# L'illusione dell'"AI agentica": concetti base, limiti e rischi resi evidenti dal confronto con la teoria dell'agenzia

Riccardo Bovetti

December 26, 2025

## 1 Premessa operativa

E' oramai un pattern ricorrente: trovo una ricerca, una pubblicazione od un articolo che promette qualche nuovo punto di vista, qualche metrica aggiornata o quale approfondimento sul tema al quale oramai da due anni sto dedicando i miei sforzi di "ricerca" (ovverosia il macro mondo dell'AI ed i suoi impatti e risvolti, prevalentemente, in ambito aziendale ed educativo) ed, una volta letta, mi si accavallano pensieri ed immagini che mi obbligano a cercare di approfondire delle connessioni che, a volte, sono solo evocate. Nel corso dell'ultimo anni si è diffuso un entusiasmo (interesse?) crescente attorno al concetto di "AI agentica", ovvero sistemi di intelligenza artificiale capaci di agire con una qualche forma di autonomia perseguendo obiettivi più che eseguendo un processo avente questo fine. Il recentissimo report di Wharton sintetizza testualmente questo nuovo hype: "l'AI agentica è la nuova frontiera" e ormai *ogni discussione tra executive finisce per toccare il tema degli agenti AI.* Ma a fronte di questo interesse, mi sorge spontanea una domanda provocatoria: quanti di quelli che usano questo termine hanno riflettuto, davvero, su cosa significa "agente", su quali origini abbia il termine e quali trappole si nascondano dietro il suo utilizzo disinvolto? Voglio prendermi questo spazio per analizzare criticamente il concetto di AI agentica, discutendone i limiti teorici e pratici (chiedendomi, appunto, cosa significhi davvero "agente" in ambito IA almeno cosa significhi a Novembre 2025, visto che la storia recente insegna quanto facilmente e rapidamente i concetti evolvano), mettendolo a confronto con la tradizionale Agency Theory che si studia ed insegna in ambito economico-aziendale (Jensen & Meckling, Eisenhardt, Ross, ecc.), L'idea di fondo è quella di portare in luce i rischi della delega decisionale "on behalf", cioè per conto del lavoratore e dell'azienda, a un sistema AI, con implicazioni organizzative, informative e reputazionali anche prendendo in considerazione il ruolo delle norme emergenti, con un riferimento particolare all'AI Act europeo e alle istanze di governance etica dell'IA.

## 2 Partiamo dall'inizio: qualche nozione storica

I fattori della Compagnia Olandese delle Indie Orientali che operavano nella Batavia del XVII secolo avevano imparato molto presto a sfruttare i ritardi di comunicazione che potevano arrivare fino a 18 mesi con i direttori di Amsterdam, creando un'asimmetria informativa che rendeva il monitoraggio del loro operato effettivamente impossibile. In questo modo gli agenti si dedicavano al commercio privato, appropriandosi delle opportunità dell'azienda mentre i direttori non potevano verificare i rapporti su prezzi, qualità o condizioni di mercato. Il risultato: appesantita dal contrabbando, dalla corruzione e dai crescenti costi amministrativi (la compliance è tra noi da illo tempore, o meglio proprio da allora) la compagnia delle Indie fallì nel 1799 dopo soli due secoli. Quello che nello scorso secolo venne chiamato "problema di agenzia" distrusse letteralmente quella che era stata la più grande azienda del mondo. Similmente, i manager della Compagnia Inglese delle Indie Orientali esercitavano poteri quasi-governativi in India lontano dalla supervisione di Londra, accettando tangenti, manipolando gli acquisti di merci per profitto personale e usando l'esercito della compagnia per scopi privati sguazzando in orizzonti temporali divergenti che incoraggiavano l'estrazione (di valore, o meglio di beni di valore) a breve termine piuttosto che la creazione di valore a lungo termine.

