Università
Ca'Foscari
Venezia

Corso di Laurea
magistrale
in Economia e
Finanza
LM-56

Tesi di Laurea

La Gestione del Rischio nei Modelli di Intelligenza Artificiale applicati al Decision Making

La tecnologia, i rischi connessi e il rispettivo framework di risk management

Relatore
Prof. Andrea Giacomelli

Laureando
Steven Bozzetto
Matricola 862125

Anno Accademico
2021 / 2022

SOMMARIO

INTRODUZIONE

Il presente elaborato vuole essere una raccolta organica di metodi e best practice che possano essere utili e direttamente applicabili a tutte quelle imprese che hanno o che dovranno avere a che fare con l'intelligenza artificiale all'interno della propria organizzazione, ed affrontarne quindi i rischi collegati. L'obiettivo è fornire la consapevolezza e gli strumenti per affrontare tali rischi: questo si traduce in un'analisi il quanto più possibile ampia delle fonti di rischio, guardando l'esperienza passata e gli errori commessi dalle IA a disposizione sul mercato e degli utenti, e lo studio di un approccio di gestione del rischio che sia in grado di intercettare tutte le fonti di rischio, comprese quelle nuove che la tecnologia porta con sé, come ad esempio il rischio di non saper spiegare come una decisione sia stata presa da una macchina.

L'IA è un'opportunità che deve essere sfruttata. Come si evidenzia più volte nell'elaborato, le aziende che utilizzano modelli predittivi, di machine learning, di computer vision o qualsiasi altro sistema "intelligente" sono in grado di efficientare non solo la propria curva di costo, ma di potenziare i propri processi di marketing, di post-selling, di distribuzione e di definizione strategica.

Sebbene il presente elaborato sia incentrato sulla gestione dei rischi legati all'utilizzo di IA nelle scelte di business, non si è voluto dare per scontato la conoscenza di alcune nozioni di base della tecnologia protagonista, al fine di fornire una traccia non tecnica di cosa effettivamente si intenda per intelligenza artificiale. Questa è la sfida principale per gli organi di amministrazione e dirigenza (il *C-level*, stando al modo di dire anglosassone): avere le competenze, la conoscenza e la consapevolezza dei rischi ed opportunità dei sistemi di IA, ed essere in grado di fare da "ponte" tra il lato business/ strategico e il lato tecnico dei data scientists. Per tale motivo, sono inserite delle nozioni e degli elementi la cui natura potrebbe sembrare puramente divulgativa, allo scopo di comprendere in pieno l'uso e le implicazioni di IA.

È il momento dell'intelligenza artificiale, e le imprese che non la adotteranno rimarranno indietro rispetto ai competitors che già ne hanno implementato soluzioni avanzate. Ciò che sembra ancora una rivoluzione tecnologica del futuro è già pienamente presente e sta cambiando le abitudini di acquisto e consumo della clientela. Dalle applicazioni in ambito marketing e after-sales alla personalizzazione dell'esperienza e del prodotto. Non solo sulla clientela, ma anche all'interno delle organizzazioni produttive stesse: dall'advance analytics, alla capacità predittiva e di pianificazione, all'automatizzazione di processi, anche quelli decisionali. Questi sono

solo alcuni esempi generici di come l'IA stia portando beneficio all'imprese che hanno deciso di investire per prime nella nuova tecnologia, guadagnandone in una diminuzione di costi, di complessità di processo, in efficienza e capacità previsionale, avendo maggior contatto e conoscenza dei propri clienti. Il tutto riflettendosi non solo nel risultato economico, ma anche nell'ambiente di lavoro, dove questa nuova tecnologia supporta le risorse umane.

Tuttavia, non è semplice le strada che porta all'adozione di "sistemi intelligenti". Le competenze necessarie si estendono sia in senso verticale che trasversale: non è sufficiente il team di esperti data scientists in grado di implementare l'algoritmo; è forte la necessita di altre conoscenze e competenze, come quelle necessarie a incastrare a regola d'arte la tecnologia con il business, con la strategia, con i lavoratori dipendenti e con i terzi. Non è solo un problema di competenze, ma anche di cultura: è necessario che le soluzioni di IA siano implementate con principi etici, che non vadano a ledere, influenzare o discriminare gli utenti che si interfacciano con essa. Inoltre, deve esserci una cultura dell'IA condivisa tra tutta l'organizzazione, istruendo e educando dipendenti e manager anche non-IT per evitare di percepire la tecnologia come minaccia anziché supporto, e saper sfruttare correttamente l'aiuto fornito.

Ed ecco che, laddove manchino competenze o cultura, nascono rischi che l'impresa deve conoscere e saper affrontare con un approccio strutturato e complessivo, che analizzi ogni fase di produzione dei sistemi e i flussi verso ciascun stakeholder coinvolto, sia esso interno all'azienda o terzo.

Nel primo capitolo del presente elaborato si vuole offrire una panoramica generale sullo stato dell'arte della tecnologia e sui suoi rischi, legati ad alcuni eventi che hanno posto l'attenzione su tematiche sensibili, come la discriminazione.

Nel secondo capitolo, si è voluto presentare una rappresentazione complessiva e generalizzata dei rischi connessi all'uso dell'intelligenza artificiale, abbozzando una tassonomia che vuole abbracciare potenzialmente tutti i fattori di rischio. Una volta identificati i rischi, si vuole presentare un framework di risk management che si adatti alle sfide poste dall'intelligenza artificiale, a partire dal Model Risk Management, utilizzato soprattutto dalle istituzioni finanziare. Tale modello è attualmente carente rispetto ai modelli di intelligenza artificiale, per cui è necessario prevedere degli adeguamenti nei suoi elementi e, soprattutto, un approccio puntuale in ciascuna fase del ciclo di vita dell'algoritmo: la mitigazione del rischio deve partire fin dalla fase di ideazione e di *proof of concept,* seguendo tutti gli step di sviluppo e continuando

successivamente alla fase di go-live.

Nel terzo capitolo, invece, verrà illustrata la direttrice normativa tracciata dalla Commissione Europea con l'Artificial Intelligence Act del 21 aprile 2021, che dirama i primi vincoli di compliance ai quali le imprese utilizzanti e/o sviluppatrici di intelligenza artificiale dovranno adeguarsi nei prossimi anni. È per i nostri scopi rilevante sottolineare come tale normativa sia stata sviluppata con un approccio basato sul rischio, prevedendo tre distinte categorie di sistemi di IA: rischio inaccettabile, rischio elevato, rischio moderato. In proporzione al grado di rischio di ciascun sistema, saranno più elevati i requisiti che il sistema dovrà soddisfare.

Prima di iniziare a parlare effettivamente dei rischi e la loro gestione, seguirà sempre in sede di introduzione una breve disgressione generale di carattere divulgativo per prendere consapevolezza di cosa sia l'intelligenza artificiale.

Cosa si intende con Intelligenza Artificiale?

Togliamoci dei pregiudizi partendo da cosa non è intelligenza artificiale: non è Terminator e neanche HAL 9000. l'IA non è un automa di forma umanoide come quello interpretato da Schwarzenegger con la propria volontà di distruggere il genero umano, e neanche un supercomputer di bordo con una propria coscienza. Ed è proprio questa la discriminante per la quale non bisogna aver paura dell'IA: il termine *intelligenza* indica la "*parte computazione dell'abilità di raggiungere obiettivi nel mondo*" (McCarthy, 2004). Le macchine intelligenti non potranno avere[1] *coscienza*, ovvero consapevolezza di sé e di ciò che sta attorno, del proprio essere e del proprio pensare e agire. È la coscienza a discriminarci dalle macchine, non l'intelligenza.

Proviamo in primis a dare una definizione teorica, cercando di declinarla passo dopo passo verso una classificazione più pratica per oggetti.

Colui che è considerato il padre dell'intelligenza artificiale, John McCarthy (2004), definisce in un suo paper l'IA come *"la scienza e l'ingegneria di creare macchine intelligenti, in particolare programmi (software) intelligenti. È correlata alla pratica di usare i computer per capire l'intelligenza umana, ma non deve limitarsi a metodi che sono biologicamente osservabili"*[2]. Declinando in altri termini, potremmo dire che l'IA è quella scienza o branca che si occupa delle "macchine (o agenti) intelligenti", ovvero

[1] Almeno per il momento, non sembra essercene possibilità

[2] Si consiglia, per capire bene la definizione e le sfumature di "intelligenza" la lettura del paper, reperibile online al seguente link: http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

qualsiasi sistema che riesca a percepire dal suo ambiente circostante elementi che permettano di massimizzare la probabilità di raggiungere gli obiettivi per cui è stato programmato.

Se volessimo dare una forma fisica all'intelligenza artificiale, allora potremmo immaginarcela come uno schermo ricco di righe di codice di programmazione che eseguono funzioni statistico-matematiche. Alla base di tutto c'è un *problema di ottimizzazione*. Non siamo in grado per ogni problema di trovare una soluzione, di trovare una formula chiusa che ci permetta di avere il risultato. Proprio per questi problemi sono stati inventati degli algoritmi che, da soli attraverso i dati, trovano un modo di risolvere il problema (il target) che è stato loro affidato. Successivamente, questi algoritmi, proprio come un cervello o, per meglio dire, un motore, possono essere inseriti in robot, nelle automobili, nei siti web, nei motori di ricerca, nei macchinari medici, nei cellulari e così via. Questo motore/cervello, denominato *machine learning* (o "ML" abbreviato) è ciò che accumuna l'intelligenza artificiale che troviamo in Alexa, in Google e nella Tesla, per quanto in apparenza possano sembrare distanti per funzioni e aspetto.

In termini pratici, possiamo definire l'intelligenza artificiale come un grande cappello che raccoglie sotto di sé sistemi che vogliono ottimizzare problemi diversi. Difatti, il machine learning è alla base di tantissime tecnologie, quali la manutenzione predittiva, video games e affini[3], l'analisi e la pianificazione in senso generale, *speech recognition* e comprendere il linguaggio naturale (*Natural Language Processing*, o *NLP*), ovvero capirne il significato, saperlo tradurre e capirne anche il sentimento che esprime ciascuna frase. Le funzioni di NLP sono quelle alla base degli smart speaker, delle chatbot e degli assistenti virtuali. Grazie al *deep learning* (o "DL"), un ulteriore innovazione, le macchine sono in grado di vedere e riconoscere oggetti attraverso la c.d. *computer vision* e *object detection*. Unendo questi sistemi è possibile sviluppare ambiti sempre più complessi, come la *Robotic Process Automation* (o RPA), la classificazione euristica, la guida autonoma e altri ambiti ancora in sviluppo.

Evoluzione Storica e livello attuale della tecnologia

Probabilmente non è ancora scontato sottolineare quanto l'intelligenza artificiale sia la tecnologia del ventunesimo secolo. Ma per quanto possa sembrare una tecnologia nuova

[3] Come il famoso Deep Blue che nel 1997 sconfisse il campione del mondo di scacchi Kasparov

è in realtà da quasi 70 anni che se ne parla.

In questi passaggi ne riscostruisce la storia Quintarelli, et al., 2020. Nel 1955, un gruppo di ricercatori universitari richiede un finanziamento di 13.500 dollari alla fondazione Rockfeller per un progetto estivo dell'università di Dartmouth il primo studio *"sulla base della congettura che ogni aspetto dell'apprendimento o qualsiasi altra caratteristica dell'intelligenza può, in linea di principio, essere descritta in modo così preciso da renderlo simulabile da parte di una macchina"*: il Dartmouth Summer Research Project for Artificial Intelligence (McCarty, Minsky, Rochester, & Shannon, 1955). Da quel momento in poi, gli studi e gli investimenti sulla nuova tecnologia procedono ad intermittenza fino ai giorni nostri. Si riveleranno fondamentali allo sviluppo dell'IA gli studi in discipline non strettamente correlate: dalle neuroscienze, come quelli di McCulloch e Pitts, i quali ipotizzarono che i neurono comunicassero tra loro con segnali simili alle funzioni logiche "AND" e "OR" (e quindi simili al computer), all'invenzione del percettrone (o "*perceptron"*) di Rosenblatt, una sorta di "neurone artificiale" il cui costrutto sarà poi fondamentale per lo sviluppo delle reti neurali. Nel 1959 Arthur Samuel, informatico dell'IBM, cogna il termine "Machine Learning", ovvero quel sistema di algoritmi che permette ad una macchina di imparare dai dati, anche se non esplicitamente programmata. Tuttavia, negli anni ottanta, gli sviluppi teorici raggiunti fino a quel momento dimostrano che l'obiettivo di raggiungere l'intelligenza umana è ancora molto lontano. I finanziamenti dedicati allo studio della tecnologia e, di conseguenza, la ricerca, vengono sospesi: questo periodo che durerà fino ai primi anni duemila viene denominato "l'inverno dell'intelligenza artificiale" (c.d. "Winter of AI").

Perché ora è esplosa l'IA e perché investirci in questo momento

Il nuovo millennio fu decisivo per la nuova, decisiva primavera dell'intelligenza artificiale. Tra i più grandi scogli che dovettero affrontare i ricercatori fino agli anni '80 ci fu il costo per l'architettura hardware sulla quale immagazzinare e processare i dati da dare in pasto agli algoritmi. Ad esempio, lo stesso Samuel all'epoca non riuscì a completare il proprio algoritmo di machine learning basato sul gioco della dama perché non c'era abbastanza memoria per calcolare e memorizzare le combinazioni di gioco. Adesso, invece, possiamo tenere comodamente un hard disk su una mano, e la capacità

di computazione delle macchine è aumentata vertiginosamente[4].

L'altra innovazione vincente fu internet e il cloud computing: grazie alla "rete", più sistemi locati in parti diverse del mondo, possono dialogare tra loro, raccogliere input e generare output in tempo reale. Grazie a queste tecnologie Amazon Alexa e gli altri *smart* speakers sono in grado di ascoltarci e risponderci. Il nostro input vocale (un saluto, una domanda) viene raccolto dal dispositivo situato sulla nostra mensola, inviato alle enormi stanze server situate negli Stati Uniti, processati attraverso algoritmi intelligenti (di *Natural Language Processing* o *NLP* per la precisione) che invieranno indietro l'output desiderato, sia esso la previsione del meteo, la riproduzione di una canzone, l'ordine di preparare il caffè alla macchinetta collegata e così via.

Ecco che cominciarono le prime applicazioni pratiche e la commercializzazione di macchine intelligenti: nel 2002 una società fondata da tre ricercatori del MIT di Boston lanciò sul mercato il primo aspirapolvere automatico, in grado di muoversi in autonomia ricordando la pianta della casa. Poi, furono le grandi società tech americane a trainare le innovazioni e le applicazioni di IA: da Google con il suo motore di ricerca in primis, ad Amazon, Facebook, Apple e Microsoft (i c.d. "GAFAM", dall'acronimo formato dalle loro iniziali).

Tuttavia, sebbene l'IA già stia permeando le nostre vite di consumatori o l'organizzazione delle imprese, è ancora un territorio economicamente ricco da sfruttare. Prendiamo come riferimento la teoria della diffusione dell'innovazione (Rogers, 1962)[5], che riassume il comportamento di consumatori e imprese durante il ciclo di vita di un nuovo prodotto o servizio. Dovremmo chiederci, in quale punto della funzione siamo in questo momento per l'IA? Consideriamo la seconda metà del ventesimo secolo come la fase di innovazione; i GAFAM possono essere considerati i portabandiera degli *early adopters*, ovvero coloro che hanno cavalvato l'innovazione e che ne hanno dato per primi un'applicabilità pratica e commerciale. Ebbene, si ritiene che al momento presente le imprese che stanno adottando e che adotteranno prontamente soluzioni di intelligenza artificiale (ad integrazione nella prorpria organizzazione o ad integrazioni dei propri prodotti o servizi) rientrano nella categoria

[4] Il paragone tra il supercomputer ASCI RED, voluto e costruito nel 1997 per mantenere e monitorare l'arsenale nucleare dal governo degli Stati Uniti, e una scheda grafica di top gamma NVIDIA oggi in commercio rende l'idea: ASCI Red aveva una velocità massima di calcolo pari a 1.3 Teraflops e il governo americano spese 55 milioni di dollari per la sua costruzione; la scheda NVIDIA Titan V ai giorni nostri ha una potenza di calcolo pari a 110 Teraflops ed è in vendita a 3.000 dollari.

[5] Un estratto sufficiente a capirne la teoria è disponibile gratuitamente su Wikipedia: https://en.wikipedia.org/wiki/Diffusion_of_innovations

degli *early majority*. Un'impresa che decidesse fornire o adottare ora servizi di intelligenza artificiale, avrebbe ancora un vantaggio competitivo sul tempo rispetto a coloro che esiteranno a farlo. Questo perché i consumatori in questa fase raggiungeranno l'apice della domanda di prodotti/servizi di IA (il segmento corrispondente alla curva gaussiana di colore blu) e l'impresa potrà sfruttare la crescita del mercato e consolidare la propria fetta prima che si generi una strenua concorrenza (come indica la funzione logistica colorata di giallo).

Questo può essere considerato il *vantaggio competitivo esogeno*, dovuto dal ciclo di vita della tecnologia, a cui poi dovrà sommarsi il *vantaggio competitivo endogeno*, ovvero quello che portà in sé l'utilizzo di intelligenza artificiale, e che verrà commentato brevemente nel prossimo capitolo.

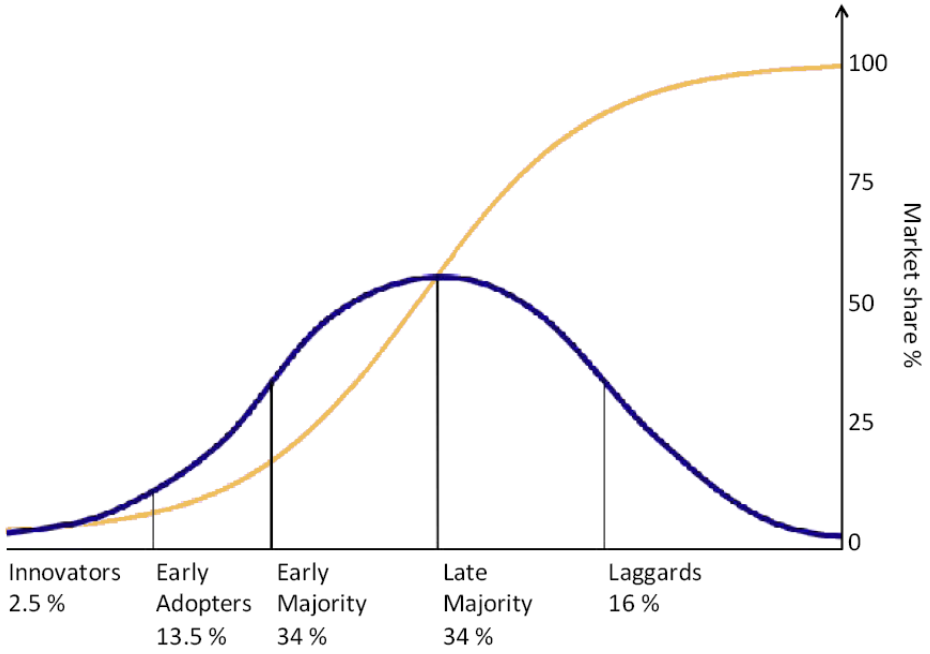

Figura 1 - Teoria della diffusione dell'innovazione (Crediti: Wikipedia)

CAPITOLO 1

IA: CONOSCERE OPPORTUNITÀ E RISCHI

1.1 Un po' di numeri: l'impatto economico

PwC mostra in uno studio condotto nel 2017 come il PIL mondiale potrebbe aumentare grazie al solo utilizzo di intelligenza artificiale fino al 14% entro il 2030, l'equivalente di 15.7 miliardi di dollari, rendendo quindi l'IA la più grande opportunità commerciale nella nostra economia segnata da velocissimi cambiamenti. I maggiori guadagni si registreranno probabilmente in Cina, le cui stime parlano di una crescita del 26% del PIL entro il 2030, seguita dal Nord America, con una crescita stimata nello stesso periodo del 14%; L'Europa, stando alle stime, inseguirà i due colossi economici globali, con una crescita attesa dell'economia pari al 9,9% per il Nord Europa e del 11,5% per il Sud Europa (in totale, circa due miliardi e mezzo di dollari). I settori maggiormente toccati saranno il commercio al dettaglio, i servizi finanziari e la sanità. L'impatto economico maggiore dell'intelligenza artificiale sarà portato principalmente da: un aumento della produttività derivante dall'automazione di processi (compreso l'uso di robot e di mezzi a guida autonoma); un aumento della produttività della forza lavoro, che sarà assistita e affiancata da sistemi di "intelligenza aumentata"[6]; un aumento delle domanda di prodotti e servizi personalizzati e/o di maggiore qualità (PwC, 2017).

