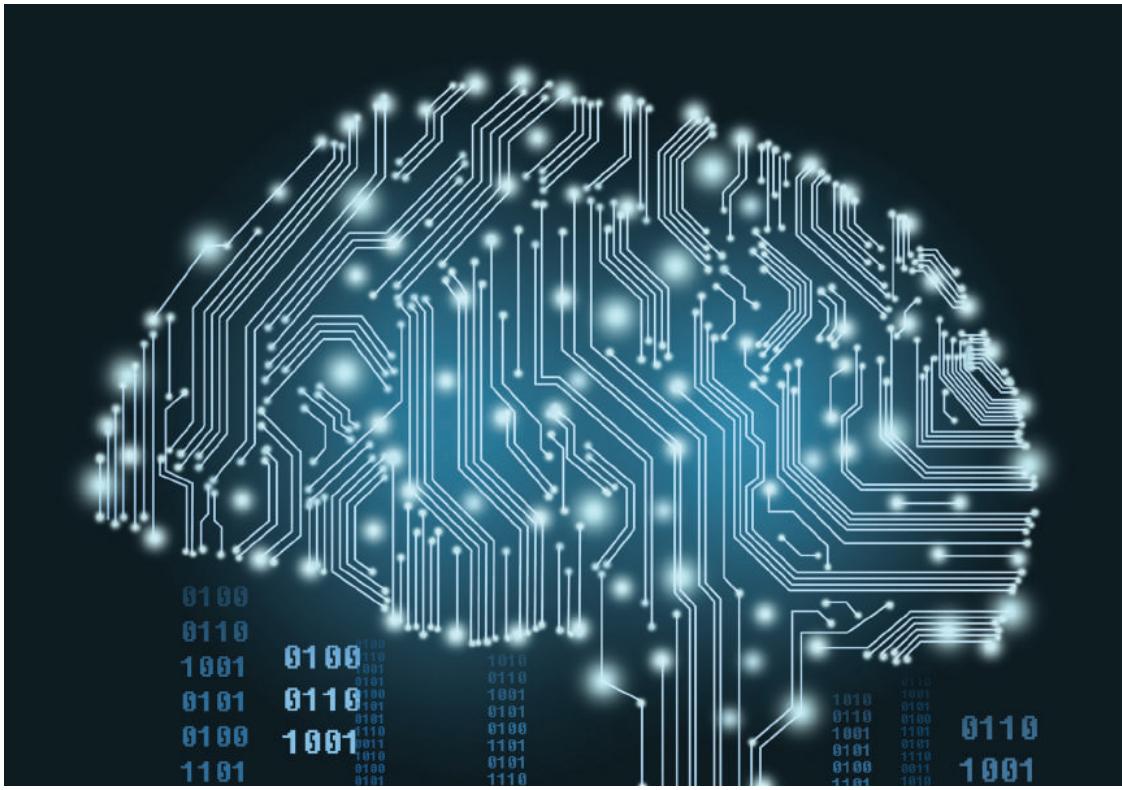
McKinsey&Company

# Intelligenza Artificiale: è tempo di agire

**Advanced Electronics** Dicembre 2017



Gaurav Batra
Andrea Queirolo
Nick Santhanam

# Intelligenza artificiale: è tempo di agire

## Presto l'Intelligenza artificiale cambierà il nostro quotidiano. Le aziende sono pronte a catturare il valore dell'ondata di innovazione in arrivo?

È un peccato per il reparto di radiologia del vostro ospedale locale. Hanno sì delle buone apparecchiature per la risonanza magnetica e un potente software di elaborazione delle immagini. Ma è qui che le macchine si bloccano. Il radiologo deve andare a cercare e leggere il file del paziente, esaminare le immagini, fare una valutazione. E se l'Intelligenza artificiale (IA da ora in poi) potesse far letteralmente decollare il processo, offrendo diagnosi e orientamenti in tempo reale, più precisi, oltre le possibilità dell'occhio umano?

Grazie ai progressi tecnologici degli ultimi anni, in poco tempo i produttori saranno in grado di offrire questo tipo di soluzioni avanzate per la risonanza magnetica. Stanno sondando nuove applicazioni dell'Intelligenza artificiale che, virtualmente, coprono tutti i settori principali: da quello industriale al pubblico. Oggi, grazie ad algoritmi migliori e archivi di dati perfezionati, il tasso di errore di calcolo di un computer nel riconoscimento delle immagini e in altre funzioni cognitive spesso è vicino o inferiore a quello umano. Anche le prestazioni hardware sono migliorate parecchio: le macchine possono elaborare volumi di dati senza precedenti. Questo e altri aspetti sono volani significativi, che aumentano la precisione dei modelli di IA.

Nell'IA, l'Apprendimento profondo (Deep learning, DL) rappresenta l'ambito con il potenziale più alto, ancora inesplorato. (Per maggiori informazioni sui tipi di IA, si veda la barra laterale, "L'evoluzione dell'IA").
Questa tecnologia si basa su reti neurali complesse, che elaborano le informazioni usando varie architetture, composte da livelli e nodi, con funzioni analoghe a quelle dei neuroni cerebrali. Ogni serie di nodi della rete esegue un modello di analisi diverso, consentendo all'Apprendimento profondo di fornire insight molto più sofisticati rispetto ai primi strumenti di IA. Con l'aumentare della sofisticazione, cresce anche la necessità di soluzioni software e hardware all'avanguardia.