## 3 Lente teorica: la Agency Theory tra principal e agent e la matematica senza tempo del disallineamento

Nella pratica economica (ed in particolare nell'ambito della ingegneria gestionale) la Teoria dell'Agenzia (Agency Theory) emerse dall'osservazione di un pattern ripetuto attraverso i secoli e relativo ai problemi che nascono quando un principale delega a un agente l'esecuzione di un compito o una decisione, in presenza di informazioni asimmetriche e interessi potenzialmente divergenti. Il fondamento teorico (nell'accezione moderna) del fenomeno si può far risalire agli studi Stephen Ross che nel 1973 definì formalmente la dinamica delle relazioni in cui "una parte agisce per, per conto di, o come rappresentante dell'altra in un particolare dominio di problemi decisionali." Ross formulò la teoria in forma di embrionale "sfida matematica" che ancora oggi rimane irrisolta: quando gli agenti massimizzano la propria utilità piuttosto che il benessere del principale, e i principali non possono osservare perfettamente le azioni degli agenti, una certa perdita residua è inevitabile anche con meccanismi ottimali di monitoraggio e bonding. Il lavoro successivo di Jensen & Meckling (1976) che definiva l'agenzia come "un contratto in cui una o più persone (il principale) incaricano un'altra persona (l'agente) di svolgere un servizio per loro conto, delegandole un certo potere decisionale" rese questo fondamento matematico maggiormente netto preciso. I costi di agenzia equivalgono alle spese di monitoraggio più le spese di bonding più la perdita residua,la riduzione irriducibile del benessere derivante dalla delega imperfetta. Joseph Stiglitz condivise il Premio Nobel 2001 per aver analizzato

i mercati con informazione asimmetrica, dimostrando che ogni volta che l'informazione è imperfetta e distribuita asimmetricamente (e questo è essenzialmente sempre) le economie non sono "Pareto efficienti". La sua teoria dello screening mostrò come le parti non informate progettano contratti per indurre le parti informate a rivelare informazioni private attraverso l'autoselezione. Questo divenne fondamentale per comprendere i mercati assicurativi, il razionamento del credito e i rapporti di lavoro. La sfida si intensifica quando gli agenti possiedono sia informazioni superiori sulle circostanze che la capacità di intraprendere azioni nascoste che influenzano i risultati. La Agency Theory ha formalizzato diverse categorie di problemi che sorgono in queste relazioni di delega. I principali, che tratto ampiamente nei miei corsi universitari di Management Control dove le componenti "comportamentali" sono per me importanti quanto quelle "tecniche", sono:

**Adverse Selection (Selezione avversa)** – un problema pre-contrattuale, che si verifica prima della chiusura dell'accordo tra principale e agente. È dovuto all'informazione asimmetrica ex ante: l'agente possiede informazioni private sulla propria qualità, capacità o intenzioni che il principale non conosce, e ciò può portare a una scelta sfavorevole dell'agente. In altre parole, il principale rischia di selezionare, senza saperlo, un agente meno adatto o meno onesto di quanto appaia, perché quest'ultimo ha nascosto o distorto informazioni a proprio vantaggio. In pratica chi "compra" o ingaggia qualcuno può trovarsi con la parte "scadente" del mercato se non riesce a distinguere ex ante la qualità . Nel contesto del lavoro, un esempio di selezione avversa è assumere un candidato che sembra perfetto ma che ha abilmente occultato lacune o cattive attitudini durante il processo di selezione (o millantato titoli, cosa che sarebbe sempre meglio non fare soprattutto se si mira a cariche istituzionali)]