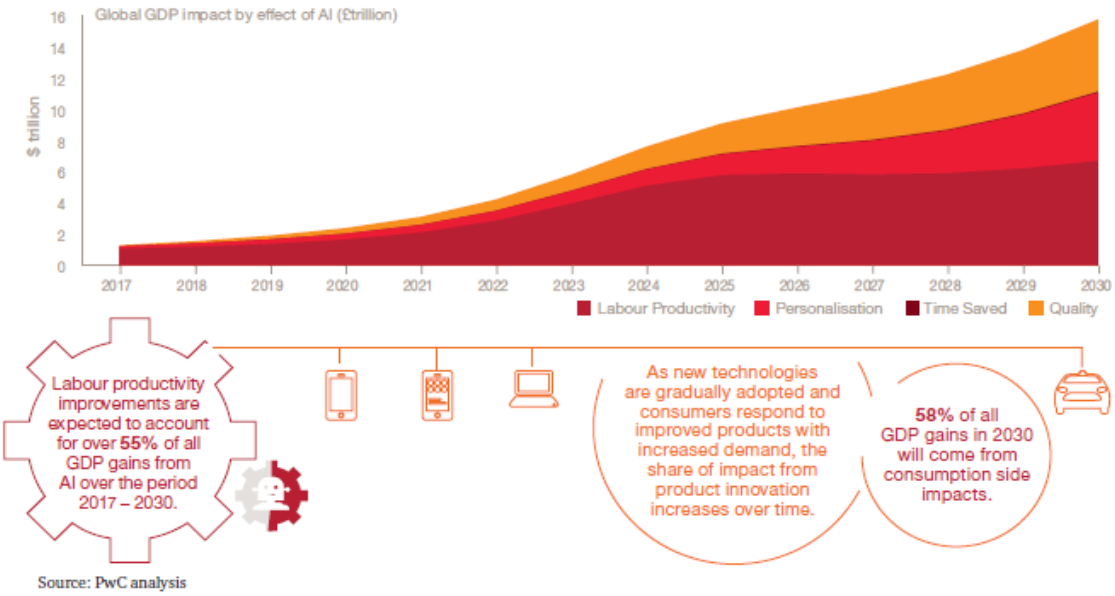

Figura 2 - L'impatto dell'IA sul PIL mondiale, previsione 2017-2030, PwC

[6] Per intelligenza aumentata, si legge nel documento, si intendono quei sistemi di IA che facilitano il decision making per le persone, e che continuamente imparano dalle interazioni con esse e dall'ambiente. Non sono quindi sistemi che prendono decisioni, ma che forniscono un output grazie al quale la decisione dell'uomo sarà migliore.

Il "McKinsey Global Survey on Artificial Intelligence 2020"[7] evidenzia in incipit come un piccolo contingente di partecipanti da diverse industrie attribuisce il 20% o più del loro EBIT all'utilizzo di AI, e pianifica di investire ancora di più nella tecnologia per cavalcare la trasformazione digitale accelerata dalla pandemia da COVID-19. In particolare, l'adozione di algoritmi di IA ha permesso alle aziende di aumentare la redditività in funzioni quali gestione e ottimizzazione dell'inventario, pricing e promozioni, customer service analytics, e previsione di domanda e vendite (sales & demand planning). Sensibili riduzioni di costi sono state riscontrate nell'ottimizzazione del talent management, l'automazione dei contact-center aziendali e l'automazione di magazzino. Il 22% dei partecipanti alla survey riconosce che almeno il 5% del proprio EBIT 2019 sia da attribuire all'utilizzo di IA (McKinsey, 2020). Nella figura successiva viene riportata una statistica significativa relativa all'aumento dei ricavi e la diminuzione dei costi raccolta per settore.

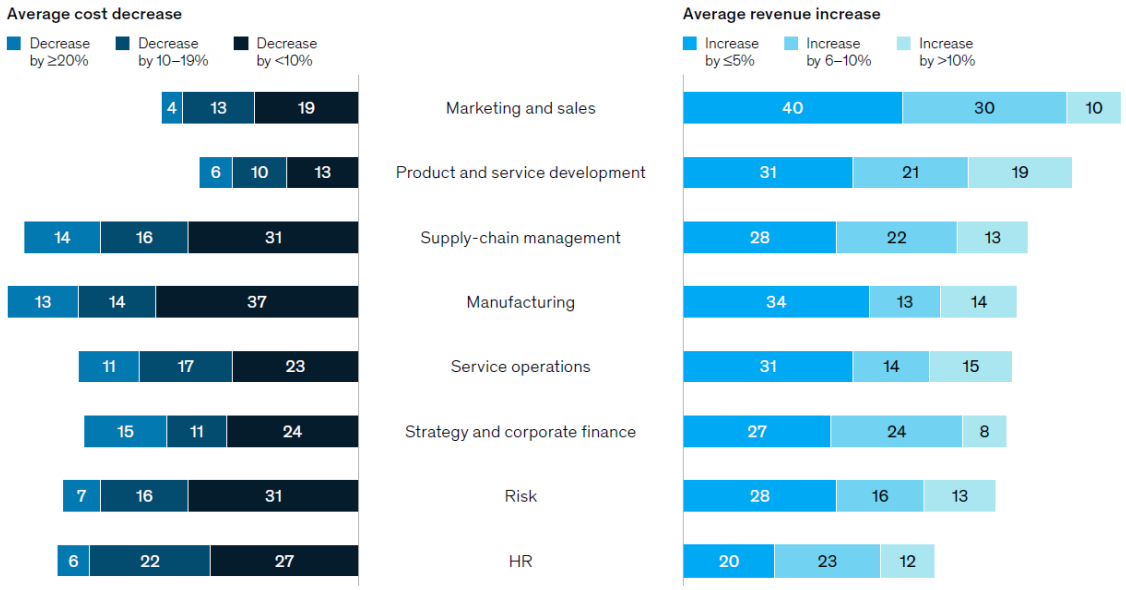

Figura 3 - Aumento medio dei ricavi e diminuzione media dei costi, McKinsey Global AI Survey 2020

Gli intervistati hanno risposto inoltre a domande inerenti al rischio dell'uso di intelligenza artificiale, evidenziando che solo una minoranza di società conosce effettivamente tutti i principali rischi e, tra queste, solo un gruppo ancora più ristretto prevede azioni di mitigazione per ciascun fattore di rischio. Solo il 41% degli

[7] Il sondaggio online è stato sottoposto a 2.395 partecipanti da diverse aree geografiche, industrie, company sizes, competenze e incarichi. Di questi, 1.151 hanno risposto che almeno in una funzione aziendale sono state adottate soluzione AI, ed è stato loro come sono state sfruttate. Per maggiori informazioni, il sondaggio è pubblicato al seguente link: https://www.mckinsey.com/business-

intervistati, infatti, ha dichiarato di riconoscere e prioritizzare i propri rischi di IA. Il sondaggio ha indagato sui dieci fattori di rischio più comunamente ricorrenti. Almeno la metà degli intervistati ritiene rilevanti e riconoscono il problema della cybersecurity e della compliance regolamentare, ma non altrettanta consapevolezza è rinvenuta in altri rischi critici dei modelli IA, quali la privacy, l'*explainability*[8], o la perdita di posti di lavoro.

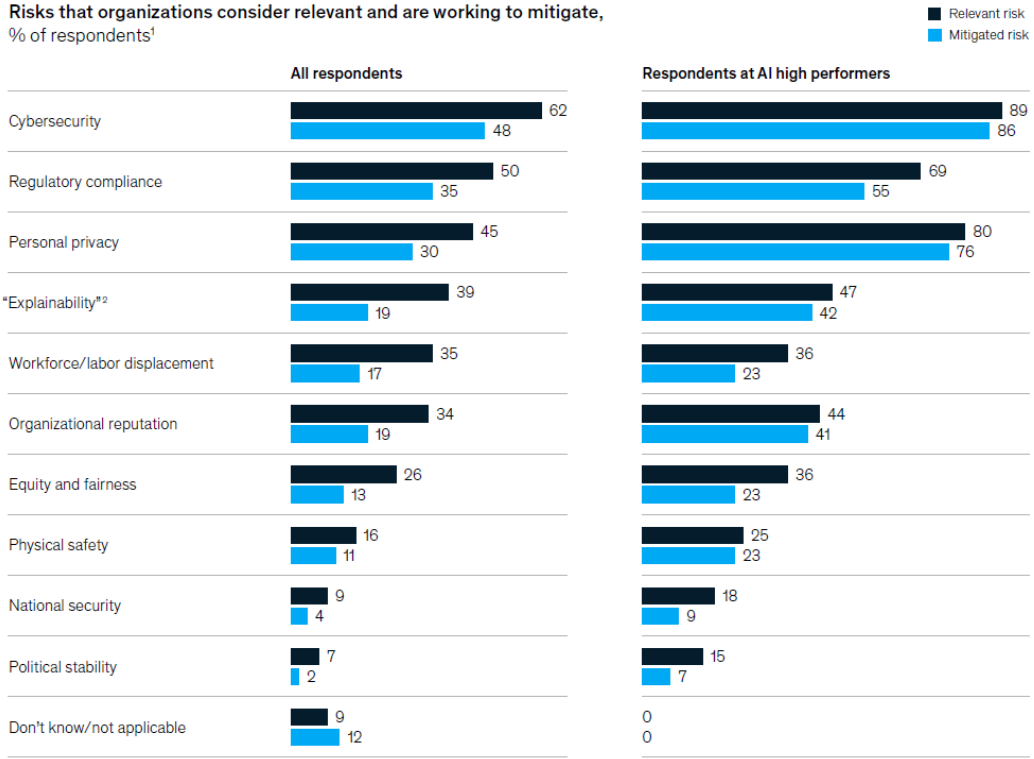

Figura 4 - I rischi dell'IA che le imprese considerano rilevanti e da mitigare, McKinsey Global AI Survey 2020[9]

Con riguardo alle risorse umane, la più grande preoccupazione è sempre stata che l'intelligenza artificiale sostituisse il lavoro umano, con conseguente rischio di licenziamenti e di esternalità negativi sulle società. Ma le risposte al questionario dimostrano che, almeno fino ad oggi, questo scenario non si è concretizzato. Meno di un terzo degli intervistati ha registrato meno del 3% di ridimensionamento della forza lavoro a causa dell'IA, e solo il 5% degli intervistati ha dichiarato che tale ridimensionamento è stato superiore al 10% della forza lavoro. Mentre le aziende di

functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020

8 L'abilità del modello di spiegare come è arrivato a generare l'output. è un tema cruciale per gli algoritmi che possono prendere decisioni al posto dell'uomo in ambiti delicati come, ad esempio, la valutazione del merito di credito o in ambito medico.

9 Il questionario suddivide in questa fase una parte degli intervistati nei c.d."AI High Performes", ovvero coloro che hanno registrato un impatto organizzativo ed economico migliore e si configurano come leader nella trasformazione gestionale con l'IA.

carattere manufatturiero (come l'automotive e industrie di assemblaggio) sono più in linea con una diminuzione di forza lavoro umana, la maggior parte degli intervistati dichiara un aumento delle assunzioni del 3% o più rispetto al passato. Tuttavia, il trend potrebbe cambiare nei prossimi tre anni: mentre il 21% degli intervistati dichiara che l'adozione di soluzioni IA produrrà un aumento dei posti di lavoro, il 34% prevede un trend inverso; un ulteriore 28% dichiara invece che l'utilizzo di IA non produrrà scostamenti sostanziali sull'impiego. Le imprese, tuttavia, si aspettano i cambiamenti che l'IA porterà sulla forza lavoro si riscontrerà in trend diversi per funzioni aziendali: si prevede maggiormente una decrescita di personale nelle funzioni quali risorse umane, manifattura, *supply chain management* e *service operations*; sarà invece in crescita il numero di impiegati nel marketing, vendite e *product development*.

Questo implica che imprese e lavoratori dovranno essere pronti a riqualificare le proprie competenze, addestrandosi ai nuovi compiti. Quasi sei intervistati su dieci affermano di aver già ri-addestrato i propri dipendenti, mentre l'83% si aspetta di dover procedere al *re-skilling* dei propri lavoratori nei prossimi tre anni.

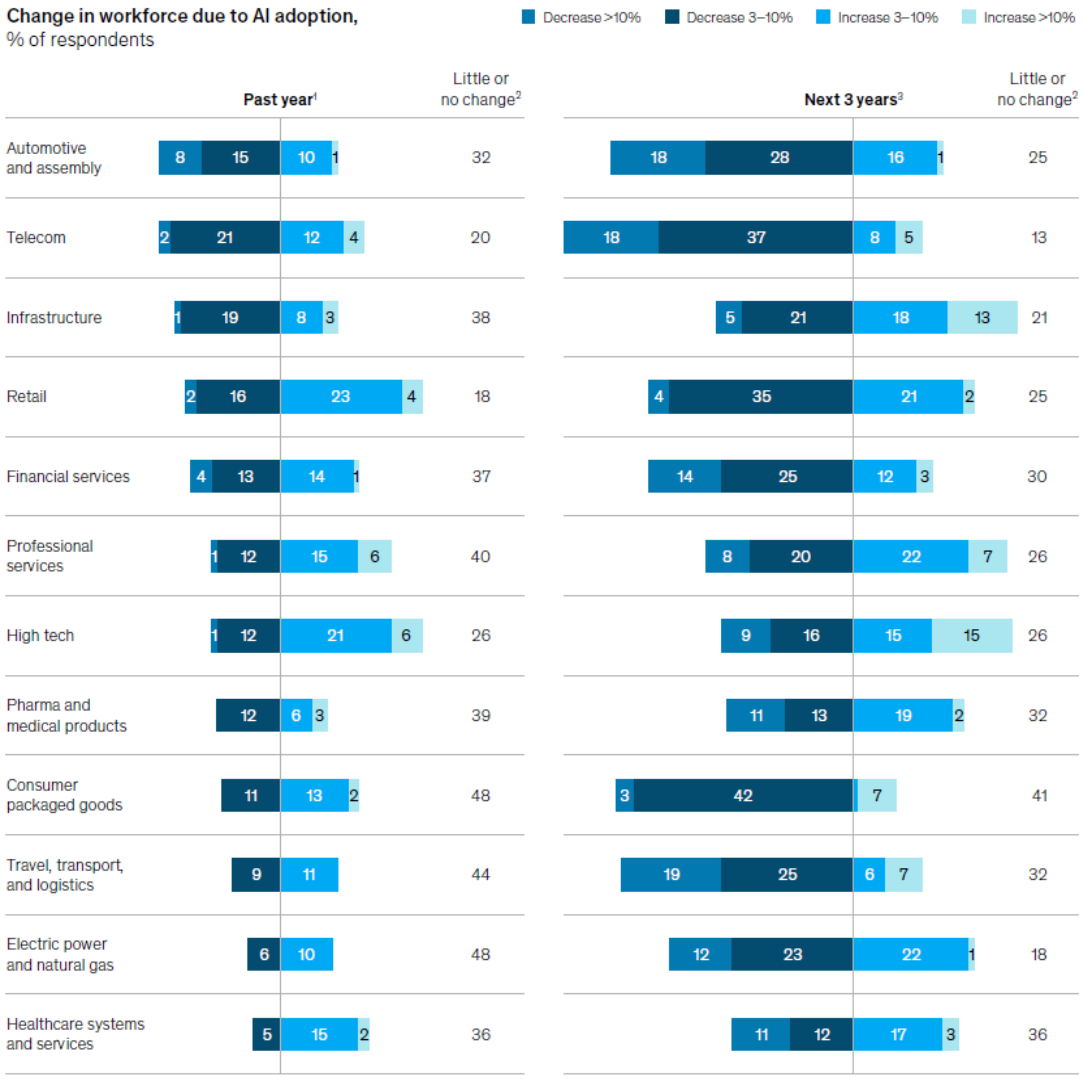

Figura 5 - Cambiamento della forza lavoro a fronte dell'adozione di IA, McKinsey Global AI Survey 2020

1.2 Stato dell'arte per settore economico: applicazione e rischi

Si è evidenziato come l'adozione di IA comporti diversi benefici, sia sotto l'aspetto remunerativo del business che sotto il profilo operativo e di processo. Si desidera in questo paragrafo andare più nello specifico ed analizzare come l'applicazione di soluzioni di intelligenza artificiale sono declinate per i settori economici maggiormente impattati, evidenziando anche i relativi rischi.