Le principali società hi-tech, ben coscienti dell'enorme potenziale dell'IA, hanno mosso i primi passi per risultare vincenti sul mercato. Ma il settore è ancora agli albori e nessuno sembra avere la chiave di volta per il successo. In che modo, allora, le imprese possono catturare il valore e cogliere i frutti di cospicui investimenti in IA?
La nostra ricerca, così come le interazioni con gli utenti finali dell'IA ,suggerisce che sei concetti cardine si dimostreranno illuminanti, una volta cessato il "polverone". Prima di tutto: all'inizio la cattura di valore si limiterà allo spazio dei consumatori e le imprese otterranno maggior valore concentrandosi sui "micro-verticali" aziendali – casi d'uso specifici in settori selezionati. La nostra analisi dello stack tecnologico suggerisce anche che le opportunità varieranno con il livello. Le aziende di maggior successo proporranno soluzioni end-to-end, spesso tramite collaborazioni o acquisizioni. Per alcuni operatori del mondo hardware, l'IA potrà rappresentare una fortunata inversione di tendenza, dopo anni in cui l'interesse degli investitori gravitava intorno ai prodotti software combinando la forte commoditizzazione che ha depresso i margini. Riteniamo che l'avvento dell'IA offra opportunità notevoli, con soluzioni sia nel cloud, sia basate su tecnologia edge, che genereranno una forte domanda negli utenti. Vogliamo soprattutto sensibilizzare le aziende sulla necessità di agire in fretta. Quelle che oggi sono disposte a grandi scommesse, a mettere in discussione le strategie tradizionali, ne usciranno vincitrici.

### Dietro le quinte del mercato dell'IA

Malgrado le aspettative e la pubblicità legate all'IA, questo mercato intimorisce anche gli analisti e gli investitori più audaci. Nel settore non è emersa una definizione standard dello stack tecnologico: è difficile decifrare uno scenario così differenziato, popolato e competitivo. Tra le centinaia di aziende che si contengono la quota di mercato, chi offre cosa?

# L'evoluzione dell'Intelligenza Artificiale

L'intelligenza artificiale (IA) nasce negli anni '50, quando il poliedrico Alan Turing crea un test per stabilire la capacità di una macchina di mimare le funzioni cognitive umane tra cui la percezione, il ragionamento, l'apprendimento e la risoluzione di problemi. Alan assiste alla diffusione dell'Apprendimento automatico (Machine Learning, ML) – dove i sistemi assorbono e "apprendono" dai dati, usando poi la base di conoscenza acquisita per fare previsioni migliori e adottare decisioni nel tempo. Nel 2010, con l'avvento delle reti neuronali profonde, si inaugura l'era dell'Apprendimento profondo.

Ogni soluzione di Apprendimento automatico e Apprendimento profondo comporta due passaggi : training e interferenza. Prendiamo il software installato sulle auto a guida automatica. Per aiutare i sistemi a rilevare eventuali ostacoli sulla strada, gli sviluppatori sottopongono alla rete neurale delle immagini – per esempio, quelle di cani o pedoni – e conducono dei test di riconoscimento. In seguito, si raffinano i parametri di rete, finché la rete neurale sarà capace di una rilevazione visiva molto accurata. Quando avrà visto milioni di immagini e sarà del tutto formata, nella fase di inferenza saprà riconoscere i cani e i pedoni.
Attualmente, il 95 per cento di workload su cloud pubblico è legato all'IA, perché la maggior parte delle applicazioni di IA è ancora relativamente immatura e servono enormi moli di dati per la raffinazione. Con il maturare dei modelli di IA, l'inferenza si farà spazio nel cloud. Si prevede che entro i prossimi tre o cinque anni, il 30-40 per cento di workload su cloud pubblico sarà legato all'inferenza dell'Apprendimento profondo, mentre il training scenderà al 60-70 per cento. L'inferenza prenderà quota anche con lo sviluppo dell'edge computing (all'interno dei dispositivi), dove l'innovazione porterà alla diffusione di chip a basso consumo con prestazioni d'inferenza all'avanguardia.

Per fare chiarezza nello scenario caotico – o almeno, così si direbbe – delle forniture, abbiamo diviso lo stack tecnologico dell'Apprendimento automatico e dell'Apprendimento profondo in sette livelli: servizi, training, piattaforma, interfaccia e hardware (Tabella 1). Alcune aziende concorrono su vari livelli, altre si concentrano solo su uno o due. Più avanti, discuteremo dei potenziali svantaggi per le aziende che condensano le proprie attività solo su livelli specifici.

## Soluzioni edge e cloud

Per tradizione, la maggior parte delle applicazioni di IA si trova nel cloud – un network di server remoti – per le funzioni sia di training, sia di inferenza. Tuttavia, l'inferenza basata su tecnologia edge diventerà una realtà sempre più diffusa per applicazioni dove una latenza di pochi microsecondi è cruciale per la missione in corso. Nei veicoli a guida automatica, per esempio, la decisione di frenare o accelerare deve avvenire con un tempo di latenza vicino allo zero: l'inferenza nell'edge sarebbe l'opzione ottimale.
L'edge computing si rivelerà la scelta più conveniente anche nelle applicazioni, dove la privacy dei dati e la larghezza di banda sono vitali: basti pensare alle tecniche di diagnostica per immagini basate sull'Intelligenza artificiale.

Con l'espandersi dell'edge computing nasceranno opportunità nuove per tutti gli operatori dello stack tecnologico, in particolare per gli sviluppatori di hardware.