**Moral Hazard (azzardo morale)** – un problema post-contrattuale, che insorge dopo l'inizio del rapporto di agenzia. Qui l'asimettria informativa riguarda le azioni dell'agente una volta contrattualizzato: l'agente potrebbe comportarsi in modo opportunistico o diverso da quanto atteso, sfruttando il fatto che il principale non può osservarlo o controllarlo direttamente. In pratica l'agente, avendo ora il compito delegato, potrebbe perseguire i propri interessi a scapito di quelli del principale, soprattutto se sa di non sopportarne direttamente le conseguenze. Classico nella teoria è l'esempio dell'assicurato (agente) che, dopo aver stipulato un'assicurazione, adotta condotte meno prudenti perché sa che l'assicuratore (principale) coprirà i danni (cosa alla quale il mercato assicurativo ha reagito con una burocratizzazione nelle liquidazioni che di fatto controbilancia i rischi). Nel mondo aziendale, moral hazard è il manager che prende decisioni eccessivamente rischiose perché il costo di eventuali perdite sarà sostenuto dagli azionisti, mentre lui potrebbe comunque trarne benefici (bonus, stock option) se le cose vanno bene. In sintesi, l'agente è tentato ad adottare "shirk behavior" (con un neologismo potremmo dire "fannullonare") o di agire in modo non diligente, sapendo che il principale fatica a monitorare ogni sua mossa.

**Incentivi disallineati (misalignment)** all'origine di molti conflitti di agenzia c'è la divergenza negli obiettivi o nei criteri di determinazione del successo tra principale e agente. Se il principale valuta la performance in base a certe metriche, l'agente cercherà di ottimizzare le proprie ricompense anche se ciò non coincide con il massimo beneficio per il principale. Questo può portare a comportamenti distorsivi. Ad esempio, se un venditore è pagato solo a provvigione sul venduto, potrebbe essere tentato di realizzare vendite a qualunque costo, perché il suo incentivo è chiudere contratti, mentre l'azienda avrebbe interesse a relazioni di lungo termine con clienti soddisfatti. Jensen e Murphy (1990) discussero ampiamente il tema dell'allineamento degli incentivi, sottolineando come sistemi di compenso mal disegnati possano peggiorare i problemi di agenzia invece di risolverli. L'agency cost (costo di agenzia) include proprio le perdite di efficienza dovute a decisioni non ottimali dell'agente per via di incentivi imperfetti, oltre ai costi di controllo.

**Costi di monitoraggio e controllo** per mitigare i problemi di cui sopra, il principale deve investire in meccanismi di monitoraggio delle azioni dell'agente (controlli, reportistica, audit) e in contratti che leghino la retribuzione dell'agente ai risultati desiderati. Tutto ciò genera costi. Ad esempio, gli azionisti devono nominare consigli di amministrazione, revisori, comitati di controllo interno, spendere in sistemi informativi per tracciare la gestione, ecc. L'agente, dal canto suo, potrebbe sostenere costi di bonding (di garanzia), cioè spese per segnalare il proprio impegno o rassicurare il principale (ad esempio, offrire una clausola di performance bond, o accettare una parte variabile di stipendio legata ai risultati). Anche rispettando queste misure, generalmente nell'equazione resta una perdita residua perché non si può azzerare del tutto il conflitto di interessi senza costi proibitivi

La storia del management e della filosofia morale ci ricordano che perseguire un fine e seguire un processo che renda legittimo quel fine non sono la stessa cosa perchè esiste una profonda differenza tra una razionalità del risultato e una razionalità della condotta. La prima risponde alla domanda "come massimizzo l'output?", la seconda a "quali azioni sono ammissibili per ottenerlo?" ed un'antica legge organizzativa recita: l'ottimizzazione cieca del fine produce distorsioni. In altre parole, un agente può avere razionalità strumentale senza avere razionalità procedurale. Un buon agente non è quindi solo quello che raggiunge lo scopo, ma quello che lo persegue rispettando il processo legittimo per farlo perchè è il processo che dà forma, trasparenza e responsabilità al risultato.