La *sanità* è uno dei settori per il quale il potenziale di crescita grazie all'IA è più spiccato: già sono utilizzati algoritmi nel supporto delle diagnosi, grazie al riconoscimento di pattern e correlazioni tra pazienti simili, oppure grazie al deep learning che, applicato alla diagnostica su base di immagini (es. radiografie), si rivela più accurata della sola analisi umana. Ciò si traduce in diagnosi più rapide ed accurate e la possibilità di applicare trattamenti personalizzati. Inoltre, con il monitoraggio

continuo, l'attività di prevenzione risulterà più efficacie e mirata, riducendo il rischio di malattia e di ospedalizzazione. Tra le varie applicazioni in sviluppo in ambito sanitario troviamo la medicina personalizzata e la telemedicina attraverso l'uso di big data: la "medicina multicanale partecipativa" si basa su "4 P": "Partecipazione", ovvero il coinvolgimento degli individui per la condivisione dei dati clinici e le informazioni sanitarie; "Prevenzione", cioè l'individuazione dei fattori di rischio dei singoli pazienti per indirizzare comportamenti preventivi; "Predizione", ovvero l'analisi predittiva per individuare i percorsi terapeutici e i farmaci più efficaci; "Personalizzazione" dei percorsi di cura e delle interazioni con i pazienti in base alle esigenze. Un ulteriore esempio dei benefici dell'uso dell'intelligenza artificiale in ambito sanitario è testimoniato dai vaccini COVID, sviluppati in tempo record anche grazie alle simulazioni effettuate dagli algoritmi di IA (Longo & Mischitelli, 2020). Oppure, un altro caso emblematico, è stato lo sviluppo del farmaco DPS-1181, formulato interamente da Centaur Chemist, un sistema di IA sviluppato appositamente per la ricerca farmacologica in soli 12 mesi. Il nuovo medicinale, nato dalla collaborazione tra un'azienda di Oxford e una casa farmaceutica giapponese, curerà diversi disturbi ossessivo-compulsivi (Mantovani, 2020). Per permettere il raggiungimento di tali orizzonti, tuttavia, è necessario affrontare almeno due barriere fondamentali: (1) le preoccupazioni relative alla privacy e alla protezione dei dati sanitari sensibili; (2) la complessità della biologia umana e la necessità di un ulteriore sviluppo tecnologico si traducono in tempi ancora non brevi di sviluppo e test delle soluzioni più avanzate, oltre che l'ottenimento della validazione come uso terapico da parte delle istituzioni sanitarie.
Data la grande mole di dati personali, particolare attenzione deve essere data a come gli algoritmi di machine learning interpolano le informazioni in essi racchiusi, evitando l'associazione di correlazioni eticamente errate: è il caso di un algoritmo largamente usato negli ospedali statunitensi per l'allocazione dell'assistenza sanitaria che, sistematicamente, discriminava i pazienti di colore: quando il team di ricerca ha eseguito i controlli statistici di routine sui dati ricevuti da un grande ospedale, sono stati sorpresi di scoprire che alle persone che si auto-identificavano come neri venivano generalmente assegnati punteggi di rischio più bassi rispetto ai bianchi ugualmente malati. Di conseguenza, i neri avevano meno probabilità di essere indirizzati ai programmi che forniscono cure più personalizzate. I ricercatori hanno scoperto che l'algoritmo assegnava punteggi di rischio ai pazienti sulla base dei costi totali dell'assistenza sanitaria maturati in un anno. L'ipotesi potrebbe essere sembrata

ragionevole, dato che i costi più elevati dell'assistenza sanitaria sono generalmente associati a maggiori esigenze sanitarie, e la persona di colore media rappresentata nel set di dati utilizzato dagli scienziati aveva costi sanitari complessivi simili a quelli della persona bianca media. Tuttavia, uno sguardo più attento ai dati ha rivelato che la persona nera media era anche sostanzialmente più malata della persona bianca media, con una maggiore prevalenza di condizioni come diabete, anemia, insufficienza renale e ipertensione. Nel loro insieme, i dati hanno mostrato che l'assistenza fornita ai neri costa in media 1.800 dollari all'anno in meno rispetto all'assistenza prestata a una persona bianca con lo stesso numero di problemi di salute cronici. Gli scienziati ipotizzano che tale fenomeno sia dovuto agli effetti del "razzismo sistemico", che vanno dalla sfiducia nel sistema sanitario alla discriminazione razziale diretta da parte degli operatori sanitari. E poiché l'algoritmo assegnava le persone a categorie ad alto rischio sulla base dei costi, quei pregiudizi sono stati trasmessi nei suoi risultati: i neri dovevano essere "più malati" dei bianchi prima di essere indirizzati per ulteriore aiuto. Solo il 17,7% dei pazienti che l'algoritmo ha assegnato per ricevere cure extra erano neri. I ricercatori calcolano che la proporzione sarebbe del 46,5% se l'algoritmo fosse imparziale. (Ledford, 2019). È chiaro quindi il carico sociale che una soluzione di IA porta con sé in un ambito delicato come la sanità: basti pensare al seguente dato: in Michigan, le persone di colore son circa il 14% della popolazione ma rappresentano il 39% dei decessi a causa COVID; in Louisiana, la popolazione di colore è circa il 33% del totale e rappresenta il 54% delle morti totali di COVID (CNBC, 2020)

Nel settore *automotive*, l'intelligenza artificiale ha mosso i primi passi attraverso alcune applicazioni di assistenza alla guida semi autonome, come i sistemi di rilevazione della stanchezza. La ricerca sulle macchine a guida autonoma è ancora in corso, sebbene alcuni modelli già adottino la tecnologia che non ancora si è rilevata sicura[10]. Ma le applicazioni non si limitano solo a questo: grande valore sarà dato alla manutenzione predittiva e all'applicazioni di sistemi IoT intelligenti, affinché gli stessi produttori o rivenditori possano diagnosticare e correggere determinate performance o problematiche della vettura senza l'intervento di un meccanico; come anche, attraverso sensori smart all'interno delle auto, di poter individuare eventi o fattori di emergenza, come il malore di un conducente, e chiamare prontamente i soccorsi. Nell'ambito

10 solo a titolo di esempio: https://www.wired.com/story/tesla-autopilot-self-driving-crash-california/#:~:text=Tesla%20now%20has%20another%20fatality%20to%20hang%20on,Wei%20Huang%2C%20died%20shortly%20afterwards%20at%20the%20hospital.

industriale, genericamente parlando, l'IA già porta benefici in termini di efficientamento dei processi, nella manutenzione preventiva degli impianti e nei sistemi di previsione della domanda e della produzione, con l'intento di migliorare la gestione del magazzino e ridurre i costi di stoccaggio e gli sprechi.

Il *settore bancario e finanziario* è un altro tra quelli maggiormente coinvolti nella digital transformation guidata dai sistemi di IA, legato soprattutto alla diminuzione dei costi e alla diminuzione del rischio operativo: modelli di machine learning sono impiegati nelle pratiche di merito di credito, definendo in maniera più efficacie il corretto pricing e forma tecnica di finanziamento, e meno costosa la valutazione del fido. Tuttavia, è essenziale assicurarsi che effettivamente l'IA risulti meno distorta rispetto alla valutazione umana: ad esempio, uno studio condotto sulla concessione dei mutui, mostra che le differenze nell'approvazione dei mutui tra gruppi minoritari e di maggioranza non sono solo dovute alla presenza di bias, ma al fatto che i gruppi minoritari e a basso reddito hanno meno dati nelle loro storie di credito. Ciò significa che, quando questi dati vengono utilizzati per calcolare un punteggio di credito e questo punteggio di credito utilizzato per fare una previsione sul default del prestito, allora quella previsione sarà meno precisa. È questa mancanza di precisione che porta alla discriminazione, non quindi i bias. I punteggi di credito comprimono una serie di dati socioeconomici, come la storia lavorativa, i record finanziari e le abitudini di acquisto, in un unico numero. Questo caso risulta ancor più sensibile considerando che, oltre a decidere le domande di prestito, le valutazioni di merito creditizio vengono ora utilizzate per prendere molte decisioni che impattano la vita, comprese le decisioni su assicurazione, assunzione e alloggio. (Blattner & Nelson, 2021). Altre applicazioni sono legate all'automazione del processo autorizzativo ed esecutivo delle dispositive cartacee (ad esempio, la mail da parte del piccolo cliente business) e delle mail di reclamo, piuttosto che la personalizzazione di formulari o documentazione a partire da modelli standard (ad esempio, il materiale contrattuale). Dal lato cliente, sarà possibile interfacciarsi con chatbot per l'evasione di pratiche bancarie, o contare su tool di gestione di portafoglio il più possibile allineati con le esigenze di investimento e la propensione al rischio grazie a strumenti di *"advance customer analytics"* (ad esempio, nella gestione congiunta di piani pensionistici, assicurazioni vita e gestione patrimoniale). Inoltre, è in aumento l'adozione di soluzioni IA nel combattere le frodi e il rischio di riciclaggio.

In ambito *retail* si registrano le più ampie applicazioni di IA, legate principalmente alla

personalizzazione sia dell'esperienza di acquisto che del prodotto stesso: dall'analisi dei comportamenti di acquisto online e offline del consumatore, ai suggerimenti personalizzati in fase di acquisto, alla previsione della domanda e all'adeguamento della produzione e del prezzo. Particolare attenzione è da dedicare all'uso di tali algoritmi, in grado potenzialmente di influenzare le decisioni di acquisto in maniera subliminale, a discapito dell'inconscio consumatore. In altri casi, potremmo a chiederci fino a che punto, eticamente, un'IA possa spingersi nel conoscere un consumatore: è il caso di un padre che, chiamando il servizio clienti di una catena di distribuzione, accusava la società di incoraggiare alla figlia non ancora maggiorenne a restare incinta. Il sistema di *recommendation* della società inviò dei coupon sconto su vestiti per bambini e culle, contornato da immagini di infanti e altri articoli correlati alla maternità. Alle scuse del manager, seguirono poco tempo dopo quelle dello stesso padre, venuto a conoscenza dell'effettiva gravidanza della figlia. L'intelligenza artificiale fu in grado di identificare la gravidanza della figlia a seguito della ricerca di alcuni prodotti "sentinella", quali lozioni inodore, integratori alimentari di calcio, magnesio e zinco. (Forbes, 2012).

L'IA è sempre più utilizzata anche nel *settore immobiliare*: l'IA si rivela adeguata nel supportare le società di *real estate* nella valutazione predittiva del valore delle proprietà: il prezzo delle case dipende da una miriade di fattori che possono cambiare in relativamente breve tempo. Per profilare un annuncio immobiliare, l'intelligenza artificiale nel settore immobiliare combina dati di mercato esistenti e informazioni pubbliche, tra cui il tasso di criminalità, la presenza di mezzi di trasporto, sorgenti luminose, il disturbo sonoro, la vicinanza di servizi essenziali e accessori (es. intrattenimento) e le tendenze di acquisto (l'intelligenza artificiale può, ad esempio, verificare gli annunci degli agenti immobiliari di case "soleggiate" valutando le disparità di luce stagionale e l'ora del giorno in cui la proprietà consente di ricevere la massima luce solare) (Lisowski, 2022). Alcuni dati, come anche il tasso di traffico presente nell'area residenziale, grazie all'IA può essere registrato ed analizzato in tempo reale, grazie all'uso delle telecamere stradali o le immagini satellitari puntualmente analizzate auotmaticamente grazie alla computer vision. Un'ulteriore applicazione è nella generazione di potenziali *lead* commerciali per le diverse proprietà, evitando le perdite di tempo con clienti le cui esigenze non sono soddisfatte pienamente dalla proprietà in scrutinio. Ad esempio, Zillow, un marketplace immobiliare online, utilizza un CRM basato sull'intelligenza artificiale che valuta diversi attributi per identificare quale casa sia in linea con le intenzioni di acquisto dei suoi visitatori online. L'algoritmo può anche

scoprire il tipo di proprietà che il potenziale cliente sta cercando. Tuttavia, la stessa Zillow è protagonista di uno dei worst case di applicazione dell'IA: come detto, ogni volta che una casa viene messa in vendita, l'intelligenza artificiale di Zillow suggerisce al proprietario il giusto prezzo per venderla, e gli altri i proprietari di case hanno potenzialmente un riferimento rapido (*benchmark*) nella valutazione della propria casa. Ma Zillow è diventata così sicura dei propri algoritmi che ha pensato di poter saltare in una nuova attività nota come iBuying, che consiste nell'acquisto diretto delle case sotto-prezzate da parte della società stessa, l'esecuzione di piccole manutenzioni o ammodernamenti, e la vendita finale, ricavandone il profitto. Tale pratica, in base all'algoritmo, prevedeva l'acquisto di 5.000 case al mese entro il 2024, ma così non fu: La società annunciò che la sua divisione per l'acquisto di case, Offers, ha perso più di 300 milioni di dollari negli ultimi mesi. Le conseguenze furono il licenziamento di circa 2.000 persone e la vendita di circa 7.000 case, di cui molte per prezzi inferiori a quelli originariamente pagati. Nel giorno della dichiarazione, Zillow perse circa 9 miliardi di dollari di capitalizzazione sul mercato. Il CEO di Zillow, Rich Barton, ha in gran parte incolpato il team di Data Science, dichiarando che "Fondamentalmente, non siamo stati in grado di prevedere i prezzi futuri delle case a un livello di precisione che rende questo un business sicuro in cui stare". Quindi, fondamentalmente, la loro intelligenza artificiale non era abbastanza buona, per diverse ragioni: in primis, il mercato immobiliare non è stabile, e tale si è dimostrato contestualmente alla crisi di Zillow; in secundis, cosa succede se, anziché vendere dopo due mesi, come l'algortimo prevedeva, la casa venisse venduta dopo sei mesi a causa dell'allungamento dei lavori di ristrutturazione? L'operazione risulta ancora profittevole? Infine, anche il modello di business: nel processo di acquisto delle case, i proprietari non avrebbero inserito l'annuncio sulla piattaforma, ma avrebbero offerto direttamente la casa alla divisione acquisti: ciò avrebbe incentivato persone con necessità di liquidare velocemente la proprietà o con alcuni difetti nascosti. (The Guardian, 2021).

Facendo questa breve overview dei settori maggiormente impattati dall'intelligenza artificiale, ci si può rendere conto di come i benefici siano accompagnati da rischi nuovi, sui quali le organizzazioni non sono ancora pienamente preparate ad affrontare o a misurare. La raccolta dei worst-case di applicazioni IA si configura come uno degli strumenti necessari al fine di acquisire consapevolezza circa l'insieme dei rischi connessi all'uso di intelligenza artificiale, soprattutto quando alla tecnologia viene delegata la facoltà di prendere decisione, anche non supervisionata dall'uomo. Le big

tech americane e asiatiche, in qualità di pionieri dell'adozione e sviluppo di IA, fanno scuola nell'uso degli algoritmi. Ad esempio, l'automazione è stata la chiave per il dominio dell'e-commerce di Amazon, sia all'interno dei magazzini che guidando le decisioni sui prezzi. Tra queste automazioni, è prevista in cantiere dal 2014 un algoritmo di IA per la revisione delle candidature per i posti di lavoro disponibili presso la società fondata da Jeff Bezos. Lo strumento di assunzione sperimentale dell'azienda ha utilizzato l'intelligenza artificiale per dare ai candidati punteggi che vanno da una a cinque stelle, proprio come gli acquirenti valutano i prodotti su Amazon. L'obiettivo era poter dare in pasto un numero più o meno elevato di curriculum e ottenere dalla macchina una classifica per assumere direttamente i migliori posizionati. Nel 2015, la società si accorse che il sistema valutava i nuovi "CV" sulla base di quelli raccolti nella decina d'anni precedente, in un settore non neutro dal punto di vista di genere: essendo la maggior parte degli impiegati di sesso maschile, l'algoritmo discriminava le candidate donne. Amazon modificò i programmi per renderli neutrali a questi termini particolari, ma questa non era una garanzia che le macchine non avrebbero escogitato altri modi di selezionare i candidati che potessero rivelarsi discriminatori. Ad esempio, l'algoritmo continuò a favorire i candidati che si descrivevano usanto verbi o aggettivi più comunemente trovati nelle descrizioni di profili maschili, come "esegiuto" o "catturato". La società di Seattle alla fine ha sciolto il team di progetto, in quanto gli stessi reclutatori di Amazon non facevano pieno affidamento al ranking fornito dal sistema (Reuters, 2018).

CAPITOLO 2

IA & MODEL RISK MANAGEMENT: MAPPARE E MITIGARE I RISCHI

2.1 – Tassonomia dei rischi legati all'IA

Identificare a priori i rischi che l'utilizzo di intelligenza artificiale non è un esercizio banale, data la molteplicità delle forme della tecnologia e della sua rapida evoluzione. A monte di una tassonomia, dovrebbe esserci una definizione univoca di cosa sia l'Intelligenza Artificiale, ma le sfaccettature e gli ambiti di applicazioni sono talmente eterogenei che non permettono una formalizzazione universalmente riconosciuta. Inoltre, anche decidere dove apporre il confine della classificazione è critico nella definizione dei rischi: alcuni sono propri della tecnologia in sé, mentre altri sono concorrenti a fattori noti (ad esempio, il business risk che si intreccia con l'utilizzo di un algoritmo a supporto delle decisioni).

Porre le definizioni più adatte a questi interrogativi è il compito su cui i vari legislatori nel mondo stanno lavorando, al fine di identificare una disciplina autonoma e dedicata necessaria, alla luce del crescente utilizzo della tecnologia.

In attesa di un panorama normativo in via di consolidazione, le autorità si sono quindi mosse sui principi cardine, definendo i valori etici, culturali e sociali che la tecnologia deve rispettare. Con principi e linee guida si intende fare riferimento all'insieme di quei valori, come la privacy, l'equità e la giustizia, rappresentanti le priorità sociali e i quali non possono essere misurati in modo consistente, in quanto dipendenti dal contesto socioeconomico. La Commissione Europea, parlando dell'*Artificial Intelligence Act*, ha deciso di adottare un approccio basato sul rischio, definendo quali rischi ed utilizzi sono accettabili o non accettabili. Tra questi ultimi, sono richieste degli specifici requisiti in funzione al rischio; per i sistemi definiti a "rischio elevato" si richiede "l'uso di dataset di alta qualità, lo stabilire un appropriata documentazione per garantire la tracciabilità, la condivisione di adeguate informazioni con l'utente, la previsione e l'implementazione di un appropriato sistema di supervisione umano, e raggiungere i più elevati standard in termini di robustezza, sicurezza, cybersecurity, accuratezza […] e requisiti minimi di trasparenza" (Commissione Europea, 2021). La proposta di regolamentazione verrà meglio discussa nel capitolo terzo del presente elaborato. I principi delineati dal Policy Observatory dell'OCSE richiedono che l'IA debba necessariamente (OCSE, 2021): essere gestita in maniera responsabile nel perseguimento del benessere delle persone e

del pianeta; rispettare lo Stato di diritto, i diritti umani e i valori democratici durante l'intero ciclo di vita del sistema (includendo in essi i principi di libertà, protezione dei dati, uguaglianza, non discriminazione, diversità, equità, giustizia sociale e diritti del lavoro); garantire la trasparenza e la divulgazione di informazioni significative, al fine di dare consapevolezza alle parti coinvolte nell'interazione con il sistema; essere robusto e sicuro per tutto il ciclo di vita del sistema. Il National Institute of Standards and Technology (NIST), istituzione governativa degli Stati Uniti d'America, ha così definito i principi rilevati per l'IA (NIST, 2021): (1) *Fariness*, o *equità*, ovvero uno standard determinato culturalmente la cui percezione differisce dalle diverse culture, e spesso determinato in termini di contenzioso, oltre che regolatori. Gli ingegneri spesso presumono che gli algoritmi di apprendimento automatico siano intrinsecamente equi perché la stessa procedura si applica indipendentemente dall'utente; tuttavia, questa percezione si è erosa di recente con l'aumento della consapevolezza degli algoritmi distorti e dei set di dati distorti. Probabilmente, l'assenza di pregiudizi dannosi è una condizione necessaria per l'equità; (2) *Accountability*, o *responsabilità*: Le determinazioni della responsabilità sono strettamente correlate alle nozioni di "rischio" e di "colpa", cioè la parte responsabile nel caso in cui si realizzi un risultato rischioso. Diversi antropologi hanno scritto ampiamente su come le percezioni del rischio e della colpa associate alla tecnologia differiscono sistematicamente tra le culture, e gli studiosi di diritto stanno sviluppando misure psicometriche della cognizione culturale che sono teorizzate per variare con queste percezioni del rischio; (3) *Trasparenza*: I tentativi di aumentare la trasparenza cercano di colmare un deficit di informazioni percepito. L'assunto di base è che le percezioni del rischio derivino da un'assenza di informazioni. La trasparenza riflette la misura in cui le informazioni sono disponibili a un decisore quando esprime un giudizio su un sistema di intelligenza artificiale e può estendersi dall'ambito da quali dati sono stati inclusi nel training del modello, alla struttura del modello, al suo caso d'uso previsto, a come sono state prese le decisioni, da chi, quando, ecc. In assenza di trasparenza, gli utenti sono lasciati a indovinare questi fattori e possono fare ipotesi ingiustificate sulla provenienza del modello. Sebbene sia impossibile rimuovere le conoscenze di base di un soggetto nella valutazione di un modello, rendere disponibili conoscenze adeguate è un requisito per costruire fiducia nell'intelligenza artificiale.

Si noti come tra questi principi non venga citata direttamente l'etica, in quanto non esiste uno standard oggettivo per i valori etici; le tre linee guida sopra citate possono

invece essere declinate in requisiti tecnici.

Delineate le linee guida, è possibile ora abozzare una tassonomia dei rischi legati all'uso di IA, cercando di porsi in un'ottica alta di business il quanto più possibile "sterile", ovvero senza porre assunzioni e limitazioni all'oggetto dell'attività di impresa e quindi sul fronte dei rischi da affrontare. In primis, sono state definite le seguenti macro-categorie di rischio, che dovrebbero essere in grado di abbracciare tutte le fattispecie aziendali verificabili sia internamente che esternamente:

- Attributi tecnici "propri" o "di *design*";
- Attributi tecnico-sociali;
- Tecnologico;
- Governance e Risorse Umane;
- Compliance
- Impatto sulla società;
- Mercato;
- Fornitore.