## Le nostre convinzioni sul futuro dell'IA

L'IA ha tutte le carte in regola per stravolgere il nostro mondo. Stando alle stime del McKinsey Global Institute, i rapidi sviluppi nel campo dell'automazione e dell'Intelligenza artificiale avranno un impatto importante sul nostro modo di lavorare e sulla nostra produttività. Per catturare valore in questo mercato in crescita, le aziende stanno sperimentando varie strategie, tecnologie e opportunità: tutti questi aspetti richiedono investimenti cospicui. Si respira ancora molta incertezza, ma le aziende che prenderanno in considerazione i punti seguenti avranno più probabilità di successo.

Tabella 1 **Lo stack tecnologico del machine-learning e del deep-learning è fatto di nove livelli distinti**

| Stack tecnologico | | | Definizione | Esempio (i) |
|---|---|---|---|---|
| Servizi | Soluzione e caso d'uso | 9 | Soluzione a problemi usando un modello di apprendimento profondo già testato | Veicoli a guida autonoma (riconoscimento visivo ) |
| Training | Tipo di dati | 8 | Dati presentati al sistema di Intelligenza artificiale (IA) in base a un'applicazione specifica | Etichettati versus non etichettati |
| | Metodi | 7 | Tecniche di ottimizzazione delle ponderazioni dei modelli per i dati forniti dall'applicazione specifica | Non supervisionati, supervisionati, rinforzo |
| Piattaforma | Architecture | 6 | Approccio strutturato per estrarre le caratteristiche dai dati, dato il problema specifico | Rete neurale convoluzionale, rete neurale ricorrente |
| | Algoritmo | 5 | Un insieme di regole che gradualmente modifica il peso della rete neurale per ottenere un'inferenza ottimale, come stabilito dal metodo di training | Back propagation, evolutivo, divergenza contrastiva |
| | Framework | 4 | Pacchetti software per definire le architetture e invocare algoritmi nell'hardware tramite l'interfaccia | Caffe, Torch, Theano |
| Interfaccia | | 3 | Classi nel framework, che determinano e facilitano la comunicazione tra il software e l'hardware sottostante | Compute Unified Device Architecture, Open Computing Language |
| Hardware | Head node | 2 | Unità hardware che organizza e coordina i calcoli degli acceleratori | Unità di elaborazione centrale (Central processing units, CPU) |
| | Acceleratore | 1 | Chip al silicio progettato per le elaborazioni altamente in parallelo richieste dall'IA | Training: Unità di elaborazione grafica (Graphic processing units, GPUs), Field-programmate gate arrays (FPGAs), Application-specific integrated circuits (ASICs)<br>Inferenza: CPU, GPU, ASIC, FPGA |

fonte: interviste a esperti; esame della letteratura

## 1.All'inizio, la cattura di valore si limiterà al settore consumatori

Le prime offerte di IA ai consumatori hanno un tratto in comune: migliorano i prodotti ma non contribuiscono direttamente alla bottom line. La maggior parte viene dai protagonisti consolidati dell'industria tecnologica, tra cui servizi online di traduzione e tagging fotografico, o assistenza vocale digitale su telefoni cellulari. Questi miglioramenti nei prodotti affascinano sicuramente gli utenti – per esempio, possono contribuire ad aumentare il tempo di permanenza su un sito web – ma non generano un incremento diretto nelle vendite o negli utili.

Quando piccole aziende creano offerte simili, spesso le vendite si rivelano limitate o inesistenti, perché i consumatori continuano a essere più attratti dalle soluzioni gratuite.

Ciò detto, i grandi operatori hanno anche un altro vantaggio: possono accedere a un bacino molto più grande di dati utenti –linfa vitale dell'IA. Pertanto, riescono a sviluppare soluzioni di IA dirette ai consumatori più precise e brillanti. Nell'immediato, i prodotti gratuiti dei grandi operatori conserveranno la quota di mercato più grossa, mentre la cattura del valore dell'IA si limiterà al settore consumatori.

Forse, non in futuro: sul mercato stanno comparendo offerte più nuove e a tariffa fissa, come gli assistenti virtuali. Con la prossima ondata di IA per i consumatori, aumenteranno le innovazioni. Le case automobilistiche e altri soggetti, per esempio, introdurranno nuovi prodotti. Se ripensiamo alle auto a guida autonoma, alcuni utenti apprezzeranno quelle in cui l'IA ha il controllo della frenatura automatica, mentre altri vorranno caratteristiche aggiuntive, per esempio sistemi di guida autonoma completa, anche a un costo più alto.

## 2. Le imprese vincenti si concentreranno sui micro-verticali nei settori promettenti

Le nostre prime analisi dei dati ottenuti dal McKinsey Global Institute, combinate con le ricerche e le opinioni degli esperti, hanno evidenziato circa 60 usi distinti di IA, nei maggiori settori. Circa 400 richiedono un certo livello di Apprendimento automatico, 300 necessitano di capacità di Apprendimento profondo. Fra le applicazioni dell'IA più interessanti, molte sono ancora in fase pilota e non implementate su scala. Sono poche le applicazioni di IA che, nei prossimi anni, potranno giovarsi di una domanda elevata, grazie a forti capacità di percezione visiva ed elaborazione:

- I governi possono usare l'IA per scansionare filmati e identificare attività sospette in luoghi pubblici, o applicare algoritmi di IA per rilevare potenziali attacchi cibernetici. Anche molte applicazioni militari, tra cui i droni, si basano sull'IA. Oltre che per la sicurezza, il ruolo dell'IA è cruciale nel controllo del traffico, con sensori e telecamere che consentono di modificare la temporizzazione e la sequenza dei segnali luminosi in base al numero di automobili sulla strada.
- Nel pubblico, le banche iniziano a usare l'IA per rilevare condotte sospette, ad esempio mdodelli che potrebbero nascondere attività di riciclaggio. Gli algoritmi di IA possono anche servire da supporto nell'elaborazione delle transazioni e nei processi decisionali, spesso con una precisione maggiore rispetto agli impiegati umani. Per esempio, gli algoritmi di IA possono rivelare che alcune caratteristiche normalmente ignorate aumentano le probabilità che una transazione sia fraudolenta.
- Nella vendita al dettaglio, l'IA contribuisce già al rilevamento antifurto e potrebbe apportare ulteriori migliorie ai sistemi di cassa automatici. Molti rivenditori al dettaglio stanno sperimentando sistemi che si avvalgono di telecamere e sensori, per rilevare quando i clienti acquistano o restituiscono articoli in negozio. All'uscita dei clienti dal punto vendita, sui loro conti si addebita il totale. Altri rivenditori usano sistemi video interni al negozio per ottimizzare la copertura da parte degli addetti alla vendita. Quando le telecamere rilevano che un cliente è davanti a un espositore, il sistema invita un addetto a prestare assistenza. In futuro, si potrà assistere a ulteriori miglioramenti in quest'area, tra cui sistemi di IA che identificano i clienti con un alto potenziale di acquisto osservando alcune caratteristiche – l'espressione facciale (quale indicatore dell'umore), l'abbigliamento e il numero di accompagnatori. Quindi, potranno allertare gli assistenti alla vendita, comunicando la posizione dei clienti nel punto vendita.

Decidere quali opportunità sondare, tra le centinaia disponibili, è certo un compito arduo per le aziende. Ma un approccio strutturato permette di restringere il campo.
In primis, occorre concentrarsi su un settore. Se da un lato l'esperienza e le capacità di un'azienda potranno influenzare la scelta, dall'altro gli operatori dovranno considerare le caratteristiche del settore, tra cui le dimensioni.

È importante valutare anche il potenziale innovativo (disruption) in un settore, stimato, analizzando il numero di casi d'uso dell'IA, i finanziamenti di capitale alle startup e l'impatto economico totale dell'IA, definito come la misura in cui queste soluzioni hanno ridotto i costi, aumentato la produttività o contribuito in altro modo alla bottom line, in un'analisi retrospettiva delle varie applicazioni. Più una soluzione di IA comporta un vantaggio economico, più sarà probabile che i clienti siano disposti a pagare per usufruirne. La tabella 2 illustra i dati che abbiamo raccolto, suddivisi per 17 settori e in base a specifici parametri dell'IA.

Il valore e la maturità dell'IA variano con il settore. Per esempio, il settore industriale potrebbe trarre enormi vantaggi dall'IA, ma le aziende non sono pronte ad adottare queste soluzioni, cosa che, invece, hanno fatto le controparti nell'automotive. Per i creatori di prodotti e servizi di IA, questo significa che la cattura del valore sarà un processo sfalsato, dove alcuni settori all'inizio saranno più remunerativi di altri.

Tabella 2 **In ogni settore, la richiesta di Intelligenza artificiale dipenderà dalle dimensioni del mercato, dai punti deboli e dalla disponibilità a spendere per le soluzioni.**

| | dimensioni | Punti deboli | | Disponibilità a pagare |
|---|---|---|---|---|
| | Dimensioni globali del settore $ trillion | Casi d'uso dell'Intelligenza Artificiale (AI) # | Aumenti di capitale nelle startup,[1] $ billion | Impatto economico medio dell'IA,[2] % |
| **Settore pubblico e sociale** | 25+ | 50+ | 1.0+ | 5–10 |
| **Vendita al dettaglio** | 10–15 | 50+ | 0.5–1.0 | 5–10 |
| **salute** | 5–10 | 50+ | 1.0+ | 15–20 |
| **Settore Bancario** | 15–25 | 50+ | 1.0+ | <5 |
| **Prodotti Industriali** | 5–10 | 50+ | 0.5–1.0 | 10–15 |
| **Materie Prime** | 5–10 | 10–30 | <0.5 | 15–20 |
| **Articoli di consumo confezionati** | 15–25 | 10–30 | 0.5–1.0 | 5–10 |
| **Automotive e Montaggio** | 5–10 | 10–30 | 0.5–1.0 | 10–15 |
| **Telecomunicazioni** | <5 | 30–50 | <0.5 | 20+ |
| **Oil & gas** | 5–10 | 30–50 | <0.5 | <5 |
| **Prodotti chimici e per agricoltura** | 5–10 | 10–30 | <0.5 | 5–10 |
| **Farmaci & prodotti medicali** | <5 | 10–30 | <0.5 | 20+ |
| **Trasporti e Logistica** | 5–10 | 30–50 | <0.5 | 5–10 |
| **Assicurazioni** | <5 | 30–50 | <0.5 | 15–20 |
| **Media & entertainment** | <5 | 10–30 | <0.5 | 15–20 |
| **Viaggi** | <5 | 10–30 | <0.5 | 5–10 |
| **Tecnologia** | <5 | 10–30 | <0.5 | 10–15 |