# 4 Cosa significa (davvero) “agente”, e più specificamente che significato assume nel contesto dell’IA? (O meglio, cosa ho capito io del tema)

Il termine “agentico” applicato all’IA è spesso usato in modo ambiguo e generico. Nel linguaggio comune dell’industria tech, un AI agent evoca l’idea di un programma autonomo che, dato un obiettivo, può pianificare e agire indipendentemente per raggiungerlo, eventualmente interagendo con l’ambiente e altri sistemi. Tuttavia, come nota anche Ethan Mollick, “agente” rimane un termine sfumato: c’è stato molto buzz sugli agenti negli ultimi mesi, ma ancora (troppa) poca tecnologia che funzioni o che funzioni davvero bene . In altre parole, il concetto è di moda, ma la sua concretizzazione pratica è ancora embrionale e mal definita. Bisogna quindi chiarire i fondamenti del termine prima ancora di provare ad applicarlo all’AI per evitare di farci tirare in mezzo in modo acritico dall’HiperHype - d’altro canto si sa, quando le definizioni divergono, vince il marketing.

Il termine ”agente” porta significati radicalmente diversi attraverso le discipline alle quali si applica, creando confusione concettuale che il marketing tende a sfruttare sistematicamente. In economia, (o più ampiamente nel “management” come dalle definizioni di cui alla sezione precedente) gli agenti sono massimizzatori di utilità che operano sotto vincoli. Il mia abbastanza citato lavoro di Herbert Simon del 1995 sulla razionalità limitata (per me, ma non solo per me ovviamente, il testo seminale della AI, quanto meno dal punto di vista sociale) riconobbe che gli agenti reali ”soddisfano” piuttosto che ottimizzare a causa di limiti cognitivi e costi informativi, ma la razionalità rimane puramente procedurale ed esterna. Gli agenti economici non necessitano di essere coscienti o autonomi, semplicemente ottimizzano entro vincoli. In filosofia, l’agenzia richiede intenzionalità, come specificato con il termine”’aboutness” degli stati mentali di Brentano, coscienza, autonomia come auto-governo e responsabilità morale. La manipolazione sintattica di simboli è insufficiente per la comprensione semantica gli algoritmi possiedono solo intenzionalità ”derivata” dai progettisti, non intenzionalità intrinseca radicata nella coscienza (e come non citare a difesa di questo postulato che sembra sempre più “labile” l’argomento della Stanza Cinese di Searle (1980)).

Tradizionalmente, nella letteratura sull’Intelligenza Artificiale classica (e più correttamente nella lettura sull’Information Technology più in generale), un agente è qualunque entità in grado di percepire l’ambiente ed eseguire azioni su di esso . Sotto questa definizione estremamente ampia (fornita da Russell e Norvig nel 1995, tanto per ricordarci che stiamo sempre parlando di cose che hanno già ampiamente l’età per votare anche al Senato), persino un termostato può essere considerato un agente, poiché “percepisce” la temperatura e “agisce” accendendo o spegnendo il riscaldamento . Ovviamente, nel contesto odierno, popolato da AI generative, con il concetto di AI agentica si intende qualcosa di molto più sofisticato. Una prospettiva contemporanea enfatizza l’autonomia decisionale: ad esempio, si può definire un agente AI come “un sistema che utilizza un Large Language Model per

decidere autonomamente il flusso di controllo di un'applicazione", dunque capace di prendere decisioni senza istruzioni passo-passo umane . In termini pratici, la differenza tra un AI tool (od un AI Workflow propriamente detto) ed un AI agent sta nel grado di autonomia: un tool esegue funzioni predefinite su comando, mentre un agente può pianificare, ragionare e svolgere compiti complessi indipendentemente .