Con attributi tecnici "propri" o "di *design*" si vogliono identificare tutti i fattori che sono sotto il diretto controllo dei data scientists e dagli sviluppatori, ovvero legati a come l'algoritmo è stato costruito ed allenato e ai dati che sono utilizzato per il suo funzionamento. Affianco a tali attributi, si differenziamo quelli tecnico-sociali, indicanti i rischi derivanti dal rapporto uomo-macchina e di come il giudizio umano sia influenzato dalle metriche e dalle logiche degli algoritmi. Il rischio tecnologico analogo a quello che già conosciamo in altri ambiti, ed è legato a dove e come il sistema di IA sia situato all'interno dell'infrastruttura IT dell'impresa, e di come essa sia in grado di renderelo sicuro e fruibile per gli scopi dell'attività di business. Con Governance e Risorse Umane si vuole abbracciare l'insieme dei rischi legati alla consapevolezza e al corretto uso dei sistemi di IA all'interno dell'azienda, e di come essa debba avere le adeguate competenze per poter capire e gestire un algoritmo di IA, sia a livello di user in prima linea, sia al livello manageriale e amministrativo. Inoltre, è indispensabile capire e prevedere l'impatto che un algoritmo di IA possa avere al di fuori dell'azienda, soprattutto se offerto come servizio ai clienti, e quindi capirne i potenziali impatti a livello etico e sociale, prevedendo anche l'impatto del rischio reputazionale conseguente. Sono da considerarsi anche i rischi di mercato, che possono essere legati, ad esempio, alla fiducia e alla crebilità delle soluzioni di IA da parte della clientela, e i

rischi drivanti dall'utilizzo di soluzioni di IA sviluppati e distribuiti da terze parti, ovvero i fornitori.

Per ciascuna delle macro-categorie sono stati elencati delle fattori di rischio puntuali (seppure non esaustivi) rilevati nella maggior parte delle casistiche e della letteratura ad oggi presente sul tema.

Attributi tecnici "propri" o "di design"

- *Accuratezza*: con "accuratezza" si intende la capacità del modello di acquisire correttamente una relazione esistente all'interno dei dait di training (un concetto analogo alla validità della conclusione statistica). A differenza degli esseri umani, i sistemi di intelligenza artificiale mancano del giudizio e del contesto per molti degli ambienti in cui vengono distribuiti. Un sistema AI/ML è generalmente efficace quanto i dati utilizzati per addestrarlo e i vari scenari considerati durante l'addestramento del sistema. Nella maggior parte dei casi, non è possibile addestrare il sistema di intelligenza artificiale su tutti i possibili scenari e dati. La mancanza di contesto, giudizio e limitazioni generali dell'apprendimento possono svolgere un ruolo chiave nell'informare le revisioni basate sul rischio e le discussioni sulla distribuzione strategica. Inoltre, una scarsa qualità dei dati potrebbe non solo limitare la capacità di apprendimento del sistema, ma potrebbe anche avere un impatto potenzialmente negativo sul modo in cui prende inferenze e decisioni in futuro. La scarsa qualità dei dati potrebbe includere dati incompleti, dati errati o inadatti, dati obsoleti o dati utilizzati nel contesto sbagliato. Tali carenze possono dare origine a previsioni potenzialmente errate o scadenti o potenzialmente provocare un fallimento. L'accuratezza può essere misurata e mitigata utilizzando utilizzando metriche standard, tra cui i tassi di falsi positivi e falsi negativi. Un metodo classico è suddividere il dataset di riferimento in e parti: training, test e validazione.
- *Affidabilità*: un modello è affidabile se il suo output è insensibile a piccoli cambiamenti nel suo input e se è privo di bias di misurazione. Le tecniche progettate per mitigare l'overfitting (ad esempio, la regolarizzazione) e per condurre adeguatamente la selezione del modello di fronte al compromesso tra bias e varianza possono aumentare l'affidabilità del modello. La definizione di

affidabilità che viene utilizzata qui è analoga alla "validità del costrutto"[11] nelle scienze sociali, anche se senza riferimento esplicito a un costrutto teorico. In particolare, questa definizione cattura la validità convergente-discriminante (se i dati riflettono ciò che l'utente intende misurare e non altre cose) e l'affidabilità statistica (se i dati possono essere soggetti ad alti livelli di rumore statistico e bias di misurazione). Le misure di affidabilità potrebbero includere i punteggi Kappa di Fleiss o i test di bontà dell'idoneità da un'analisi fattoriale (NIST, 2021).

In relazione ai bias, il tema verrà meglio approfondito nel capitolo successivo.

- *Robustezza*: un modello è robusto se si applica a più impostazioni oltre le quali è stato addestrato. Le minacce alla robustezza del modello possono essere mitigate riconoscendo esplicitamente i limiti della strategia di campionamento con cui sono stati selezionati i dati di training, test e holdout e assicurando che i modelli non vengano applicati "off label" (ad esempio, in domini che non sono rappresentativi di questi dati di training). Pertanto, la robustezza è analoga al concetto di "validità esterna"[12] nelle scienze sociali.
- *Sicurezza e Resilienza*: un modello insensibile agli attacchi contraddittori o, più in generale, ai cambiamenti imprevisti nel suo ambiente o utilizzo, può dirsi resiliente e sicuro. Questo concetto ha una certa relazione con la robustezza, tranne per il fatto che va oltre la provenienza dei dati per comprendere usi imprevisti o esplicitamente ostili del modello o dei dati. Mitigare questi rischi è un'area di ricerca aperta, ma può trarre vantaggio da approfondimenti sulla progettazione flessibile del sistema.

Tutti questi concetti devono essere allargati anche in un'*ottica temporale* rispetto al ciclo di vita del modello: un modello che rispetti tutti i requisiti di accuratezza, affidabilità, robustezza, sicurezza e resilienza può perdere ciascuna di queste qualità nel tempo, ad esempio a fronte dei cambiamenti intercorsi nell'ambiente da cui provengono i dati, oppure per la mancanza del modello di "dimenticare" i dati passati. Per tale motivo, si richiede un monitoraggio continuo dei fattori di rischi, al fine di assicurarsi che il modello risponda correttamente alle domande per cui è stato

11 Con "validità del costrutto" nelle scienze sociali si intende l'adeguatezza delle inferenze fatte sulla base di osservazioni o misurazioni (spesso punteggi dei test), in particolare se un test misura il costrutto (o astrazione) previsto.

12 La "validità esterna" si riferisce alla capacità dei risultati di un esperimento di essere generalizzati oltre lo studio immediato.

costruito e manutenerlo costantemente.

Attributi tecnico-sociali

- *Explainability (o "chiarezza")*: aumentare l'explainability significa cercare di fornire una descrizione programmatica di come vengono generate le previsioni del modello. L'assunto di base è che le percezioni del rischio derivino da una mancanza di conoscenze tecniche di base da parte dell'utente. Anche date tutte le informazioni necessarie per rendere un modello completamente trasparente, un essere umano deve applicare le competenze tecniche di cui dispone per capire come funziona il modello. L'explainability si riferisce alla percezione dell'utente di come funziona il modello, come ad esempio quale output può essere previsto per un determinato input. I rischi dovuti alla mancanza di chiarezza possono sorgere se gli esseri umani deducono erroneamente il funzionamento di un modello e non funziona come previsto. Questo rischio può essere gestito da descrizioni di come funzionano i modelli per i livelli di abilità degli utenti. I modelli c.d. "black box", per i quali non è chiaro come viene generato il risultato, sono esempi di modelli non chiari. Si sottolinea, infine, che il concetto di chiarezza è diverso dal concetto di trasparenza: la trasparenza non garantisce l'explainability, soprattutto se l'utente non ha una comprensione dei principi tecnici del machine learning.
- *Interpretabilità*: l'ipotesi di base è che le percezioni del rischio derivino da una mancanza di capacità di dare un senso o contestualizzare l'output del modello in modo appropriato. Ad esempio, i modelli sono sviluppati per un particolare uso funzionale. L'interpretabilità del modello si riferisce alla misura in cui un utente è in grado di determinare l'aderenza a questa funzione e le conseguenti implicazioni di questo output su altre decisioni consequenziali. Un esempio per capire bene il concetto è considerare un modello che decida se la privacy o la sicurezza siano rispettate o meno in dato luogo o contesto: il risultato può essere condiviso o meno a seconda degli individui che lo analizzano in base alla propria concezione valoriale. I rischi per l'interpretabilità possono spesso essere affrontati attraverso la comunicazione dell'interpretazione prevista dai progettisti di modelli, sebbene questa rimanga un'area aperta di ricerca. Tuttavia, la prevalenza di diverse interpretazioni può essere facilmente misurata con

strumenti psicometrici. L'interpretabilità è il collante che lega la trasparenza – le informazioni fornite insieme all'output di un modello – alle determinazioni che hanno a che fare con i valori (ad esempio, privacy, sicurezza). Dato uno stimolo trasparente mostrato a un decisore, possono quindi applicare i loro valori per interpretarlo e determinare se è, ad esempio, "sicuro" o "non sicuro".

La trasparenza non garantisce l'interpretabilità. Spesso, l'interpretabilità è associata a rappresentazioni semplici, mentre la trasparenza può creare un sovraccarico di informazioni. L'interpretabilità si basa sulla capacità dell'utente di "collegare i puntini" date le informazioni fornite da un sistema più trasparente. Ciò significa che ci devono essere punti da collegare, ma anche che le informazioni sono presentate in modo tale che l'utente possa creare una comprensione coerente dell'uso del modello nel contesto.

- *Privacy*: le caratteristiche tecniche specifiche di un sistema possono promuovere la privacy affinché i valutatori possano identificare come il trattamento dei dati potrebbe creare problemi legati alla privacy. Tuttavia, le determinazioni della probabilità e della gravità dell'impatto di questi problemi sono contestuali e variano tra culture, individui e regolamentazioni. Inoltre, garantire l'equità può richiedere la violazione della privacy e viceversa (poiché le determinazioni di equità spesso richiedono l'ottenimento di dati che alcuni considerano privati).
- *Sicurezza ("Safety")*: nel contesto dei dispositivi medici e dei farmaci, la sicurezza è una determinazione categorica fatta da esperti del settore: un farmaco è considerato "sicuro ed efficace" o non lo è. Queste determinazioni sono fatte in relazione allo stato dell'arte sul campo e rispetto alle aspettative della società. Le agenzie di regolamentazione, come la Food and Drug Administration degli Stati Uniti (FDA), in genere mantengono misure per la sicurezza in un determinato contesto; tuttavia, queste misure sono soggette a revisione, spesso con il contributo dei professionisti (NIST, 2021).
- *Bias*: le distorsioni o pregiudizi sono un elemento costante nell'IA, meglio si vedrà nel prossimo capitolo dedicato. I bias non possono essere eliminati completamente. Piuttosto, i bias dannosi devono essere identificati e, per quanto possibile, compresi, misurati, gestiti e ridotti.

Tecnologico

- *Cybersecurity*: esistono potenziali debolezze nei sistemi di AI e nell'ambiente in cui essi sono situati. La maggior parte dei potenziali attacchi contro i sistemi di intelligenza artificiale possono essere raggruppati nelle seguenti quattro categorie (AIRS, 2021):
    - *Data Privacy Attacks*: un utente malintenzionato è potenzialmente in grado di dedurre il set di dati utilizzato per addestrare il modello, compromettendo così potenzialmente la privacy dei dati. Un hacker potrebbe potenzialmente dedurre informazioni riservate dal set di dati di training analizzando i parametri o interrogando il modello;
    - *Training Data Poisoning*: l'avvelenamento dei dati è la contaminazione dei dati utilizzati per addestrare il sistema AI / ML, influenzando negativamente il suo processo di apprendimento o output. L'avvelenamento dei dati potrebbe potenzialmente essere utilizzato per aumentare il tasso di errore del sistema AI / ML o per influenzare potenzialmente il processo di riqualificazione. Alcuni degli attacchi in questa categoria sono noti come attacchi "label-flipping" e "frog-boil";
    - *Adversarial Inputs*: i sistemi di intelligenza artificiale che utilizzano input da sistemi / utenti esterni, interpretano l'input ed eseguono le azioni per cui sono programmati. Un hacker potrebbe potenzialmente utilizzare un input dannoso o un payload esplicitamente progettato per aggirare il classificatore dei sistemi di intelligenza artificiale;
    - *Model Extraction*: in questi attacchi, un hacker cerca di rubare il modello stesso. Gli attacchi di estrazione del modello sono potenzialmente quelli più impattanti, poiché il modello rubato potrebbe essere utilizzato come "strumento" per creare rischi aggiuntivi. La ricerca su tali attacchi indica che, data la capacità illimitata di interrogare il modello, l'estrazione potrebbe avvenire senza richiedere alti livelli di sofisticazione tecnica e potrebbe essere eseguita ad alta velocità.
- *Infrastruttura IT*: La dipendenza significativa, in alcuni casi, delle applicazioni di intelligenza artificiale dai big data aumenta il rischio rappresentato dall' infrastruttura IT esistente, in quanto quest'ultima potrebbe non essere compatibile con l'IA (ad esempio sistemi incapaci di elaborare big data, data

streamline insufficiente, capacità manutentiva, obsolescenza, etc.).

Governance e Risorse Umane

- *Competenze*: un'organizzazione deve possedere le giuste competenze e grado di conoscenza dei sistemi di IA che sono in essa compresi. Le competenze non devono essere possedute solo dal dipartimento di data science e IT di sviluppo degli algoritmi, ma si richiedere una conoscenza sufficiente per tutti gli utenti coinvolti direttamente o indirettamente nell'uso affinché abbiano consapevolezza delle implicazioni di un sistema di IA (garantendo quindi la *chiarezza*, *l'interpretabilità*, la *privacy*, la *sicurezza* e il riconoscimento di *bias*). Questa necessità si riflette nel continuo re-skilling delle risorse umane dell'impresa e nella definizione di nuove conoscenze necessarie richieste nella fase di recruitment. È essenziale per il corretto funzionamento del sistema che l'uomo sia "incluso nella regola" dell'algoritmo che, essendo in continuo apprendimento, necessità di essere messo in discussione. Si deve evitare l'errore di escludere l'uomo dai processi e dalle decisioni dell'IA. Un'IA che decide solo su regole e processi rischia di aumentare la burocrazia all'interno dell'impresa (c.d. "*Algocracy*") (Duranton *(BCG)*, 2020)[13].
- *Correttezza d'uso*: con rischio di correttezza d'uso si fa riferimento ad un uso improprio dell'algoritmo di IA, ovvero un uso diverso di un modello rispetto al motivo per il quale è stato costruito ed allenato. Inoltre, il fatto che la tecnologia sia teoricamente in grado di eseguire un compito non significa che debba necessariamente essere applicata per quel compito: Se applicare una soluzione IA per un problema specifico è una questione aziendale ampia, che racchiude considerazioni commerciali, normative e di accettazione del cliente. che vanno oltre la capacità tecnica della tecnologia di eseguire un compito. Un esempio: sarebbe improprio utilizzare il solo output di un modello che analizza il merito di credito di un cliente per approvare o meno un'erogazione, in quanto è richiesta la supervisione e la revisione del personale specializzato. Per mitigare tale rischio è utile redigere tenere aggiornato un inventario che consenta all'organizzazione di identificare e tenere traccia dei sistemi di AI/ML,

13 Si consiglia la visione del TED Talk di Sylvain Duranton, Managing Director & Senior Partner, Global Leader, BCG GAMMA Paris, disponibile su Youtube al seguente link: https://www.youtube.com/watch?v=2KMk1lJGPlk

monitorando le informazioni, gli usi e rischi associati se presenti. Tale inventario potrebbe descrivere lo scopo per il quale il sistema è progettato, l'uso previsto e le eventuali restrizioni su tale uso. Gli inventari potrebbero anche elencare gli elementi di dati chiave per ciascun sistema AI/ML, inclusi eventuali sistemi/modelli di alimentazione, proprietari, sviluppatori, validatori e date chiave associate al ciclo di vita AI/ML. Le organizzazioni possono trarre vantaggio dall'implementazione di protocolli, strutture, framework e strumenti per aiutare a mantenere un inventario dei sistemi accurato e completo.

- *Governance & Responsabilità (Accountability)*: L'organo amministrativo e le figure manageriali devono essere le prime per avere consapevolezza dell'intelligenza artificiale e stabilire un'adeguata governance dell'uso e dei rischi con le rispettive responsabilità. Si assume che il sistema non possa essere mai considerato responsabile. Al fine di stabilire un'adeguata governance, un'organizzazione dovrebbe:
    - dare la propria definizione di cosa sia IA, al fine di contornare concretamente quale sia l'infrastruttura oggetto della policy;
    - costruire un inventario che tracci i sistemi di IA utilizzati all'interno dell'organizzazione (come sopra descritto nella correttezza d'uso);
    - stabilire policy e standard che assicurino l'utilizzo appropriato degli algoritmi di IA, comprendendo i principi etici (che potrebbero essere ripresi dalle istituzioni pubbliche elaborano tali ragionamenti).
    - Adottare un framework utile all'organizzazione a imparare, governare, monitorare ed acquisire esperienza riguardo l'adozione dell'IA. Si parlerà di framework nei paragrafi 2.3 e 2.4.

  In relazione alla responsabilità, si raccomanda di stabilire chiaramente le figure di responsabilità in funzione del rischio e dell'utilizzo dell'algoritmo e stabilire un processo centrale di monitoraggio e di escalation per fornire un'esposizione sufficiente all'interno di un'organizzazione alle decisioni prese e un'opportunità per sollevare preoccupazioni o sfide quando appropriato. Questa struttura dovrebbe consentire al sistema di monitoraggio di adattarsi alle mutevoli esigenze dell'organizzazione, man mano che l'adozione dell'IA matura o si verificano cambiamenti sostanziali nel settore. Un esempio è rappresentato dalle tre linee di controllo adottare generalmente dalle istituzioni finanziarie: La

maggior parte di tali enti segue un modello a tre linee di difesa, che separa i gruppi di prima linea, che sono generalmente responsabili dei rischi aziendali (la prima linea), da altri gruppi di supervisione del rischio e di sfida indipendenti (la seconda linea) e di assicurazione (la terza linea). Il governance framework dovrebbero garantire che nello sviluppo e nell'utilizzo del sistema di IA siano soddisfatti requisiti di business (primo livello), supervisione (secondo livello) e garanzia (terzo livello) sufficienti.

Le nuove esigenze in tema di governance e responsabilità dell'IA ha posto l'attenzione sulla creazione di nuovi, appositi organi aziendali, da considerare utili in relazione alla struttura esistente:

- *ML Operations*: un team operativo ML esegue il provisioning dei dati per l'analisi da parte del team di data science. Possono anche creare e mantenere set di dati allo scopo di addestrare i sistemi di intelligenza artificiale.
- *Data Science*: alcune organizzazioni hanno pratiche mature di Data Science. Oltre alle responsabilità assegnate, il team di Data Science potrebbe gestire l'inventario del sistema di intelligenza artificiale e il controllo della versione.
- *Centro di Eccellenza IA*: un Centro di Eccellenza (CdE) può fornire una piattaforma di condivisione delle conoscenze in un'organizzazione. A seconda dell'organizzazione, un CdE potrebbe creare una visione collettiva e creare e condividere le *best practice* maturate. Inoltre, il CdE potrebbe mantenere l'impegno con l'industria per condividere e apprendere le migliori pratiche.
- *Comitato Etico*: un comitato di revisione etica può rivedere i progetti di intelligenza artificiale in conformità con i principi etici di un'organizzazione, ad esempio, un algoritmo considerato ad alto rischio.

Compliance

- *Protezione dei dati*: l'organizzazione deve assicurarsi che l'utilizzo dei dati personali impiegati per alimentare i modelli di IA/ML rispetto quanto previsto dalla normativa sul trattamento dei dati (in Europa il *GDPR*) e assicurare la trasparenza nei sistemi di decisione automatica. Per mitigare questo rischio, è

possibile introdurre del "rumore" ("*noise*") nel dataset affinché i dati sensibili vengano criptati.