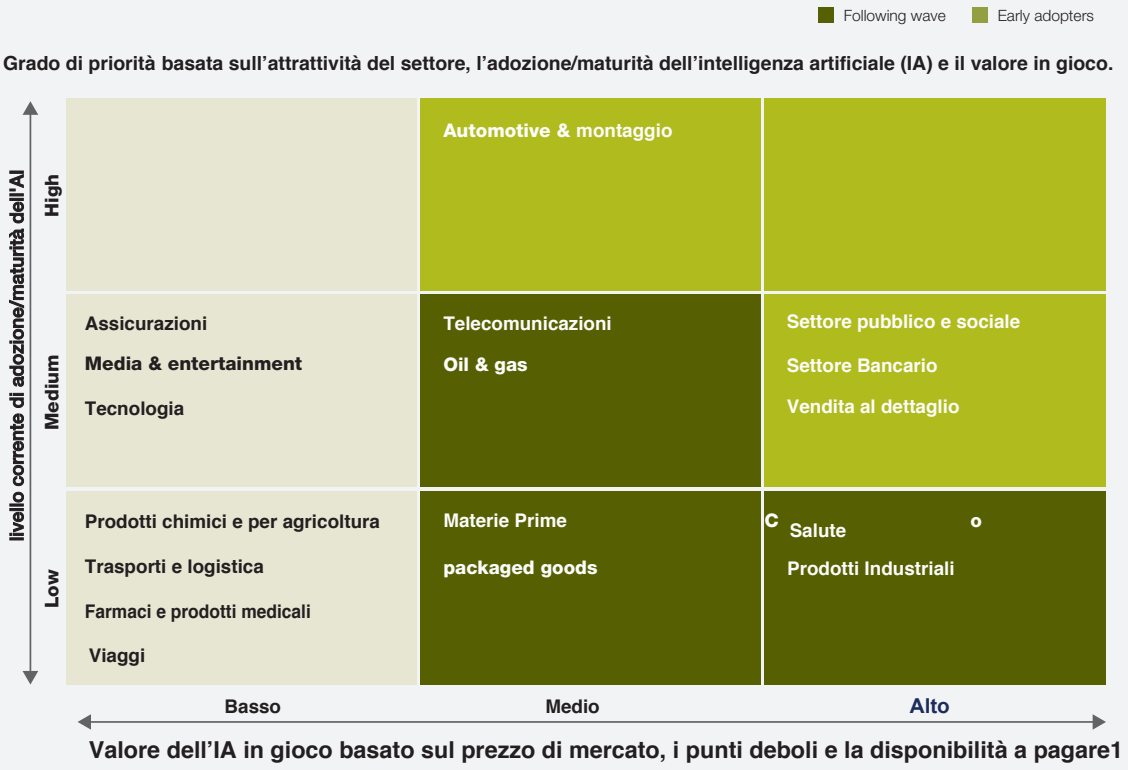
1. Per le startup transettoriali, la quantità di capitale si considera distribuita in base alle dimensioni complessive del settore.
2. L'impatto economico è la somma del valore relativo a tutti i casi d'uso diviso per le dimensioni complessive del settore.

Fonte: Crunchbase; interviste a esperti; IDC; IHS; analisi del McKinsey Global Institute

Combinando il valore in gioco con la maturità, appare chiaro che al momento tanti settori offrono opportunità straordinarie per l'IA: il pubblico, il bancario, quello della vendita al dettaglio e l'automotive (Tabella 3). La prominenza del pubblico potrà sorprendere, in tempi in cui i governi tagliano il budget. Eppure molti funzionari ritengono che l'IA possa migliorare l'efficienza e l'efficacia dei processi e sono propensi ai finanziamenti. Nel progettare le strategie di IA, i fornitori dovrebbero orientare gli investimenti considerando i potenziali utenti delle soluzioni di IA interessati a fare da apripista.

*Microverticali* Una volta scelto uno o più settori cui mirare, le aziende devono fare un altro passo: selezionare casi d'uso particolari – che chiameremo microverticali – su cui concentrarsi. Gli acquirenti non s'interessano all'IA solo perché è una tecnologia nuova ed entusiasmante – al contrario, vogliono che l'IA generi un solido ritorno sui loro investimenti (Return on investement, ROI) risolva problemi specifici, li faccia risparmiare o aumenti le vendite. Per esempio, un impianto di produzione che vuole ridurre il tempo di fermo macchine non cercherà semplicemente un fornitore di IA il cui nome sia noto nello spazio industriale. Al contrario, cercherà un'azienda con una solida esperienza nella manutenzione preventiva, in grado di offrire soluzioni sicure.

Tabella 3 **Il valore dell'Intelligenza artificiale in gioco varia con i settori – come la disponibilità ad adottare le soluzioni.**




1 I punti si sono identificati in base al numero di casi d'uso e ai finanziamenti di capitale alle startup. La disponibilità a pagare si è basata sul valore economico totale dell'IA per un settore.

Fonti: Crunchbase; interviste a esperti; IDC; IHS; analisi del McKinsey Global Institute

Se un fornitore di IA cercasse di proporre una soluzione orizzontale – che i clienti potrebbero applicare in moltissimi casi slegati tra loro — la proposta di valore non sarebbe così accattivante. Gli utenti si chiederebbero se il ROI della soluzione giustifichi la spesa elevata, specialmente se applicata a vari casi d'uso ritenuti di scarso rilievo o trascurabili.