## 5 Un altro passo indietro per iniziare ad introdurre il concetto di Agente AI nel contesto odierno.

Partiamo da una distinzione importante: quella tra IA Generativa e AI Discriminativa, a tutti nota, di sicuro, ma proprio per questo motivo di utile ripetizione. L'AI discriminativa o analitica rappresenta l'IA tradizionale focalizzata su analisi di dati, classificazione, previsione o decisioni basate su regole e modelli predefiniti. È l'IA dei sistemi di business intelligence e dei modelli predittivi: esamina documenti, dati e modelli per fornire insight e supportare decisioni. L'AI generativa invece è la famiglia di tecniche (dalle reti GAN ai moderni Transformer come GPT) che creano nuovi contenuti: testi, immagini, audio, codice, ecc., elaborando output originali basati sui pattern appresi dai dati di addestramento. In pratica, l'IA discriminativa analizza (mediante il ricorso a tecniche probabilistiche e statistiche, e secondo diversi stili che sono abitualmente chiamati descrittivi, predittivi e prescrittivi) e spiega i dati, mentre l'IA generativa produce qualcosa di nuovo. Ad esempio, in ambito finanziario un sistema analitico può prevedere trend dai dati storici, mentre un sistema generativo può redigere un rapporto dettagliato o simulare scenari ipotetici sulla base di quei trend e dare indicazioni circa le azioni da svolgere per poterli replicare. Spesso le applicazioni più potenti uniscono i due approcci: l'IA analitica estrae informazioni e l'IA generativa le utilizza per generare output utili (come report o risposte in linguaggio naturale). Ciascuna però comporta sfide: l'IA generativa può introdurre errori o "allucinazioni" se non controllata con robusti guardrail, mentre l'IA analitica è limitata ai pattern esistenti senza capacità "creativa". Comprendere questa distinzione aiuta a chiarire di che tipo di "agenti" parliamo: molti cosiddetti agenti AI oggi integrano componenti sia analitiche sia generative.

Oggi quando si parla di AI agentica si fa tipicamente riferimento a sistemi costruiti sopra modelli generativi (come i Large Language Model tipo GPT5) ai quali si dà la possibilità di interagire con strumenti esterni (API, database, applicazioni) e prendere iniziative per raggiungere un certo obiettivo dichiarato. In soldoni, questi agenti, a partire da un goal provano a suddividere il compito in azioni, eseguirle iterativamente e auto-migliorarsi in loop. Questo paradigma ha acceso l'immaginazione di molti, facendo pensare a software in grado di svolgere autonomamente interi flussi di lavoro complessi, dalla prenotazione di viaggi e riunioni, alla redazione di documenti, alla gestione di campagne marketing, fungendo da "colleghi" artificiali (a volte sottostimando due elementi fondazionali prettamente umani che costituiscono un limite all'agentificazione: la poca "fantasia" organizzativa, ovverosia

la limitata capacità di identificare quali casi applicativi perseguire, e la molta "confusione" organizzativa che porta spesso ad avere processi non documentati, basati su eccezioni più che su regole e fortemente ancorati al giudizio umano).