- *Regolamentazione*: sebbene non esista ancora un insieme di norme dedicate ai sistemi di intelligenza artificiale, altre regolmanentazioni esistenti (es. legate all'uso del trading algoritmico o lo stesso GDPR) possono essere adottate in analogia alla fattispecie in oggetto di legge. Infine, si richiede una costante attenzione alle nuove normative in materia che senza dubbio saranno emesse, assicurandosi quindi che gli algoritmi siano sviluppati in linea con i principi e i requisiti richiesti per legge.

Impatto sulla società

Valutazione delle esternalità: è importante valutare le conseguenze delle applicazioni di IA che sono vendute o rese disponibili a utenti esterni per comprenderne le implicazioni etiche e culturali e verificare se sono in linea con la policy aziendale. Gli esempi più lampanti sono quelli rappresentati dai social media: se il meccanismo che suggerisce agli utenti i contenuti da visualizzare propone post che incitano alla violenza o all'odio? O se il sistema viene utilizzato in maniera impropria per influenzare le decisioni e le opinioni degli utenti? Tra questi, il caso più noto è lo scandalo dei Cambridge Analytica[14]. Oppure, l'utilizzo di riconoscimento facciale che Facebook ha deciso recentemente di non utilizzare[15]. Se sottovalutate oppure non controllate, le esternalità negative, oltre che ad un danno alla società, si configurano spesso come danno reputazionale che influisce sia sull'azienda ma anche sulla credibilità stessa delle soluzioni di IA (vedasi poi il rischio mercato). Alcuni ambiti oggetto di analisi e di continuo studio in questo senso sono le macchine autonome (automobili autonome, armi letali automatiche, …) sistemi di sorveglianza che sfondano la sfera della privacy e il loro uso improprio, sistemi legati al mondo della salute e delle diagnosi, robotica intelligente.

Mercato

Si potrebbe definire un rischio di mercato legato alla credibilità e agli investimenti nella tecnologia, per quanto essi possano apparire in completa ascesa. Tuttavia, si può

14 https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal
15 Tra le fonti, si segnala l'articolo de "Il Post": https://www.ilpost.it/2021/11/03/facebook-riconoscimento-facciale-privacy/

configurare un aumento del rischio sistemico derivante dall'accentramento delle soluzioni di IA in pochi grandi fornitori.

Fornitore

Tutti i fattori di rischio che sono stati precedentemente catalogati devono essere valutati anche per sistemi di terzi, non costruiti internamente ma forniti da *system provider*. Un algoritmo fornito esternamente non esime da responsabilità l'organizzazione che lo adotta nel caso in cui si rivelasse una "black box". Il sistema esterno deve potersi integrare con l'impresa in conformità con il framework e i principi da essa identificati, esibendo la massima trasparenza abbattendo i problemi di agenzia che possono intercorrere tra impresa cliente e system providers. Inoltre, sono da considerarsi i rischi legati all'utilizzo di risorse *open-source* quali librerie, dataset e linguaggi di programmazione, piuttosto che il rischio di default del fornitore stesso (Deloitte, 2018).

Tassonomia dei rischi legati all'Intelligenza Artificiale	
Categoria di rischio	***Sottocategoria***
Attributi tecnici di modello	Accuratezza
	Affidabilità
	Robustezza
	Sicurezza e Resilienza
Attributi tecnico-sociali	Explainability
	Interpretabilità
	Privacy
	Sicurezza (Safety)
	Bias
Tecnologico	Cybersecurity
	Infrastruttura IT
Governance & Risorse Umane	Competenze
	Correttezza d'uso
	Governance e Responsabilità
Compliance	Protezione dei dati
	Regolamentazione
Impatto sulla società	
Mercato	
Fornitore	

Tabella 1 - Tassonomia dei rischi legati all'uso di IA

2.2 – Dati e bias

Con il termine "*bias*" vogliamo intendere qualsiasi tipo di distorsione o scostamento rispetto ad un valore atteso o alla realtà che possiamo riscontrare sia a livello statistico-matematico che umano. È una caratteristica insita nella natura cognitiva dell'uomo, in quanto condizionata dall'ambiente e dall'esperienza. In quanto umana, i bias si ritrovano sistematicamente in tutti i dati esistenti al mondo (Xiang, 2019). Ad esempio, un decisore umano potrebbe pesare eccessivamente una decisione sulla base della propria esperienza personale (c.d. "ancoraggio"), oppure potrebbe sovrastimare un'informazione che è a sua disposizione (c.d. "*availability*"; un esempio è pensare che fumare non faccia male in quanto conosciamo persone che fumano diversi pacchetti al

giorno e sono in salute) (Baer & Kamalnath, 2017). Oppure, abbiamo la tendenza a selezionare fonti ed evidenze che sono a supporto di una nostra teoria (c.d. "*confirmation bias*"), o ancora, non siamo proni ad accettare un cambio nel nostro credere in quanto abituati ad avere quella convinzione (c.d. "conservazione"). Quando si sente parlare di algoritmi razzisti, è auspicabile che non sia stato il data scientist a voler creare una logica che discriminasse persone di una determinata etnia, piuttosto che di un determinato sesso. È molto più probabile, quindi, che i dati utilizzati per allenare l'algoritmo enfatizzino una realtà distorta rispetto a quella attesa. L'IA al servizio delle Risorse Umane di Amazon che preferiva assumere i maschi di etnia bianca, non era razzista in sé, ma ha raccolto gli stessi bias dell'uomo, osservando sostanzialmente che all'interno dell'azienda stessa la maggior parte dei dipendenti fossero maschi bianchi. Questo, potenzialmente, perché alcuni algoritmi di ML intercettano relazioni non lineari che sono troppo complesse per essere identificate e riviste dagli esseri umani. Queste relazioni hanno il potenziale di causare discriminazioni di trattamento creando proxy per lo status di classe protetta[16] o l'etnia. In una certa misura, queste preoccupazioni sono state attenuate dai progressi nelle tecniche di *reverse engineering* che consentono ulteriori informazioni su queste relazioni complesse.

Anche la specificazione errata del sistema può anche causare esiti discriminatori. In questo caso, l'esito predittivo è indipendente sia del risultato che dello stato della classe protetta, ma l'effetto "classe" è incorporato nella previsione. Ad esempio, supponiamo che un sistema di merito di credito includa se una persona tende a fare acquisti in un discount. È probabile che una tale variabile catturi una misura della ricchezza, che può essere un ragionevole predittore della capacità di rimborso, ma può anche catturare involontariamente una discriminazione. Inoltre, se è più probabile che il negozio si trovi in quartieri di minoranza, il sistema potrebbe esacerbare ulteriormente questo effetto. In questo modo, la variabile può fungere da *proxy* per il quartiere, che a sua volta funge da *proxy* per l'etnia.

L'algoritmo di IA riuscirà a prendere decisioni meno distorte rispetto a quella di un umano o accentuerà i bias? Questa è la domanda da porsi.[17] In molti casi è stato

16 Con "classe protetta" si fa riferimento all'etnia, la religione, il sesso, l'orientamento sessuale o qualsiasi altra caratteristica che possa essere oggetto di discriminazione

17 La domanda non è banale e richiede un notevole sforzo: in primis, nel dare una definizione di "bias", o "*distorsione*" ed essere in grado di identificare tali oggetti. Poi, assumendo di aver eliminato tutti i bias da un algoritmo, possiamo assumere che esso sia anche *equo*? O *inclusivo*? Queste domande fanno riflettere su quanto l'IA debba essere una pratica multidisciplinare e non solo tecnica, comprendendo esercizi di filosofia ed etica.

dimostrato che l'IA può ridurre la soggettività umana nell'interpretare i dati, perché gli algoritmi di *Machine Learning* imparano *solo* dalle variabili che migliorano l'accuratezza delle loro previsioni (McKinsey Global Institute, 2019). Ad esempio, nel 2002 un team di economisti ha studiato l'impatto di algoritmi di sottoscrizione automatica nell'industria dei prestiti ipotecari. Il risultato principale fu che quei sistemi predicono il default più accuratamente di quanto i sottoscrittori manuali facciano. O ancora, uno studio condotto dalla Columbia Business School ha studiato le performance di un algoritmo di job-screening presso una software house: quando l'algoritmo si esprimeva sulle persone che la società avrebbe dovuto intervistare, il sistema favoriva più candidati "non usuali" rispetto ai recruiter umani; in poche parole, l'algoritmo risultava molto meno distorto rispetto ai candidati le cui caratteristiche erano meno presenti all'interno dell'azienda (come quelli senza le referenze personali o quelli senza un titolo di studio da un'università prestigiosa) (Miller, 2018). Inoltre, a differenza delle decisioni umane, le decisioni prese da un'IA possono principalmente essere spiegate, esaminate ed interrogate.

L'intelligenza artificiale vuol simulare la meccanica del nostro cervello attraverso le reti neurali; se quindi i *bias* influenzano l'intelligenza umana, essi influenzano anche l'intelligenza artificiale, per almeno una serie di ragioni:

1. Gli algoritmi di Machine Learning sono proni ad incorporare i bias dei loro stessi creatori umani. Ad esempio, potrebbero formalizzare dei parametri distorti da coloro che hanno popolato / generato i dati. Gli algoritmi che predicono dei comportamenti umani e che necessariamente si basano su dati storici, tali saranno influenzati da distorsioni passate. Gli algoritmi potrebbero quindi anche esacerbare questi bias: è quello che accade nei Social Media, quando l'algoritmo filtra notizie o contenuti in base alle nostre preferenze ed alimenta il *confirmation bias*.
2. I dataset utilizzati potrebbero facilmente essere incompleti o limitati: alcuni modelli americani sui mutui ipotecari costruiti prima della crisi finanziari non potrebbero matematicamente accettare delle variazioni negativi nei prezzi delle case (in relazione ai tassi di interesse negativi che, finché non sono apparsi nel mondo reale, nessun algoritmo avrebbe potuto prevedere) (Baer & Kamalnath, 2017).

Alcuni ricercatori distinguono il modello di IA in due algoritmi differenti: il "*trainer*", che può essere *biased* a causa dei dati sottostanti o per il processo di training, e lo

"*screener*", che semplicemente esegue le predizioni in base al *trainer*. Un bias può avere origine in ogni fase del ciclo di vita del set di dati: creazione, design del modello, campionamento, raccolta ed elaborazione (Gao, Santinelli, & Singer, 2020). Il punto del ciclo di vita del dataset in cui emerge la distorsione è strettamente legato alla causa principale della distorsione:

- *Creazione*: Bias esistono già nei dati a causa delle dinamiche storico-sociali. Ad esempio, il bias di creazione può essere trovato nelle pratiche di reclutamento. Nel 2015, le donne detenevano il 47% di tutti i posti di lavoro negli Stati Uniti, ma solo il 24% delle posizioni di scienza, tecnologia, ingegneria e matematica (STEM). Se un'azienda costruisce un sistema di reclutamento basato sull'intelligenza artificiale per le posizioni STEM utilizzando i dati storici di reclutamento senza adottare misure di mitigazione appropriate, è probabile che il sistema abbia un pregiudizio nei confronti delle candidate donne a causa di pregiudizi storici e mancanza di rappresentazione diversificata nei dati di formazione;
- *Design*: Il modo in cui viene progettato un sistema di intelligenza artificiale o machine learning può introdurre distorsioni dei dati. Se la progettazione del sistema e (compresa la progettazione del prodotto, la progettazione di esperimenti, la progettazione di sondaggi, ecc.) è intrinsecamente distorta, anche i dati raccolti saranno distorti;
- *Campionamento (Sampling)*: I bias da campionamento nascono quando la popolazione del dataset di training non sono rappresentativi della popolazione totale oggetto del sistema di IA. Un esempio: Nel 1936, una previsione elettorale di un quotidiano predisse la vittoria di Alfred Landon con i 57% dei voti rispetto al 43% che avrebbe ottenuto Franklin D. Roosevelt, quando quest'ultimo diventò Presidente con un grande margine (62%). Sebbene la dimensione del campionamento del sondaggio fosse molto grande (2,4 milioni), l'esito è stato minato dal bias di campionamento. Questo perché il campione era basato su elenchi telefonici, iscrizioni a club ed elenchi di abbonamenti a riviste. Nel bel mezzo della Grande Depressione, ciò significava per lo più elettori della classe media e alta, che non erano rappresentativi della popolazione votante.
- *Raccolta*: i bias da collezione possono emergere da tre fonti:
    - In primo luogo, le persone che raccolgono o etichettano il set di dati

possono avere un pregiudizio personale. Questo è un evento comune nello sviluppo di modelli di riconoscimento delle immagini, come le reti neurali convoluzionali, perché richiedono un enorme set di addestramento etichettato ("taggato"). Molte aziende utilizzano servizi di etichettatura dei dati online per ridurre al minimo i costi di generazione di set di dati etichettati per il ML. Se i lavoratori che stanno facendo l'etichettatura fossero viziati da pregiudizi inconsci o formazione culturale, tali potrebbero infiltrarsi nel processo di etichettatura dei dati. Ad esempio, i lavoratori che hanno avuto un'esposizione limitata alla comunità LGBTQ+ o una formazione inadeguata potrebbero etichettare erroneamente una foto di una coppia dello stesso sesso;

  - Secondo, gli "*outliers*" (punti dati che differiscono in modo significativo dagli altri) o i dati errati vengono raccolti a causa di un errore della macchina. Ad esempio, parti di sensori su una macchina sono rotte e restituiscono valori anomali. Senza controllo e mitigazione, questi valori anomali potrebbero compromettere l'addestramento del modello e persino le prestazioni dell'intero sistema di intelligenza artificiale;
  - In terzo luogo, gli utenti sono spesso riluttanti a valutare i prodotti. Netflix ha scoperto, infatti, che quando ha sostituito il suo sistema di valutazione a cinque stelle con un sistema di pollice in su / pollice in giù, il numero di valutazioni è aumentato del 200%. Inoltre, pochissime persone scrivono recensioni e le persone sono molto più propense a scrivere recensioni se si sentono molto fortemente sul prodotto in questione. Di conseguenza, le recensioni degli acquisti online sono solitamente polarizzate e i clienti nel mezzo sono meno rappresentati.

- *Elaborazione*: la distorsione dei dati può essere introdotta anche quando i dati vengono elaborati in preparazione per il training del modello. Varie tecniche di ingegneria dei dati possono contribuire a migliorare le prestazioni del modello. Questi includono la compilazione di valori mancanti, la normalizzazione e la "tokenizzazione" (spezzando il contenuto del testo in pezzi più piccoli, "token"). Ma il pregiudizio può essere introdotto nei dati se tali tecniche vengono implementate senza una piena comprensione del contesto specifico. Prendiamo, ad esempio, un set di dati demografici in cui al 10% delle entità mancano i dati della colonna che rappresenta le misurazioni dell'altezza. Riempire i buchi con

un valore mediano di tutti i dati in quella colonna introdurrà pregiudizi nel set di dati perché le differenze nelle altezze di uomini e donne non sono state prese in considerazione.

Le imprese possono prendere delle contromisure per eliminare i bias o per proteggersi contro gli algoritmi distorti. Ci sono tre misure che un'organizzazione deve adottare (Baer & Kamalnath, 2017):

1. Gli utenti dei sistemi di intelligenza artificiale devono avere consapevolezza dell'output che genera l'algoritmo e di capirne il significato. Non possono limitarsi a schiacciare un pulsante e devono andare oltre il risultato, supervisionando il sistema;
2. I data scientists che sviluppano gli algoritmi devono utilizzare campioni che minimizzino la presenza di bias. Senza addentrarsi in concetti troppo tecnici, è sufficiente sottolineare che i dati storici disponibili sono spesso inadeguati, e nuovi, *unbiased* dati devono essere generati attraverso appositi esperimenti controllati;
3. I dirigenti devono essere consapevoli di quando usare e quando non usare gli algoritmi di IA, capendo il *trade-off* tra i due scenari: l'IA offre velocità ed è poco costosa, mentre altri sistemi, quali alberi decisionali, semplici regressioni o il puro *human decision-making* sono approcci che possono garantire più flessibilità e trasparenza.

Dei test possono verificare che distorsioni passate provenienti da decisioni umane non vengano incluse all'interno del sistema. Inoltre, alcuni algoritmi avanzati possono correggere statisticamente alcuni concetti ben definiti di errore, ma essi non possono distinguere tra errori con elevato impatto sul business da quelli con un'importanza marginale.

2.3 – Model Risk Management: Framework ed estensione ai modelli di ML

Storicamente, il settore dei servizi finanziari è stato costantemente sotto la lente regolamentare in riferimento ai modelli di *decision making*. Per affrontare i rischi e l'impatto dall'ampio utilizzo di tali modelli, soprattutto dopo la recessione del 2008, il settore è stato soggetto ad una rigorosa regolamentazione del Model Risk Management (MRM). Ciò ha portato l'industria finanziaria all'acquisire esperienza nella pratica e a maturare l'expertise dell'uso del MRM. Il MRM può essere definito come "*il*

framework metodologico e organizzativo che si propone di definire e identificare i modelli, assegnare la relativa priorità (Tier) ed assicurarne il presidio nell'ambito di una piattaforma centralizzata (Model Inventory), nonché procedere con le attività di valutazione, monitoraggio e mitigazione del rischio" (AIFIRM, 2021).

Già prima della regolamentazione, molte banche effettuava una qualche forma di controllo, compresa la convalida indipendente, sui modelli che incidevano sul bilancio (ad esempio modelli di valutazione dei derivati) o sul capitale regolamentare (ad esempio modelli di rischio di credito o di rischio di mercato). La normativa, come ad esempio la Supervisory Letter 11-7 negli Stati Uniti, ha ampliato l'ambito di applicazione atteso dalle banche a tutti i modelli utilizzati in qualche modo per informare il processo decisionale aziendale e ha introdotto il concetto di rischio modello come rischio da gestire come altri tipi di rischio ben noti (PwC, 2020). Ciò in quanto, citando la FED, "l'utilizzo di modelli presenta inevitabilmente un rischio intrinseco, che è potenzialmente foriero di conseguenze negative di decisioni basate su modelli errati o utilizzati in modo improprio. Il rischio di modello può essere definito come la perdita potenziale derivante da decisioni erronee assunte sulla base delle stime ottenute da un modello, a causa di errori nello sviluppo, nell'attuazione o nell'utilizzo del modello stesso". Sempre la normativa SR 11-7 – "Supervisory Guidance on Model Risk Management" definisce un "modello" come "un metodo, sistema o approccio quantitativo che applica teorie, tecniche e ipotesi statistiche, economiche, finanziarie o matematiche per elaborare i dati di input in stime quantitative". Un modello è, inoltre, costituito da tre diverse componenti:

- un componente di input delle informazioni, che fornisce ipotesi e dati al modello;
- una componente di elaborazione, che trasforma gli input in stime;
- una componente di reporting, che traduce le stime in utili informazioni aziendali (AIFIRM, 2021).

Ecco, quindi, che vi rientrano tutte le tecnologie sottostanti l'intelligenza artificiale.

Dopo la regolamentazione, la maggior parte delle banche (sia piccole che grandi) hanno implementato il proprio framework MRM, che include, generalmente, i seguenti elementi:

- *Test e documentazione dello sviluppatore* (controllo di prima linea): spetta allo sviluppatore del modello (il c.d. *Model Owner*) fornire la giustificazione della

scelta della metodologia e la prova del test prima che il modello venga pubblicato. Si prevede inoltre che il modello sia ben documentato, compresi i dettagli tecnici e le ipotesi chiave del modello. Il Model Owner, anche per delega a funzioni di supporto, è il responsabile dell'identificazione del modello e del suo censimento all'interno dell'inventario (come poi descritto);

- *Revisione e convalida di una parte indipendente* (controllo di seconda linea): tutti i componenti di un modello devono essere esaminati da una parte indipendente. Una funzione indipendente all'interno del team di model risk dovrebbe effettuare una revisione e un test indipendente dettagliato. Gli inevitabili conflitti di interesse che sorgono nello sviluppo del modello danno origine all'importante principio della sfida efficace, che è un importante motore della revisione di seconda linea.