## 3. Per avere successo sul fronte IA, le aziende devono offrire soluzioni end-to-end

Per eccellere nell'Intelligenza artificiale, le aziende devono offrire o organizzare soluzioni end-to-end su tutti i nove livelli dello stack tecnologico: molte aziende clienti, infatti, oggi lottano per implementare soluzioni frammentarie.

Un ospedale, per esempio, preferirà acquistare un sistema che includa sia le apparecchiature per la risonanza magnetica, sia il software di IA per la diagnostica, piuttosto che reperire i componenti separatamente e poi cercare di farli lavorare insieme.

Oltre ad aumentare le vendite, i fornitori di soluzioni end-to-end possono avere un ascendente strategico sui clienti e accelerare l'adozione delle tecnologie in questione. Nvidia, per esempio, offre una piattaforma chiamata Drive PX: di fatto è un modulo, non solo un chip, che offre una soluzione end-to-end per la guida automatica.

La piattaforma unisce processori, software, telecamere, sensori e altri componenti per offrire immagini in tempo reale dell'ambiente intorno a un veicolo. È in grado anche di stabilirne la posizione su una mappa e pianificare un itinerario sicuro da percorrere.

I grandi operatori del mercato hardware e software, spesso ampliano il proprio portfolio di prodotti di IA, muovendosi attraverso tutto lo stack e acquisendo altre aziende. Questo tipo di operazione è frequente in tutti i settori, ma soprattutto nell'IA, dove servono soluzioni end-to-end. Dal 2012, si sono registrate più di 250 acquisizioni di aziende private con esperienza in IA, di cui 37 solo nel primo trimestre del 2017.1 Per competere con questi giganti, molte startup stringono partnership, al fine di posizionarsi quali integratori di sistemi per le soluzioni di IA.

### 4. Nello stack tecnologico dell'IA, la maggior parte del valore sarà legato alle soluzioni o all'hardware

Rispetto allo stack tecnologico dell'IA, la nostra analisi delle prossime tendenze suggerisce che ogni livello genererà direttamente una diversa quantità utile, o valore. Il valore si concentrerà per lo più in due aree (Tabella 4) All'inizio — e in un certo senso, curiosamente, visti gli andamenti del settore — molte delle opportunità migliori verranno dall'hardware (nodi di head, acceleratori di inferenza e acceleratori di training). Si stima che complessivamente queste componenti costituiranno dal 40 al 50 per cento del valore totale per i soggetti che vendono soluzioni di IA.

Exhibit 4 **Most opportunities for monetization will come from the hardware and use case/solution layers of the artificial intelligence stack.**

| Stack tecnologico | Livelli | Fonte di valore in Intelligenza Artificiale |
|---|---|---|
| Servizi | Soluzione e caso d'uso | 40–50% |
| Training | Dati | 0–10% |
| | Metodi | |
| Piattaforma | Architettura | 0% |
| | Framework | |
| | Algoritmo | |
| Inferfaccia | | 0% |
| Hardware | Head node | 40–50% |
| | Acceleratore (training e inferenza) | |

Fonte: Intervista ad Esperti, McKinsey analysis.

In molti altri settori, l'hardware è stato commoditizzato. Ma questa non sarà la tendenza dell'IA nel prossimo futuro, perché gli hardware ottimizzati per risolvere tutti i problemi dei microverticali offriranno prestazioni maggiori, considerando il costo totale di esercizio, rispetto agli hardware di largo consumo, come le unità centrali di elaborazione generiche (CPU). Per esempio, gli acceleratori ottimizzati per reti neurali convoluzionali sono più adatti alla ricognizione delle immagini e, quindi, saranno scelti dai produttori di dispositivi medici. Ma gli acceleratori ottimizzati per le reti di memoria a breve termine sono migliori per il riconoscimento vocale e la traduzione linguistica e, pertanto, attireranno i produttori di sofisticati assistenti virtuali. Ogni caso d'uso ha esigenze leggermente diverse, quindi necessiterà di un hardware parzialmente customizzato.

Secondo un'altra tendenza che si discosta dalla norma, difficilmente il software (inteso come piattaforma e livelli di interfaccia) sul lungo termine sarà l'unico elemento distintivo in ambito IA. Come con l'avvento degli acceleratori dell'Apprendimento profondo, l'hardware da solo o unito al software consentirà di migliorare significativamente le prestazioni: diminuirà, per esempio, la latenza o il consumo energetico.

Pertanto, in questo scenario gli operatori dovranno fare molta attenzione alle scelte di hardware.

Un altro 40- 50 per cento del valore dell'IA verrà dai servizi, comprendenti soluzioni e casi d'uso. Gli integratori di sistemi, che spesso hanno accesso diretto ai clienti, cattureranno la maggior parte dei guadagni unendo le soluzioni, attraverso tutti i livello dello stack.
Nell'immediato, altre aree dello stack dell'IA non genereranno molti utili anche se potrebbero creare del valore indiretto che potrebbe orientare la crescita nel mondo dei consumatori e delle imprese. il che renderebbe questo livello dello stack relativamente più attraente in futuro."