  I controlli di prima e seconda linea devono concentrarsi sulla solidità concettuale del modello, evidenziando le ipotesi e i limiti del modello, i pro e i contro degli approcci alternativi e conducendo test approfonditi del comportamento del modello (ad esempio, analisi di sensibilità, stress test, casi limite e back-test ove pertinente);
- *Revisione periodica e monitoraggio continuo delle prestazioni* del modello, per confermare che il modello continui a funzionare come previsto e che la convalida più recente sia ancora sufficiente. L'*outcomes analysis*, come il backtesting di un modello statistico, è particolarmente rilevante in questo caso;
- *Model Inventory*: l'inventario dei modelli è definito come uno degli elementi fondamentali per la realizzazione di un framework di MRM; In particolare, la normativa della Federal Reserve stabilisce la necessità di "mantenere un inventario a livello aziendale di tutti i modelli, che dovrebbero aiutare una banca (o un'impresa, più in generale) a valutare il rischio del suo modello nell'aggregato", fornendo anche indicazioni sulle informazioni da archiviare nell'inventario dei modelli: "l'inventario dovrebbe descrivere lo scopo e i prodotti per i quali il modello è progettato, l'uso effettivo o previsto e le eventuali restrizioni d'uso. È utile che l'inventario elenchi il tipo e la fonte di input utilizzati da un determinato modello e dai componenti sottostanti (che possono includere altri modelli), nonché gli output del modello e l'uso previsto. Dovrebbe inoltre indicare se i modelli funzionano correttamente, fornire una

descrizione di quando sono stati aggiornati l'ultima volta ed elencare eventuali eccezioni ai criteri. Altri elementi includono il nome delle persone responsabili di vari aspetti dello sviluppo e della convalida del modello; le date delle attività di convalida completate e pianificate; e il lasso di tempo durante il quale si prevede che il modello rimanga valido". Inoltre, "il perimetro di applicazione dell'inventario deve essere riferito a tutti gli strumenti che hanno un impatto sul fronte regolamentare, reputazionale e sulle decisioni aziendali di business. Il processo di completamento dell'inventario potrà, quindi, basarsi su un approccio graduale che preveda il censimento degli strumenti regolamentari di misurazione del rischio in un primo momento e la tracciatura di tutti gli strumenti sviluppati e utilizzati nelle diverse aree dell'organizzazione in fasi successive, a prescindere dalla tipologia (non solo modelli di rischio) e dalla finalità (regolamentare e manageriale). […] Il patrimonio informativo può essere principalmente ricondotto alle seguenti aree tematiche: caratteristiche generali, che comprende le informazioni che consentono di identificare i modelli ed i loro utilizzi presso l'Istituto; elementi di governance, volti ad individuare i ruoli e le conseguenti responsabilità delle Funzioni coinvolte nel processo di governo del rischio modello; ottimizzazione del portafoglio modelli, riguardante l'assegnazione del Tier al modello; valutazione e mitigazione del rischio modello; aspetti documentali inerenti alle diverse fasi del ciclo di vita del modello" (AIFIRM, 2021).

- *Model Risk Tiering*: la classificazione dei modelli consente un'efficace definizione delle priorità, in modo che la maggior parte delle risorse vengano spese per i modelli con il model risk più impattante;
- *Model Usage controls*: la verifica che l'uso previsto del modello sia allineato con l'utilizzo effettivo;
- *Governance del modello e reporting*: ciò richiede un comitato di rischio che approvi modelli della stessa classe di rischio a fronte di una discussione sui rischi identificati. Questo può essere considerato il canale principale per la segnalazione delle metriche di rischio del modello al senior management. È fondamentale che il senior management sia consapevole dei principali problemi di rischio del modello e delle fonti materiali di rischio del modello nell'azienda;
- *Third-party vendors*: i modelli sviluppati da fornitori esterni sono trattati

secondo lo stesso standard dei modelli sviluppati internamente, con l'obbligo per il terzo di fornire documentazione, giustificare la progettazione del modello e l'esito dei test;

- *Internal Auditing*: l'internal audit deve procedure ad ispezioni regolari al fine di assicurare la coerenza e l'efficacia del risk management in tutta l'azienda (controlli di terza linea).

La valutazione del rischio modello è il processo finalizzato a misurare la gravità del rischio modello, sulla base di una metodologia qualitativa e/o quantitativa predefinita, che tiene conto sia della numerosità e gravità delle debolezze del modello e sia dell'efficacia di eventuali azioni di mitigazione del rischio, lungo tutto il ciclo di vita del modello. Il suo risultato consiste in uno *score* di rischio che esprime la misura di rischio sintetica attribuibile a ciascun modello. Tale score combina generalmente elementi di valutazione di tipo qualitativo e quantitativo e si basa su differenti approcci di quantificazione, anche di tipo teorico.

Lo score si può ottenere come combinazione tra fattori di *inherent risk* e fattori di mitigazione.

Tra i fattori di inherent risk sono compresi, ad esempio, basi dati e relative caratteristiche (volumetria, presenza di missing, profondità storica ecc.), assunzioni sottostanti alla specificazione del modello, metodologie di stima adottate, presenza di moduli intermedi, risultati dei diagnostici che derivano dal processo di applicazione delle assunzioni e delle specifiche ai dati disponibili; inoltre, sono compresi altri fattori di rischio modello, come l'incertezza, la complessità e la conformità ai requisiti regolamentari. I fattori di rischio possono essere valutati specificatamente in riferimento sia alle singole componenti di un modello (ad esempio dati, model design, processi/governance e utilizzo) e sia alle diverse fasi del ciclo di vita di un modello (ad esempio sviluppo, implementazione/IT systems, manutenzione/calibrazione).

Tra i fattori di mitigazione sono da considerare, in particolare, le attività di *maintenance* sui modelli, le azioni mitiganti poste in essere per indirizzare le anomalie individuate, i controlli di Convalida e di Audit, oltre ai controlli di secondo livello dedicati ed altri processi decisionali e derogatori. Questi potrebbero essere legati anche ad alcune *practices* come, ad esempio, l'esercizio di deroghe nei processi decisionali qualora le Funzioni Operative non si ritrovino nei risultati messi a disposizione. Tali azioni mitiganti vengono valutate in base alla loro efficacia, sia in termini di esecuzione che di

tempestività.

Per quanto riguarda il processo di valutazione, la responsabilità di valutare i diversi fattori di rischio è in primo luogo in capo al Model Owner, che nella prassi è solitamente chiamato ad esprimersi attraverso questionari di *self assessment* sottoposti a controlli successivi:

- controlli della Convalida Interna: validazione iniziale e periodica del modello, rilascio di *finding* e raccomandazioni, realizzazione di follow up periodici;
- controlli di secondo livello: verifiche afferenti alla corretta applicazione del modello ai processi operativi condotti da Funzioni dedicate;
- revisione della Funzione Internal Audit: realizzazione di audit sul modello con relativo rilascio di *opinion*, raccomandazioni e relativo monitoraggio.

In pratica, lo *score* di model risk è la misura che esprime sinteticamente il rischio associato a ciascun modello e può essere definito come il valore "netto" risultante dalla valutazione dei rischi specifici ed ulteriori, ridotta per effetto delle mitigazioni, e risulta fondamentale per l'organizzazione al fine di orientare il presidio complessivo sui modelli e le strategie di gestione correlate. Lo score di model risk assegnato può essere poi collegato ad un processo decisionale, al fine di indirizzare le strategie di gestione dei modelli, lungo l'intero ciclo di vita, e massimizzare il presidio complessivo dell'impresa in termini di *model governance* (AIFIRM, 2021). Tale procedura diviene fondamentale per allineare il rischio misurato con il *risk appetite* definito a priori dall'organizzazione.

L'approccio del MRM è decisamente strutturato negli istituti finanziari, dato che concorre alla determinazione dei requisiti patrimoniali fissati per legge, ma molto meno nelle imprese non regolamentate. Soprattutto, rispetto alla pratica che si è sviluppata nelle imprese più esperte, il MRM soffre rispetto all'intelligenza artificiale alcune debolezze:

- Il MRM è tipicamente basato su una valutazione periodica (ad esempio, ogni anno o a frequenze maggiori), assumendo che i modelli siano statici tra una revisione e l'altra. I modelli di IA imparano dai dati, e la loro logica cambia ogni volta che vengono riaddestrati per imparare su nuovi dati;
- I processi tradizionali di MRM sono hanno un carattere sequenziale (o c.d. "a cascata") e richiedono in media dalle sei alle 12 settimane di lavoro dopo che il modello sia stato sviluppato. Questo processo, oltre a ritardare l'utilizzo del

modello, poco si adatta alla metodologia agile[18] di solito utilizzata per lo sviluppo dei modelli di IA;

- Il MRM è spesso focalizzato sui tradizionali rischi finanziari (ad esempio, il rischio di credito o di mercato) e potrebbe non coprire completamente la vasta e nuova gamma dei rischi derivanti dalle diverse sfaccettature dell'IA, come il rischio reputazionale e i rischi per il consumatore;
- Alcune applicazioni di IA, come le chatbot, soluzioni di NLP e di HR analytics, possono essere qualificate come modelli sotto la definizione regolatoria in ambito bancario. Tuttavia, queste applicazioni sono totalmente diverse rispetto ai modelli tradizionali (come quelli di asset o capital pricing, stress-testing e credit-risk) e non è facilmente applicabile il MRM (McKinsey, 2020).

Tuttavia, il framework e le "best practice" del Model Risk Management sono probabilmente un punto di partenza necessario, seppur non sufficiente, per la gestione del rischio del modello di IA. È necessario in primis stabilire quali sono i rischi specifici che sfidano il MRM classico:

- Un rischio che è più pronunciato nei sistemi di intelligenza artificiale è la natura dinamicamente in evoluzione della tecnologia (nuove tecniche continuano a essere generate a un ritmo rapido) e la sua relativa complessità. Ciò può rendere più difficile per i professionisti avere una comprensione completa dei pro e dei contro delle diverse tecniche, creando implicazioni per la consapevolezza e la mitigazione del rischio;
- Con riguardo all'interpretabilità, la capacità di spiegare le ragioni alla base delle raccomandazioni di un modello è anche strettamente correlato alla verificabilità e alla responsabilità dei modelli, che sono altre aree di grande interesse per i responsabili politici e i consulenti politici;
- Rispetto ai modelli tradizionali a cui è applicato il MRM, devono essere considerate l'esistenza dei bias e il concetto di equità del sistema;
- Mentre i modelli in altre aree possono essere costruiti sulla base di principi noti o relazioni matematiche stabilite, i modelli di intelligenza artificiale hanno bisogno di molti dati da cui imparare. Ciò significa che l'uso di questa tecnologia

18 La metodologia "agile" si contrappone a quella tradizionale "a cascata": mentre nella seconda il team segue in maniera precisa degli step ben definiti di sviluppo (a cui si può sovrapporre il MRM), la metodologia agile consiste nella programmazione del lavoro in sprint, condotti parallelamente dai membri del team a seconda dei task che vengono percepiti importanti e/o urgenti. Non essendoci una

è strettamente correlato alla gestione di grandi quantità di dati che comportano tutti i rischi rilevanti per i dati: privacy, diritti di proprietà intellettuale, idoneità dei dati di formazione, ecc. Le aziende o le divisioni che potrebbero non essere state storicamente impegnate in attività ad alta intensità di dati potrebbero dover costruire tali capacità e strumenti di gestione del rischio insieme all'adozione dell'IA nelle loro operazioni regolari.

- L'uso frequente di modelli *vendor* o pre-addestrati comporta rischi legati alla progettazione del modello, alla trasparenza e alla rilevanza dei dati di training. Per tali strumenti di terze parti, altri rischi come i bias e la fairness possono essere più difficili da tracciare e controllare.
- Altri rischi peculiari riguardano:
    - Stabilità, ovvero il modello che è generalmente accurato, non distorto e spiegabile ma dà risultati incoerenti in casi limite;
    - Cybersicurezza;
    - Continuous learning, ovvero il miglioramento del modello in base di nuovi dati, il tutto *in automatico*. L'utilizzo dell'apprendimento automatizzato in produzione senza una sufficiente comprensione della natura dei modelli sottoposti a addestramento non deve essere raccomandato. Mantenere la capacità interna di comprendere la natura dei rischi dei modelli in uso è essenziale.

Ecco, quindi, degli spunti per integrare il framework del MRM affinché risulti più solido per i modelli di machine learning o di altro tipo di intelligenza artificiale (PwC, 2020):

1. Tutte le istanze di applicazioni AI devono essere inserite nell'inventario dei modelli e contrassegnate con il *tag* "AI". L'uso di questa etichettatura è quello di consentire l'applicazione dei seguenti sei passaggi a tutti i modelli di IA;
2. I criteri di valutazione del rischio del modello dovrebbero essere aggiornati per i modelli di IA in modo da includere quegli specifici fattori di rischio dell'IA, come, ad esempio, l'*explainability*, chiarezza, l'interpretabilità o capire come un modello influisce sugli individui o se utilizzi variabili eticamente sensibili;
3. Deve essere registrato come vengono affrontati i problemi di *bias* e *fairness*, comprovare l'analisi con un contraddittorio eseguito da una parte indipendente

consequenzialità, ecco che non può essere applicato il tipico workflow

dal gruppo di lavoro proprietario (in genere, da una funzione di seconda linea). Devono essere definite e utilizzate tecniche standardizzate per rilevare la distorsione del modello rispetto a fattori sensibili e per affrontare tali distorsioni;

4. Laddove il modello si comportasse come una *black box* ola mancanza di specifiche assunzioni porti a una mancanza di una chiara comprensione del motivo delle decisioni di un modello di IA, è necessario un approccio definito per determinare se le tecniche di interpretabilità utilizzate sono appropriate. In alcuni casi, le tecniche e i *tool* dedicati all'*explainability* dovrebbero essere classificate come modelli, con l'approvazione e i controlli pertinenti.
5. Per ogni uso distinto di uno strumento di intelligenza artificiale, è necessario registrare un processo decisionale sul fatto che questa sia una scelta appropriata: il costo di una minore interpretabilità giustifica il guadagno in termini di prestazioni rispetto a modelli lineari più semplici? Gli utenti hanno una sufficiente esperienza e la gestione del rischio? Si tratta di un'applicazione standard del settore? Ci sono problemi di privacy dei dati?
6. Il modello di intelligenza artificiale può cambiare il comportamento man mano che apprende da dati aggiuntivi. Mentre il monitoraggio continuo delle prestazioni del modello è una parte ben consolidata dei framework MRM esistenti, il fatto che i modelli di intelligenza artificiale stiano prendendo decisioni basate su *features* ad alta dimensione potrebbero non essere ovvie per l'occhio umano; ciò implica che i risultati del modello possono cambiare qualitativamente man mano che arrivano nuovi dati di addestramento. Per questo motivo, i test di accuratezza del modello, così come la distorsione e la chiarezza, devono essere monitorati regolarmente su una base più frequente di quanto potrebbe essere il caso per i modelli tradizionali.
7. Un comitato per i rischi dell'IA dovrebbe essere istituito in sovrapposizione al modello di governance esistente. Ciò prenderebbe in considerazione la popolazione complessiva dei modelli di IA e l'implementazione del *framework MRM IA* esteso. Il comitato dovrebbe monitorare i rischi emergenti dell'IA, evidenziare le aree di governance debole e garantire che l'uso dell'IA rientri nella propensione al rischio dell'impresa.

In termini di risorse, per garantire un'efficace MRM AI, le organizzazioni che utilizzano tali sistemi devono mantenere personale con competenze pertinenti in tutte e tre le linee

di difesa e in tutta la gerarchia verticale, compresi gli organi decisionali competenti. Le organizzazioni con una significativa esperienza nel MRM e il rispettivo livello di consapevolezza del rischio del modello (es. banche ed assicurazioni) dovrebbero essere in grado di adattarsi ai rischi del sistema di IA con aggiustamenti piuttosto modesti. Altri, per i quali questa tecnologia è un mezzo per applicare per la prima volta il processo decisionale assistito da modelli su larga scala e che devono ancora sviluppare un pensiero consapevole del rischio del modello, dovranno intraprendere un viaggio più lungo. Per quest'ultimo, attingere alle migliori pratiche MRM esistenti potrebbe essere particolarmente utile. Se un'impresa è nuova della tecnologia e volesse adottare una soluzione di IA per la propria organizzazione, un approccio pratico per la valutazione della preparazione di un'organizzazione include la risposta alle seguenti domande (PwC, 2020):

- Sono in grado di spiegare come funziona il modello o la tecnologia selezionata e quali sono i suoi vantaggi e svantaggi relativi?
- Sia i proprietari dei modelli che il senior management sono consapevoli dei rischi e dei limiti della tecnologia?
- Qual è il loro livello di esperienza con questa tecnologia?
- Hanno visto solide prove industriali e accademiche per l'applicazione della tecnologia selezionata per il caso in questione?
- Quale parte della creazione del modello controllano (ad esempio costruendolo da zero, riutilizzando modelli esistenti da casi simili, utilizzando prodotti "acceleratori" di fornitori terzi, utilizzando modelli pre-addestrati "pronti all'uso" o modelli di fornitori)
- Hanno il controllo sui dati su cui è stato addestrato il modello e sono sicuri che i dati di addestramento siano rappresentativi della popolazione target?
- Hanno valutato il costo relativo della maggiore complessità e della diminuzione dell'interpretabilità rispetto al potenziale guadagno in termini di prestazioni, tenendo conto del contesto di utilizzo del modello?
- La gestione è soddisfatta del risultato di questa valutazione in un senso aziendale più ampio (non puramente il grado di prestazioni tecniche del modello)?

Infine, è stato rilevato che molte delle questioni rilevanti per il rischio non sono di natura tecnica ma, piuttosto, gestionale, e richiedono pertanto un grado significativo di giudizio manageriale. Pertanto, mentre l'automazione svolge un ruolo importante per il

monitoraggio delle prestazioni del modello, i componenti chiave per una solida convalida del sistema di intelligenza artificiale e la gestione del rischio sono ancora politiche di governance formalizzate e risorse umane con le competenze appropriate per implementare tali politiche.

2.4 – Derisking by Design

Come si è evidenziato, i modelli tradizionali di identificazione, misurazione e gestione del rischio sono in difficoltà di fronte alle sfide che gli algoritmi di IA implicano, mentre è chiaro l'impatto che possono avere i fattori di rischio non adeguatamente valutati. Quando si tratta di intelligenza artificiale, il risk management non può essere pensato dopo, o limitarsi ad essere una funzione di *model-validation*, come avviene per i modelli tradizionali negli istituti finanziari. Le imprese necessitano di costruire una gestione del rischio direttamente nello sviluppo dei propri progetti di IA, affinché il controllo sia costante e contemporaneo tra la produzione e l'implementazione nell'impresa. Questo approccio viene denominato da McKinsey[19] "Derisking by design", ovvero "mitigando dall'inizio".

La sfida di gestire i rischi associati all'IA non è banale, soprattutto per la compresenza di alcuni fattori: l'IA è difficile da tracciare all'interno dell'impresa. L'uso di tali soluzioni è in continuo aumento e, in alcuni casi, queste sono decentralizzate all'interno dell'azienda stessa, rendendo quindi difficile per i risk manager da tracciare. Inoltre, alcune soluzioni di IA sono incorporate all'interno di hardware o software prodotti da terzi (ad esempio, il software CRM), utilizzate in unità di business diverse; queste introducono potenzialmente un nuovo tipo di rischio non controllato. Inoltre, la gestione del rischio connesso a IA implica delle scelte nell'implementazione non banali per imprese senza una funzione di risk management consolidata. Non è immediato chiedersi per imprese che non hanno mai maturato un expertise come dovrebbe essere gestito il rischio reputazionale per un'impresa globale: accentrato o decentrato per unità nazionali? Oppure, come combinare la gestione dei rischi di IA con gli altri rischi, come data privacy, cybersecurity e eticità dei dati?

Per vincere la sfida senza condizionare lo sviluppo *agile* delle soluzioni di intelligenza artificiale, ecco che bisogna adottare un nuovo approccio, ovvero identificare e mitigare i rischi sin dalla primissima fase di produzione dei sistemi IA. Tale approccio permette

19 *"Derisking AI by design: How to build risk management into AI development"* (J. A. Baquero, R.

agli sviluppatori e ai loro diretti interessati in azienda di costruire modelli che sono consistenti con i valori e il *risk appetite* dell'impresa. Strumenti e concetti come l'interpretabilità del modello, *bias detection* e misurazione delle performance sono incorporati nei modelli, così che il controllo sia costante in tutto lo sviluppo dell'IA e condiviso in tutta l'impresa. Ma non solo: lo stesso controllo deve sussistere anche quando l'IA sia prodotta e introdotta da terzi. Diventa essenziale entrate in stretto rapporto con i fornitori di sistemi sin dalla fase di ideazione, affinché essi possano capire i potenziali rischi e possano mitigarli. Inoltre, sarà critico capire come i sistemi verranno aggiornati nel tempo e controllare come il modello cambierà di conseguenza. Grazie a questo approccio, si riducono i ritardi di implementazione, grazie all'inserimento di controlli già nelle prime fasi, e ci sarà un ulteriore velocità di pre-implementazione, dato che la maggior parte dei rischi saranno già stati identificati e mitigati.

McKinsey ha individuato otto distinte dimensioni del *model governance*, per i quali, in ciascuna fase, è possibile identificare ed incorporare diversi controlli per ciascuna fonte di rischio (nella lista sono stati inseriti a titolo di esempio alcuni possibili controlli):

A. Ideazione della soluzione;

controlli: *analisi di scopo, metriche di valutazione, valutazione ambientale con i dati disponibili.*

B. Ottenimento di dati affidabili per costruire ed allenare il modello;

controlli: *data-pipeline testing, analisi delle fonti, controlli e score statistici, fairness dei processi e dell'utilizzo dei dati, generazione automatica di documenti.*

C. Costruire un modello che raggiunga una buona performance nel risolvere il problema identificare nella fase di ideazione;

controlli: *model-robustness review, test sui KPI aziendali, controlli antidata-leakage, label-quality assessment, disponibilità dei dati nella produzione*.

D. Valutare le performance del modello, coinvolgendo le misure di business;

controlli: *standardized performance testing, feature-set review, rule-based threshold setting, verifica degli output del modello da esperti in materia, istituire requisiti o restrizioni in base al business o ai KPI, risk assesment, predictive-outcome fairness.*

Burkhardt, A. Govindarajan, & T. Wallace, A cura di)

E. Spostare il modello nell'ambiente di lavoro;

controlli: *non-functional-requirements check-list, validazione della fonte dei dati, full data-pipeline test, operational-performance thresholds, inserimento di messaggi di warning su interfacce esterne.*

F. Utilizzo del modello da parte delle unità di business coinvolte;

controlli: *responsabilizzazione e formazione dei colleghi, escalation mechanisms, workflow management, audit-trail generation.*

G. Inventory management di tutti i modelli;

controlli: *meccanismi di ricerca, automated inventory statistical assessment and risk overview by department.*

H. Monitoraggio costante.

controlli: *degradation flagging, schedulazione dei training, test periodici come il test di ipotesi bayesiano, automated logging, and audit-trail generation, continua verifica che l'algoritmo lavori sempre come è stato inteso e sia appropriato all'ambiente corrente.*

Risk management by design embeds controls across the algorithmic model's life cycle.

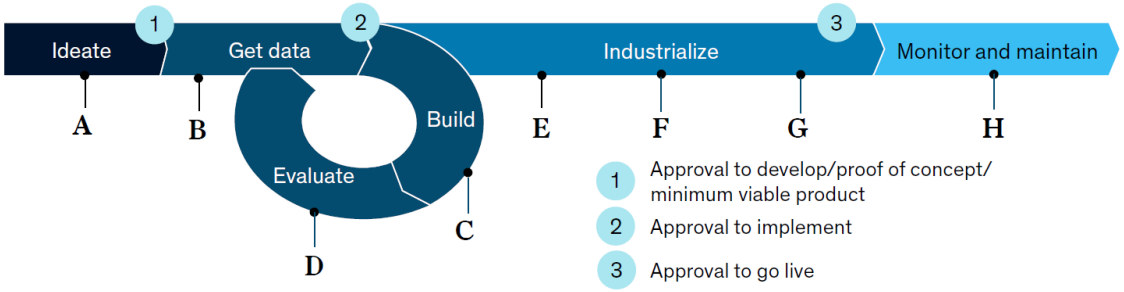

Figura 6 - Controllo dei rischi lungo il ciclo di vita dell'algoritmo

Affinché il processo di gestione del rischio sia applicato efficacemente, non è sufficiente che il risk manager lavori anch'esso *agile*: infatti, anche il resto dell'organizzazione deve essere coinvolta nella cultura del rischio, e coordinarsi costantemente con la funzione di risk management. Quest'ultima deve: sviluppare tutti gli elementi del model risk management e definire gli standard sulla costruzione e il rischio connessi ai sistemi di IA o ML; provvedere linee guida e monitorare il rischio di compliance, ad esempio, stabilendo quali dati dei clienti non possono essere utilizzati dai modelli; provvedere linee guida e mitigare i principali rischi non finanziari, quali quello reputazionale e terzi parti legati ai modelli di IA/ML. Le altre funzioni, anche quelle più tecniche, devono avvedersi degli impatti anche economici e organizzativi, pur

non essendo prettamente in linea con il bagaglio di competenze strettamente richieste dal ruolo. In particolare:

- I data scientists e gli sviluppatori possono introdurre controlli e contribuire alla trasparenza dei modelli, anche utilizzando tool di risk management (ad esempio, *explainability*, *bias testing*, dashboard di monitoraggio delle performance);
- I data engineers e strategist devono assicurarsi la qualità dei dati e l'applicabilità di nuove *features* (ad esempio, attraverso il c.d. *feature engineering*[20]).
- Il team IA in toto devono chiedersi preventivamente quali rischi possono sorgere durante il ciclo di vita del modello e avere chiaro quale sia il loro ruolo nel controllo del rischio e sapere come interagire con il resto dell'organizzazione attraverso protocolli prestabiliti;
- Il resto del team IT deve invece assicurarsi la cyber-sicurezza e la qualità dei sistemi (hardware e software) nei quali sono implementati tutti gli ingredienti degli algoritmi.
- Lato business, gli utilizzatori aziendali devono verificare la verosimiglianza degli output predittivi in base all'esperienza vissuta e assicurarsi che il modello sia stato utilizzato per il business-case per il quale è stato progettato.

Per permettere questa integrazione e coordinamento tra il team IA e il risk management attraverso il ciclo di vita del modello, si richiede una piattaforma condivisa, ispirata all'inventario dei modelli del MRM, che includa i seguenti elementi:

- Una documentazione standard concordata che soddisfi i bisogni e sia chiara a tutti gli stakeholders (sviluppatori, risk manager, business, compliance);
- Un unico tool di processo (*workflow*) per coordinare e documentare l'intero ciclo di vita, dall'ideazione, agli stadi iterativi di sviluppo, alla produzione e al rilascio, fino al ritiro del modello;
- L'accesso agli stessi dati, lo stesso ambiente di sviluppo, stessa tecnologia e lo storico dei test e revisioni;
- Una dashboard che con alta frequenza (se non in tempo reale) monitori i KPI critici, sin dal momento della produzione del modello;

20 Con feature engineering si intende quella pratica con la quale "a ritroso" i data scientist possono tracciare i features che l'algoritmo di intelligenza artificiale ha identificato per elaborare il proprio output: una "feature" è una proprietà o attributo condiviso da unità indipendenti di dati, per i quali viene applicato un algoritmo di IA/ML.

- Un set di strumenti a supporto dell'*explainability*, utili ad interpretare il comportamento di tutte le tecnologie IA, specialmente per quelle che appaiano meno trasparenti.

È evidente che quanto richiesto vada a stravolgere la classica organizzazione aziendale, data l'integrazione di compiti e conoscenze richieste sia da una parte che dall'altra. Ed è proprio l'integrazione la vera sfida per l'impresa. Laddove ci sia o meno una funzione di rischio esistente, la direzione può prendere alcune iniziative base per mettere in pratica il "derisking by design":

- Articolare i valori etici e la visione dell'impresa. Gli executive dovrebbero immaginare da una prospettiva top-down come l'azienda utilizzerà i dati, gli analytics e l'IA; ciò comprende un chiaro inquadramento di come i modelli debbano essere impiegati, riconoscere i rischi associati, e stabilire linee guida e confini che formano le basi del risk-management (come, ad esempio, accentrare il coordinamento del rischio e decentralizzare invece lo sviluppo degli algoritmi);
- Creare il processo di sviluppo dei modelli, come precedentemente indicato, dall'ideazione al monitoraggio, stabilendo tutti i controlli intermedi e i principi di produzione;
- Stabilire la governance e i ruoli chiave; identificare le persone responsabili dei rischi sia dal lato team analytics che del team di risk management, oltre che le responsabilità del management stesso;
- Adottare un modello agile che porti a lavorare affiancati sia i risk managers che il team di sviluppo, affinché capiscano le rispettive responsabilità e *modus operandi*. Così si consente di risolvere i conflitti da disallineamento dovuto al framework nel modo più efficiente possibile, e rendere il processo di sviluppo sempre più fluido;
- Rendere disponibili strumenti a garanzia della trasparenza e dell'interpretabilità dei sistemi;
- Formare e diffondere continuamente la conoscenza dell'IA e dei rischi connessi trasversalmente a tutta l'organizzazione.

CAPITOLO 3
I PRIMI PASSI DELLA REGOLAMENTAZIONE EUROPEA: L'ARTIFICIAL INTELLIGENCE ACT

3.1 – I razionali dell'Unione Europea per l'Artificial Intelligence Act

A partire dal lancio della Strategia Europea per l'Intelligenza Artificiale nell'aprile 2018. la politica della Commissione è stata quella di rendere l'Unione Europea un hub di livello mondiale per l'intelligenza artificiale, garantendo allo stesso tempo che l'IA sia incentrata sull'uomo e affidabile. La Commissione ha pianificato l'investimento di un miliardo di euro all'anno nell'IA, mobilitando ulteriori investimenti da parte del settore privato e degli Stati membri per raggiungere 20 miliardi di EUR di investimenti all'anno nel corso di questo decennio. In particolare, la promozione dell'innovazione basata sull'IA è strettamente legata all'attuazione della strategia europea per i dati, compresa la recente proposta di legge sulla "Data Governance Act", poiché l'IA può prosperare solo quando vi è un accesso agevole a tale risorsa. Soprattutto le piccole e medie imprese avranno bisogno di un accesso equo ai dati per rendere possibile un'ampia diffusione dell'IA nell'economia dell'UE. Questo è quanto afferma la Commissione Europa nel suo documento accompagnatorio "*Fostering a European approach to Artificial Intelligence*", evidenziando nuovamente i vantaggi nell'adozione di tale tecnologia: ottimizzare i processi industriali, rendendoli più resilienti, efficienti e più ecologici e a consentire soluzioni innovative di autoapprendimento e in tempo reale, dalla manutenzione predittiva ai robot collaborativi, dai *digital twins* alla realtà aumentata. Si prevede che nuove opportunità commerciali e un maggiore dinamismo economico creeranno nuove opportunità di lavoro e supereranno le potenziali perdite di posti di lavoro.

Il 21 aprile 2021, dopo tre anni di ricerca e *white papers*, l'Unione Europea ha pubblicato una proposta per il primo quadro giuridico per l'Intelligenza Artificiale (AI): The Artificial Intelligence Act. Tale disciplina propone norme al fine di migliorare la trasparenza e ridurre al minimo i rischi per la sicurezza e i diritti fondamentali prima che i sistemi di IA possano essere utilizzati nell'Unione Europea. La sua architettura si basa su una serie di componenti fondamentali che, nel loro complesso, costruiscono un *approccio normativo europeo proporzionato e basato sul rischio*.

In primo luogo, fornisce una definizione tecnologicamente neutra dei sistemi di IA che sia a prova di futuro, nella misura in cui può coprire tecniche e approcci che non sono

ancora noti o sviluppati. All'articolo 3, "Definizioni", il testo definisce "sistema di intelligenza artificiale" come qualunque software che sia sviluppato con una o più delle tecniche o approcci definiti in allegato I[21] e in grado, per un determinato insieme di obiettivi definiti dall'uomo, di generare risultati quali contenuti, previsioni, raccomandazioni o decisioni che influenzano gli ambienti con cui interagiscono.

In secondo luogo, per evitare un eccesso di regolamentazione, la proposta si concentra sui cosiddetti casi d'uso "ad alto rischio", in cui i rischi che i sistemi di IA pongono sono particolarmente elevati. Il fatto che un sistema di IA sia classificato come ad alto rischio dipende dalla sua destinazione d'uso del sistema e dalla gravità del possibile danno e dalla probabilità che si verifichi. I sistemi ad alto rischio includono, ad esempio, l'affidabilità creditizia dell'IA o sistemi destinati ad essere utilizzati per reclutare persone o valutare le loro per il processo decisionale giudiziario.

In terzo luogo, la proposta prevede che i sistemi di IA ad alto rischio debbano rispettare una serie di requisiti specificamente progettati, tra cui l'uso di serie di dati di alta qualità, la creazione di una documentazione appropriata per migliorare la tracciabilità, la condivisione di informazioni adeguate con l'utente, la progettazione e l'attuazione di adeguate misure di controllo umano e il raggiungimento degli standard più elevati in termini di robustezza, sicurezza, sicurezza informatica e precisione. I sistemi di IA ad alto rischio devono essere valutati per verificarne la conformità a tali requisiti prima di essere immessi sul mercato o messi in servizio. Per integrarsi agevolmente con i quadri giuridici esistenti, la proposta tiene conto, se del caso, delle norme settoriali in materia di sicurezza, garantendo la coerenza tra gli atti giuridici e la semplificazione per gli operatori economici.

Il progetto di regolamento proposto stabilisce il divieto di una serie limitata di usi dell'IA che violano i valori dell'Unione europea o violano i diritti fondamentali (c.d. *rischi inaccettabili*). Il divieto riguarda i sistemi di IA che distorcono il comportamento di una persona attraverso tecniche subliminali o sfruttando vulnerabilità specifiche in modi che causano o possono causare danni fisici o psicologici. Tale divieto si estende anche per l'attribuzione di uno *score* sociale generico dai sistemi di IA da parte delle

21 Machine learning, sia esso di tipo supervised, supervised e reinforcement learning, o deep learning; approcci logici e basati sulla conoscenza, tra cui la knowledge representation, la programmazione induttiva (logica), le basi di conoscenza, i motori di inferenza e deduttivi, il ragionamento (simbolico) e i sistemi esperti; approcci statistici, stime bayesiane e metodi di ricerca e di ottimizzazione. All'articolo 4, il legislatore si riserva la facoltà di aggiornare la presente lista, seguendo gli sviluppi del mercato e della tecnologia.

autorità pubbliche. Per il caso specifico dei sistemi di identificazione biometrica a distanza (ad esempio strumenti di riconoscimento facciale per controllare i passanti negli spazi pubblici), il regolamento proposto stabilisce un approccio più rigoroso. L'uso in tempo reale a fini di contrasto sarebbe in linea di principio vietato in spazi accessibili al pubblico, a meno che non sia eccezionalmente autorizzato dalla legge.

Secondo il regolamento proposto, altri usi dei sistemi di intelligenza artificiale (c.d. a *basso rischio*) sono soggetti solo a requisiti minimi di trasparenza, ad esempio nel caso di chatbot, sistemi di riconoscimento delle emozioni o deep fake. Ciò consentirà alle persone di fare scelte informate o di ritirarsi da una determinata situazione (Commissione Europea, 2021).

3.2 – L'approccio risk based: rischio inaccettabile, elevato, moderato

L'Artificial Intelligence Act, al titolo secondo, esordisce definendo le pratiche proibite di intelligenza artificiale, dando quindi una risposta sul piano etico al potenziale uso di tali sistemi. Viene data una risposta chiara: la società non vuole di sopportare determinati rischi, definiti inaccettabili, e per i quali qualsiasi forma di mitigazione viene considerata non sufficiente a fronte del potenziale impatto che le pratiche identificate portino. Nello specifico, la normativa "proibisce l'immissione sul mercato, la messa in servizio o l'uso di un sistema di IA che:

- impieghi tecniche subliminali al di là della coscienza di una persona al fine di distorcerne materialmente il comportamento in un modo che causi o è probabile che causi a quella persona o ad un'altra persona danni fisici o psicologici; oppure, sistemi di IA che sfruttino le vulnerabilità di un gruppo specifico di persone a causa della loro età, disabilità fisica o mentale, al fine di distorcere materialmente il comportamento di una persona appartenente a tale gruppo in modo da causare o possa causare a tale persona o a un'altra persona danni fisici o psicologici;
- valuti o classifichi dell'affidabilità delle persone fisiche in un determinato periodo di tempo in base al loro comportamento sociale o alle caratteristiche personali o di personalità note o previste, con il punteggio sociale che porta a uno o entrambi i seguenti elementi: (1) trattamento pregiudizievole o sfavorevole di determinate persone fisiche o di interi gruppi di persone in contesti sociali estranei ai contesti in cui i dati sono stati originariamente

generati o raccolti; (2) trattamento pregiudizievole o sfavorevole di determinate persone fisiche o di interi gruppi di persone che sia ingiustificato o sproporzionato rispetto al loro comportamento sociale o alla loro gravità;

- esegua l'identificazione biometrica remota in tempo reale in spazi accessibili al pubblico a fini di attività di contrasto, salvo casi il cui uso sia strettamente necessario, quali la ricerca di vittime specifiche di reato, prevenzione di una minaccia specifica e imminente per la vita o l'incolumità fisica delle persone, o il rilevamente e identificazione nei confronti di un autore o sospettato di una tipologia strettamente difinita di reati (Commissione Europea, 2021);

L'articolo 6 dell'atto normativo definisce le regole di classificazione per i sistemi di IA ad alto rischio, stabilendo che, "*un sistema di IA è considerato ad alto rischio se sono soddisfatte entrambe le condizioni seguenti:*

a) il sistema di IA è destinato a essere utilizzato come componente di sicurezza di un prodotto, o è esso stesso un prodotto, disciplinato dalla normativa di armonizzazione dell'Unione elencata nell'allegato II[22]*;*

b) il prodotto, il cui componente di sicurezza è il sistema di IA, o il sistema di IA stesso in quanto prodotto è soggetto a una valutazione della conformità da parte di terzi ai fini dell'immissione sul mercato o della messa in servizio di tale prodotto ai sensi della normativa di armonizzazione dell'Unione elencata nell'allegato II."

Ancora, sono considerati sistemi IA ad altro rischio anche i sistemi elencati nell'allegato III della normativa. Tale allegato comprende sistemi compresi in ciascuna delle seguenti aree:

- Identificazione biometrica e categorizzazione delle persone fisiche;
- Gestione e funzionamento delle infrastrutture critiche (ad esempio, gestione del traffico, forniture di acqua, gas ed elettricità);
- Istruzione e formazione professionale (ad esempio, nella valutazione personale al fine dell'ammissione in ambiti educativi e professionali);
- Occupazione, gestione dei lavoratori e accesso al lavoro autonomo (ad esempio, algoritmi impiegati nella selezione del personale, o determinanti la promozione o terminazione del rapporto di lavoro;
- Accesso e fruizione dei servizi privati essenziali e dei servizi e benefici pubblici

22 trattasi di una lista di normative armonizzate dell'Unione Europea

(come, ad esempio, l'accesso al credito);

- Applicazione della legge;
- Gestione della migrazione, dell'asilo e del controllo delle frontiere;
- Amministrazione della giustizia e dei processi democratici (Commissione Europea, 2021).

La Commissione Europea si riserva il potere di "aggiornare l'elenco di cui all'allegato III aggiungendo sistemi di IA ad alto rischio".