Per esempio, dati e metodi, entrambi elementi di training, attualmente arrivano a costituire al massimo il 10 per cento del valore del fornitore tipico di IA. Ciò accade perché la maggioranza dei dati proviene da utenti di soluzioni di IA anziché da fornitori terzi. Potrebbe infine emergere un mercato dei dati nel mondo dei consumatori e delle aziende,

## 5. Architetture hardware specifiche saranno gli elementi distintivi fondamentali, per il cloud e l'edge computing

Con la crescita dell'IA, l'hardware riacquista fascino, dopo anni in cui gli interessi di investitori e aziende ruotavano intorno all'hardware. Le nostre discussioni con gli utenti finali suggeriscono che vi sarà un forte interesse per le soluzioni sia di edge, sia di cloud computing, secondo il caso d'uso. Il cloud resterà l'alternativa più popolare per molte applicazioni visti i vantaggi di scalabilità.
Nell'hardware nel cloud, le preferenze di clienti e fornitori per la tecnologia ASIC o le unità di elaborazione grafica (GPU) sono variabili: il mercato probabilmente resterà frammentato.
Detto ciò, si considera importante e crescente anche il ruolo dell'inferenza nell'edge computing, laddove una latenza ridotta e le questioni di privacy sono fondamentali o ci sono problemi di connettività. Nell'edge, le tecnologie ASIC saranno più apprezzate dai clienti perché offrono un'esperienza utente più ottimizzata, con consumi energetici inferiori e migliori capacità di elaborazione, per molte applicazioni. Nella periferie (l'edge) delle reti aziendali si assisterà a una sana concorrenza tra Field Programmable Gate Array, GPU e tecnologie ASIC.
Le ASIC potranno comunque presentare un vantaggio, date le loro prestazioni superiori per watt, elemento decisivo in edge computing. Riteniamo che potranno prevalere sulle applicazioni aziendali specifiche, se il livello della domanda sarà sufficientemente alto da giustificarne i costi elevati di sviluppo.

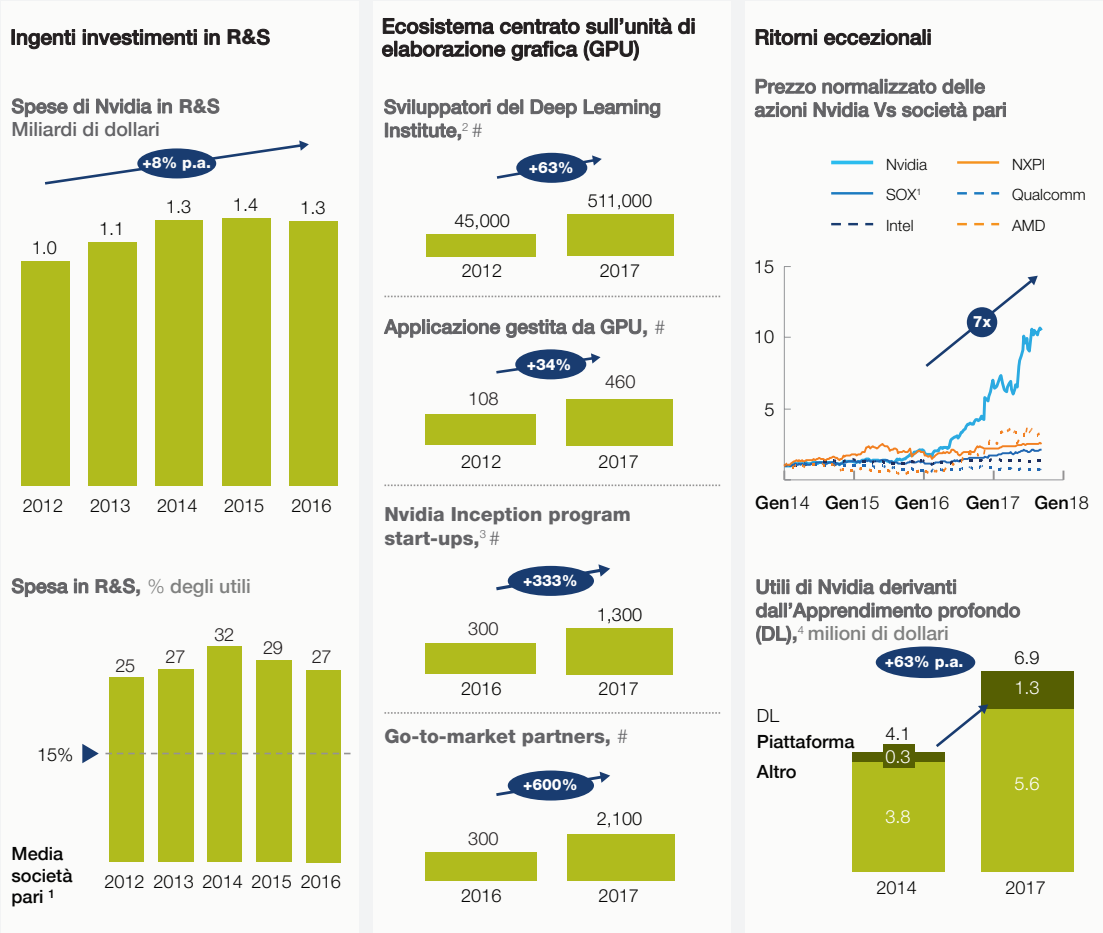
## 6. Il mercato sta già decollando — le aziende devono agire ora e rivalutare le strategie esistenti

Le aziende tecnologiche forse nutrono delle incertezze rispetto all'evoluzione dell'IA, ma riconoscono l'enorme opportunità legata all'Apprendimento profondo e vogliono coglierla.

Alla luce degli sviluppi tecnologici ancora in corso e delle strategie diametralmente opposte attuate dai vari operatori, è ancora difficile dire quale potrà essere la chiave di volta del successo. I protagonisti del mercato stanno già facendo le loro mosse. Le aziende leader vanno in direzioni che sfidano gli schemi comuni.