Al Titolo III, Capo 2, la proposta di normativa delinea i requisiti a cui devono sottostare i sistemi di intelligenza artificiale ad alto rischio, evidenziando in primis che "nel garantire conformità a tali requisiti si tiene conto della finalità prevista del sistema di IA ad alto rischio e del sistema di gestione dei rischi". L'articolo 9 continua poi affermando che tale framework di risk management debba essere costituito da "un processo iterativo continuo eseguito nel corso dell'intero ciclo di vita di un sistema di IA", comprendente le fasi classiche della gestione del rischio, quali l'identificazione e l'analisi dei rischi noti e prevedibili associati a ciascun sistema ad altro rischio, la stima e valutazione dei rischi che possono emergere sia in caso di utilizzo conforme alle finalità che uso improprio ragionevolmente prevedibile, la valutazione di altri eventuali rischi derivanti dall'analisi dei dati raccolti dal sistema di monitoraggio successivo all'immissione sul mercato, e l'adozione di adeguate misure di gestionr dei rischi. Ulteriori requisiti sono poi meglio definiti negli articoli successivi del Titolo III Capo 2, rigurdanti i dati e la loro governance, la documentazione tecnica, la conservazione delle registrazioni (il *log* degli eventi), la trasparenza e fornitura di informazioni agli utenti, la sorveglianza umana e i requisiti di accuratezza, robustezza e cibersicurezza. Il tutto declinato anche nell'ottica del fornitore di sistemi AI (Titolo III, Capo 3). Gli obblighi di sorveglianza imposti dal progetto di regolamento a coloro che costruiscono e vendono sistemi ad alto rischio sul mercato o li utilizzano comprendono:

a) "*Conformity Assessments*", che sono valutazioni d'impatto algoritmiche che analizzano set di dati, bias, il modo in cui gli utenti interagiscono con il sistema e la progettazione e il monitoraggio complessivi dei risultati del sistema;
b) Garantire che questi sistemi siano spiegabili, supervisionabili e funzionino in modo coerente per tutta la loro durata, anche nei casi limite;
c) Istituzione di una pratica di gestione del rischio informatico a livello di organizzazione che includa rischi specifici dell'IA, come gli attacchi informatici

ai sistemi IA.

I sistemi considerati come a rischio minimo avrebbero, invece, un numero significativamente inferiore di requisiti, principalmente sotto forma di obblighi specifici di trasparenza, come rendere gli utenti consapevoli del fatto che stanno interagendo con una macchina in modo che possano prendere una decisione informata sulla continuazione, chiarire se un sistema utilizza il riconoscimento delle emozioni o la classificazione biometrica e notificare agli utenti se l'immagine, i contenuti audio o video sono stati generati o manipolati dall'IA per rappresentarne falsamente il contenuto, come un video generato dall'IA che mostra un personaggio pubblico che fa una dichiarazione che non è mai stata, di fatto, fatta (c.d. *deep fake*). L'obbligo di creare tale consapevolezza si applicherà ai sistemi di tutte le categorie di rischio.

Un punto di attenzione è la portata extra-territoriale del Regolamento: oggetto della normativa sono tutti i sistemi di IA che sono presenti/stanno per entrare nel mercato o sono utilizzati in Europa: ciò implica che qualsiasi sistema di IA che fornisca un output all'interno dell'Unione Europea sarebbe soggetto ad esso, indipendentemente da dove si trova il fornitore o l'utente, andando quindi a influenzare i grandi software provider extra-europei, soprattutto quelli statunitensi e asiatici.

3.3 – Le implicazioni del Regolamento per le imprese

Sebbene non ci sia ancora data certa riguardo l'entrata in vigore effettiva della normativa, visto anche il lungo percorso che ha impiegato la GDPR, è chiaro che fin da ora le imprese debbano cominciare a adeguarsi al futuro standard tracciato dall'AI Act.

Con queste basi in atto, le organizzazioni possono intraprendere tre azioni per iniziare a costruire sistematicamente il loro programma completo di gestione del rischio legato all'IA: creare un inventario dei sistemi di intelligenza artificiale e delle misure di mitigazione del rischio basate su una tassonomia standard; condurre le valutazioni di conformità e stabilire un sistema di governance dell'IA.

Le organizzazioni dovrebbero creare e mantenere inventari completi contenenti descrizioni di tutti i sistemi di intelligenza artificiale associati a casi d'uso attuali e pianificati, insieme a classificazioni dei rischi per ciascun sistema. Possono quindi utilizzare questi inventari per mappare i loro sistemi di intelligenza artificiale rispetto alle normative esistenti e potenziali future per identificare e affrontare eventuali lacune nella conformità.

Secondo la proposta di Regolamento UE, le organizzazioni sarebbero tenute a condurre una valutazione della conformità per tutti i sistemi di IA ad alto rischio. Analogamente alle valutazioni d'impatto sulla privacy attualmente richieste in varie regioni, la valutazione della conformità è una revisione di ciascun sistema di IA per vedere se soddisfa le normative applicabili e altri standard pertinenti. Tali valutazioni dovrebbero includere tutte le informazioni richieste ai sensi della proposta di regolamento dell'UE, come le seguenti: (1) documentazione riassumente le varie scelte ed assunzioni nella costruzione dell'algoritmo, incluse le limitazioni e il livello di accuratezza; (2) i rischi insisti nel sistema, comprese le fonti prevedibili di conseguenze indesiderate, come potenziali discriminazioni o violazioni dei diritti fondamentali; (3) eventuali misure di attenuazione dei rischi integrate nel sistema o ad esso applicate, come la sorveglianza umana.

Due componenti chiave di un sistema di governance di successo sono un comitato inter-funzionale dedicato responsabile di garantire la conformità al rischio di IA e audit indipendenti dei sistemi di IA. Il comitato dovrebbe essere composto da professionisti provenienti da una varietà di funzioni, tra cui la cibersicurezza, il diritto e la tecnologia, per affrontare adeguatamente l'intera gamma di rischi dell'IA. Questo organo di governance stabilisce gli standard di rischio a cui i team di intelligenza artificiale devono aderire, garantisce e controlla i sistemi di intelligenza artificiale e i processi di sviluppo per la conformità e consiglia i team aziendali e di sviluppo su specifici compromessi o decisioni necessarie per conformarsi agli standard normativi e organizzativi. Diverse organizzazioni, come l'*International Organization for Standardization* (ISO) e il *National Institute of Standards and Technology* (NIST) degli Stati Uniti, stanno già pubblicando standard di sviluppo e implementazione dell'IA responsabili e spingendo per l'adesione internazionale o nazionale; I comitati di governance dell'IA possono utilizzarli come risorse utili per la definizione di standard organizzativi e benchmarking. (McKinsey & Company, 2021)

CONCLUSIONI

L'obiettivo del presente elaborato è di guardare all'utilizzo dell'intelligenza artificiale evidenziandone in particolare i rischi connessi al suo utilizzo. Mentre i vantaggi economici ed organizzativi sono chiari alle imprese, non sono altrettanto chiare le conseguenze che lo sviluppo e l'utilizzo di un sistema di IA porta con sé. Le grandi aziende tech, pionieri nell'adozione e sviluppo di IA, hanno avuto più volte occasione di fallire, anche clamorosamente, nell'implementare sistemi che si sono rivelati poi fallosi, e ripercuotendosi poi a livello di danni economici e reputazionali. Non tutte le imprese, soprattutto quelle medio/piccole, non possono permettersi di sbagliare, e devono fare tesoro dei fatti finora accaduti e degli studi tecnici ancora in corso. Partendo dagli studi e dagli eventi, si è potuto tracciare un quadro generale il quanto più possibile ampio per raggruppare tutti i potenziali rischi dell'intelligenza artificiale. Non solo quelli endogeni, di carattere tecnico-implementativo, che sono stati in questo elaborato toccati ad un alto livello, ma anche quelli esogeni, legati alle caratteristiche dell'ambiente nel quale l'algoritmo risiede, nell'insieme delle proprie condizioni naturali e dalle regole che governano l'ambiente. A cavallo di queste due dimensioni, l'utilizzo dei dati, fonte essenziale dello sviluppo dell'IA, ma portatori di distorsioni di cui spesso neanche l'uomo è consapevole. Difatti, seppure i rischi connessi all'IA possano essere facilmente tracciati, non è conclusa la sfida in capo a noi umani nel percepire e definire il concetto di equità e rispondere dilemmi etici e morali, sui quali non possiamo permetterci di trascurare e semplificare come soliti fare nella definizione di modelli matematici. Emblematici sono i casi degli algoritmi razzisti o sessisti che attraverso i dati hanno interpolato la reale condizione sociale, facendo da cartina tornasole.

Dopo aver identificato le diverse fonti di rischio, abbiamo definito un framework utile per gestirle: a partire dal Model Risk Management, è stato definito un modello che permetta adeguatamente di tener traccia di tutte le informazioni riguardanti gli sviluppi e revisioni degli algoritmi e i rischi ad essi connessi, attraverso lo strumento dell'inventario. Grazie al *model inventory*, è possibile effettuare un'attività di *tiering*, grazie alla quale è possibile concentrare i controlli e le valutazioni in proporzione al rischio associato ad ogni sistema di IA. Rispetto al tradizionale *MRM*, basato su valutazioni periodiche e focalizzato sui tradizionali rischi finanziari, è stata ravvisata la necessità di integrare il più possibile il processo di identificazione e mitigazione del rischio all'interno del ciclo di vita dell'algoritmo: con "*derisking by design*" è stata definita la buona pratica per la quale la "cultura del rischio" debba essere estesa e

condivisa a tutte le parti coinvolte nello sviluppo di sistemi AI. Tale approccio, infatti, responsabilizza tutti gli stakeholders coinvolti, dagli sviluppatori alle funzioni di business più operative, nel prevedere e/o segnalare anomalie potenziali ed occorse in ogni momento: dalla definizione del primo *proof-of-concept* alla raccolta dei dati, fino alla sorveglianza del sistema in funzione, comprese tutte le revisioni intermedie. Particolare attenzione è da attribuire alle caratteristiche peculiari della tecnologia: in primis, la supervisione continua dell'algoritmo, dato che, col tempo, le condizioni ambientali per le quali l'algoritmo è stato sviluppato possono cambiare; poi, la capacità dell'algoritmo di saper spiegare come ha prodotto l'output e la propria decisione. Altri elementi contornano in modo non banale il framework descritto nell'elaborato: dalla revisione dell'algoritmo e delle assunzioni da parte di un team diverso rispetto a quello di sviluppo (sia esso interno o esterno all'azienda), all'istituzione di un comitato di governance che esprima pareri e definisca linee guida sul piano etico e funzionale dei modelli. Se tali considerazioni possono considerarsi dovute e adeguate a un'organizzazione di grande dimensione, non banale è l'adozione di tali presidi da parte di piccole e medie imprese: questa situazione si configura come una potenziale opportunità di mercato per glie enti che possiedano le competenze adeguate allo svolgimento di consulenze e certificazioni, alle quali le imprese non strutturate possano esternalizzare tali funzioni.

Infine, con particolare attenzione alla compliance, è stato analizzato il framework normativo a tendere che la Commissione Europea ha pubblicato nell'aprile 2021: l'approccio basato sul rischio adottato dall'Istituzione europea, ha delineato gli standard di adozione di sistemi di intelligenza artificiale considerati "ad alto rischio", ponendo dei vincoli di requisito che ciascun sistema in oggetto dovrà rispettare, quali l'obbligo di trasparenza, la valutazione di conformità, l'adeguata informativa all'utente e i requisiti di accuratezza, robustezza e cibersicurezza. Tali elementi andranno quindi a riassumersi nel framework di MRM "adeguato", costituendo delle nuove *milestone* che dovranno essere raggiunte nello sviluppo e mantenimento dei sistemi di IA.

BIBLIOGRAFIA

AIFIRM. (2021, Marzo). *Position Paper - AIFIRM.* Tratto da AIFIRM: https://www.aifirm.it/wp-content/uploads/2016/03/2021-Position-Paper-28-Model-Risk-Management.pdf

AIRS - Artificial Intelligence/Machine Learning Risk and Security. (2021, Dicembre 31). *Wharton University of Pennsylvania: Artificial Intelligence Risk & Governance*. Tratto da Wharton University of Pennsylvania: https://ai.wharton.upenn.edu/artificial-intelligence-risk-governance/#_ftn1

Baer, T., & Kamalnath, V. (2017, Novembre 10). *Controlling Machine-Learning Algorithms and their Biases.* Tratto da McKinsey&Company: https://www.mckinsey.com/business-functions/risk-and-resilience/our-insights/controlling-machine-learning-algorithms-and-their-biases

Blattner, L., & Nelson, S. (2021, May 5). *How Costly is Noise? Data and Disparities in Consumer Credit.* Tratto da Cornell University: https://arxiv.org/abs/2105.07554

CNBC. (2020, May 27). *As U.S. coronavirus deaths cross 100,000, black Americans bear disproportionate share of fatalities.* Tratto da CNBC: https://www.cnbc.com/2020/05/27/as-us-coronavirus-deaths-cross-100000-black-americans-bear-disproportionate-share-of-fatalities.html

Commissione Europea. (2021, Aprile 21). *Agenda Digitale.* Tratto da Agenda Digitale: https://imgcdn.agendadigitale.eu/wp-content/uploads/2021/05/03154811/regulation_annex_ai_875FDD6D-CC6A-E50A-8E48824677EFED42_75789.pdf

Commissione Europea. (2021, Aprile 21). *Communication on Fostering a European approach to Artificial Intelligence.* Tratto da Commissione Europea: Scaricabile dalla seguente pagina: https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence

Commissione Europea. (2021, Aprile 21). *EUR-Lex.* Tratto da EUR-Lex: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206&qid=1644678082070

Deloitte. (2018, Gennaio 01). *AI and Risk Management - Innovating with Confidence.* (T. Bigham, V. Gallo, S. Nair, M. Lee, S. Soral, T. Mews, . . . M. Fouché, A cura di) Tratto da Deloitte.com: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/financial-services/deloitte-uk-ai-and-risk-management.pdf

Duranton, S. (2020, Febbraio 14). *How humans and AI can work together to create better businesses*. Tratto da Youtube: https://www.youtube.com/watch?v=2KMk1lJGPlk

Forbes. (2012, February 16). *How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did.* Tratto da Forbes: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=6b9a3e866686

Gao, A., Santinelli, M., & Singer, S. (2020, Agosto 12). *Getting to the Root of Data Bias in AI.* Tratto da BCG GAMMA: https://medium.com/bcggamma/getting-to-the-root-of-data-bias-in-ai-a8179a54f45e

Ledford, H. (2019, October 24). *Millions of black people affected by racial bias in health-care algorithms.* Tratto da Nature: https://www.nature.com/articles/d41586-019-03228-6/

Lisowski, E. (2022, February 15). *Artificial intelligence in real estate: Use cases.* Tratto da addepto: https://addepto.com/ai-in-real-estate-use-cases/

Longo, A., & Mischitelli, L. (2020, Novembre 11). *Vaccini anti-covid: così l'intelligenza artificiale ha accellerato la ricerca.* Tratto da Agenda Digitale.eu: https://www.agendadigitale.eu/sanita/vaccini-anti-covid-cosi-lintelligenza-artificiale-sta-aiutando-la-ricerca-e-le-terapie/

Mantovani, R. (2020, Febbraio 14). *Salute, Il primo farmaco sviluppato dall'intelligenza artificiale.* Tratto da Focus: https://www.focus.it/scienza/salute/il-primo-farmaco-sviluppato-da-intelligenza-artificiale

McCarthy, J. (2004). *What Is Artificial Intelligence?* Standford University, Computer Science Deparment, Standford, CA. Tratto da http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

McCarty, J., Minsky, M., Rochester, N., & Shannon, C. (1955). *A Proposal For The Dartmouth Summer Research Project On Artificial Intelligence.*

McKinsey & Company. (2021, August 10). *What the draft European Union AI regulations mean for business.* Tratto da McKinsey & Company: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/what-the-draft-european-union-ai-regulations-mean-for-business

McKinsey. (2020, Agosto 13). *Derisking AI by design: How to build risk management into AI development.* (J. A. Baquero, R. Burkhardt, A. Govindarajan, & T. Wallace, A cura di) Tratto da McKinsey: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/derisking-ai-by-design-how-to-build-risk-management-into-ai-development

McKinsey. (2020, November 17). *The State of AI in 2020.* (T. Balakrishnan, M. Chui, B. Hall, & N. Henke, A cura di) Tratto da McKinsey & Company: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020

McKinsey Global Institute. (2019, Giugno 1). *Notes from the AI frontier: Tackling bias in AI (and in humans).* (J. Silberg, & J. Manyika, A cura di) Tratto da Notes from the AI frontier: Tackling bias in AI (and in humans): https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/mgi-tackling-bias-in-ai-june-2019.pdf

Miller, A. P. (2018, Luglio 26). *Harvard Business Review.* Tratto da Harvard Business Review: https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

NIST. (2021, October 15). *Draft - Taxonomy of AI Risk.* Tratto da National Institute of Standard and Technology: https://www.nist.gov/system/files/documents/2021/10/15/taxonomy_AI_risks.pdf

OCSE (Organizzazione per la Cooperazione e lo Sviluppo Economico). (2021, Dicembre 31). *OECD AI Principles overview*. Tratto da oecd.ai: https://oecd.ai/en/ai-principles

PWC. (2017). *Artificial Intelligence Study - Sizing the Prize.* (A. S. Rao, & G. Verweij, A cura di) Tratto da PWC: https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html

PwC. (2020). *Model Risk Management of AI and Machine Learning Systems.* Tratto da PwC: https://www.pwc.co.uk/data-analytics/documents/model-risk-management-of-ai-machine-learning-systems.pdf

Quintarelli, S., Corea, F., Ferrauto, C. G., Fossa, F., Loreggia, A., & Sapienza, S. (2020). *Intelligenza Artificiale - Cos'è davvero, come funziona, che effetti avrà.*

(S. Quintarelli, A cura di) Torino: Bollati Boringhieri editore.
Reuters. (2018, October 11). *Amazon scraps secret AI recruiting tool that showed bias against women.* Tratto da Reuters: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
Rogers, E. M. (1962). *Diffusion of innovations.*
The Guardian. (2021, November 4). *The $300m flip flop: how real-estate site Zillow's side hustle went badly wrong.* Tratto da The Guardian: https://www.theguardian.com/business/2021/nov/04/zillow-homes-buying-selling-flip-flop?utm_source=podia&utm_medium=broadcast&utm_campaign=822477
Xiang, M. (2019, Marzo 18). *Human Bias in Machine Learning*. Tratto da Towards Data Science: https://towardsdatascience.com/bias-what-it-means-in-the-big-data-world-6e64893e92a1

Ringraziamenti

Un sincero ringraziamento al mio relatore, Andrea Giacomelli, per la libertà e gli spunti offertomi per la stesura del presente elaborato, venendo incontro ai miei interessi e fornendomi una direttrice per lo sviluppo della tesi.

Ringrazio l'Università Ca' Foscari, dai docenti che mi hanno accompagnato nella mia crescita personale, arricchendomi della loro conoscenza ed esperienza, a tutti i miei colleghi e amici che hanno contribuito alla mia maturazione: un particolare ringraziamento all'associazionismo studentesco e a JEVE Ca' Foscari, che completano il mondo universitario arricchendolo delle esperienze che la sola istruzione non può fornire.

Ringrazio i miei genitori, Adelina e Dante, che sempre mi hanno supportato negli studi e nelle mie passioni, che senza il loro contributo, affetto e guida non avrei mai potuto raggiungere così come ho fatto. Non bastano le parole per esprimere tutta la mia stima per voi.

Ringrazio Chiara, con cui ho condiviso in toto il bello e il meno bello di questa meravigliosa esperienza, e con cui avrò la fortuna di vivere le mie prossime avventure.

Ringrazio tutte le meravigliose persone che mi hanno supportato e che hanno sempre creduto in me: da mia sorella, alla mia famiglia, ai miei amici, alla musica.