Per esempio, Nvidia ha aumentato la spesa in R&D per soluzioni di IA di circa l'8 per cento l'anno, dal 2012 al 2016, in cui ammontavano a1,3 miliardi di dollari (Tabella cinque). Questi importi rappresentano circa il 27 per cento degli utili totali di Nvidia — molto più della quota media del gruppo d'imprese, del 15 per cento — e rivelano la volontà di Nvidia di intraprendere un percorso diverso rispetto a molte aziende di semiconduttori che stanno tagliando drasticamente i costi di R&D. Nvidia ha fatto passi importanti anche per creare un ecosistema di prodotti end-to-end centrato sulle proprie GPU.

Tabella 5 **Nvidia sta investendo in modo massiccio nell'intelligenza artificiale, acquisendo quote di mercato e ottenendo ritorni eccezionali**




[1] Il gruppo di società pari comprende diverse aziende di semiconduttori nell'indice PHLX Semiconductor Sector (SOX).
[2] Sviluppatori di GPU formati con programmi di Nvidia.
[3] Startup partecipanti all'Nvidia Inception Program.
[4] Gli utili di Nvidia generati dall'Apprendimento profondo sono in termini di Datacenter (HPC, artificial intelligence) e auto.

Fonte: Company financials; Nvidia;ricerche di stampa; S&P; McKinsey analysis

L'azienda sta investendo moltissimo per formare i propri sviluppatori con le competenze utili a usare le GPU per l'Apprendimento profondo, finanziando startup che contribuiscono a diffonderne l'utilizzo, stringendo partenariati per creare soluzioni che incorporano i prodotti dell'azienda, e aumentando il numero delle applicazioni gestite da GPU. Altre imprese che seguono strategie altrettanto originali, potrebbero ottenere ritorni eccezionali.

Il successo di Nvidia dimostra che le aziende tecnologiche non potranno eccellere nell'IA mantenendo lo status quo. Devono rivedere ora la loro strategia e fare le grandi scommesse che serviranno a sviluppare offerte solide di IA. Con una posta in gioco così alta, le aziende non possono permettersi un piano nebuloso e provvisorio per catturare il valore.
Quali le riflessioni principali da fare mentre si va avanti? La nostra indagine suggerisce alcune idee emergenti sulle questioni classiche in materia di strategia d'impresa:

- ***Dove Competere.*** Nel decidere dove competere, le aziende dovranno guardare sia ai settori, sia ai microverticali. Dovranno scegliere i casi d'uso che si adattano alle loro capacità e presentano un vantaggio competitivo. Dovranno indirizzarsi alle necessità più urgenti di un settore, come la rilevazione di operazioni fraudolente con carta di credito.

- ***Come competere.*** Le aziende dovranno cercare adesso i partener adatti, o procedere ad acquisizioni per costruire ecosistemi intorno ai propri prodotti. I fornitori di hardware dovranno spostarsi verso l'alto dello stack, i soggetti che si occupano di software dovranno fornire soluzioni chiavi in mano. È ora anche di guardare con occhi nuovi ai modelli di monetizzazione. I clienti si aspettano che i fornitori di IA assumano parte del rischio, in un acquisto: questo potrebbe generare opzioni di prezzo creative. Per esempio, un'azienda potrebbe vendere un'apparecchiatura per la risonanza magnetica dotata anche di IA al solito prezzo, chiedendo di pagare un extra solo per eventuali immagini elaborate con l'uso dell'Apprendimento profondo.

- ***Quando Competere.*** Le aziende high-tech sono famose per le soluzioni sofisticate e avveniristiche ma, se parliamo di IA, concentrarsi sulla perfezione potrebbe essere deleterio. I nuovi entranti potranno migliorare e guadagnare rapidamente in scala, per poi diventare un riferimento nel settore. Le aziende dovranno concentrarsi su soluzioni forti, che consentano di stabilire una presenza nello scenario attuale, piuttosto che ambire alla perfezione. Con un primo successo alle spalle potranno guardare oltre, a opportunità più speculative.

Se le aziende aspettano due o tre anni prima di stabilire una strategia di IA e scommettere sul futuro, rischiano di non trovare poi lo slancio giusto in un mercato in rapida evoluzione. La maggior parte conosce la posta in gioco e vuole avanzare, ma è priva di una strategia forte. I sei principi chiave esposti in questo lavoro possono orientare le imprese nella direzione giusta, aiutarle a fare il salto. La domanda fondamentale è: quali operatori prenderanno questa direzione, prima che la finestra delle opportunità si chiuda?

[1] "The race for AI: Google, Baidu, Intel, Apple in a rush to grab artificial intelligence startups," CB Insights, 21 giugno 2017, cbinsights.com.

**Gaurav Batra** è partner dell'ufficio di McKinsey a Washington, DC. **Andrea Queirolo** è associate partner dell'ufficio di New York e Nick Santhanam è senior partner nell'ufficio della Silicon Valley.

Gli autori ringraziano **Eylul Harputlugil, Sam Morton, Harish Soundararajan** e **Ben Byungchol Yoon** per i contributi a quest'articolo.

Copyright © 2017 McKinsey & Company. Tutti i diritti riservati.

Traduzione a cura della Regione Toscana
Direzione Attività Produttive

Contatto per la distribuzione:
Gaurav Batra
Numero di telefono: +1 650 842 5612
Email: Gaurav_Batra@McKinsey.com

Dicembre 2017
Progetto Sydney Design Studio Copyright
© McKinsey & Company