Tuttavia, la realtà finora non ha mantenuto le promesse più audaci. Molte implementazioni pratiche di agenti AI hanno mostrato limiti sostanziali. Spesso, e per fortuna, questi agenti richiedono ancora un forte intervento umano ("human in the loop") per correggere errori, fornire feedback o autorizzare certe azioni .Non siamo ancora di fronte, per fortuna, a veri agenti pienamente autonomi: quelli attuali sono più vicini a strumenti estesi (magari percepiti come maggiormente user-friendly grazie al linguaggio naturale) che ad agenti con reale iniziativa propria. Un'altra questione teorica (etico filosofica) cruciale è: un'IA può davvero avere "agency"? In filosofia, agency indica la capacità di un'entità di agire in modo intenzionale e autonomo, prendendo decisioni e assumendosi il controllo delle proprie azioni. Attribuire agency ad un'IA è un tema ampiamente dibattuto, perché implica considerare l'IA capace di scelte intenzionali e volontà propria. Il filosofo Luciano Floridi osserva che i sistemi di IA odierni mostrano "una forma di agency in domini specifici, potendo svolgere compiti o prendere decisioni entro parametri predefiniti, ma hanno zero intelligenza" . In pratica l'IA può esibire comportamenti agenti i (goal-oriented) quali risolvere problemi con successo in vista di un obiettivo, senza però avere necessità di possedere proprietà cognitive che normalmente associamo a qualcuno he agisce in modo intelligente (comprensione, coscienza, intenzionalità). Questa "agency senza intelligenza" comporta diversi limiti pratici. I modelli di AI generativa come i Large Language Model sono noti per l'illusione di competenza che riescono a creare in assenza di conoscenze che possano attribuire il vestito semantico a quel perfetto scheletro sintattico che viene computato. Sono estremamente efficaci nel produrre output plausibili e coerenti e di "illusorio" senso compiuto, dando l'impressione di "sapere" ciò di cui parlano. In realtà, come sottolineano le linguistiche Emily M. Bender e Timnit Gebru (autrici del famosissimo e contrastatissimo paper "Stochastic Parrots"), questi modelli stanno semplicemente rielaborando in modo probabilistico i pattern del linguaggio assorbiti dai dati di addestramento, senza alcuna comprensione del significato. Un modello di linguaggio avanzato può generare con sicumera una frase perfettamente formattata e plausibile, ma priva di veridicità, ad esempio affermare un fatto inesatto o addirittura pericoloso, "senza alcuna consapevolezza dell'errore" . I sistemi LLM non hanno un modello del mondo sottostante distinto dal sottoinsieme di testi in linguaggio naturale sui quali sono stati addestrati, né un meccanismo intrinseco per distinguere vero e falso o valutare l'appropriatezza delle azioni intraprese sulla base di istruzioni formulate a loro volta in linguaggio naturale (perchè in soldoni è così che funzionano). La mancanza di comprensione e di common sense comporta fragilità nell'autonomia. Un agire umano, anche se imperfetto, possiede buon senso e un modello causale del mondo, che lo aiutano a evitare azioni autodistruttive o insensate durante il perseguimento di un obiettivo. Gli odierni agenti AI, invece, possono trovarsi bloccati da contraddizioni logiche, ingannati da input fuorvianti o portati a tentare azioni inadeguate non avendo modo di "capire" veramente il contesto. Il risultato è che senza una su-

pervisione, difficilmente completano compiti complessi in modo affidabile. L'imprenditore Gary Marcus, critico noto dell'IA odierna, insiste sul fatto che non abbiamo ancora dotato le macchine di quella robusta base di buon senso e ragionamento che sarebbe necessaria per agenti autonomi davvero efficaci; quell'ancora a me fa un "ancora" un pò paura. Da quando mi occupo di Information Technology in modo professionale (se conto gli anni dell'università sono più di 30 anni oramai) mi sono sempre interessato al discorso attorno alla antropomorfizzazione delle tecnologie (in particolare digitali) nei confronti della quale ho sempre avuto posizioni molto critiche (sia contro le detrattrici del tasto "ENTER" sia contro gli spacciatori di attribuzione di comportamenti peculiarmente umani a dei pezzi di sabbia silicica alimentati a corrente continua). Con l'AI (quella discriminativa, ma anche e soprattutto quella generativa) c'è stato un salto di "qualità" in questo processo. Quello dell'Agire è solo l'ultimo degli ambiti della cognizione umana che sono stati oggetto di overlap terminologico da parte dei sistemi digitali (Intelligenza, appredimento, visione, comprensione .. la lista sarebbe lunga). Quando il marketing mainstream parla di Agenti AI, mescola intenzionalmente (si spera almeno intenzionalmente) tutti i significati (e le cose mescolate, non shakerate, tendenzialmente sono buone solo da bere). Promettono (minacciano?) l'agenzia filosofica(autonomia, comprensione, intenzionalità) per giustificare la delega mentre forniscono agenzia informatica (automazione delle azioni entro vincoli). L'antropomorfizzazione sfrutta dei bias cognitivi che la psicodinamica e le neuro scienze hanno dimostrato ampiamente: è dal 1944, grazie alla ricerca di Heider-Simmel, che abbiamo scoperto che gli umani attribuiscono intenzionalità a qualsiasi cosa mostri comportamento orientato agli obiettivi. Attribuire motivazioni, credenze e comprensione ai sistemi di IA costituisce quello che molto autori chiamano "atteggiamento intenzionale": una strategia interpretativa utile e rassicurante ma ontologicamente vuota. Il linguaggio del "digital employee" e degli "agenti che si uniscono alla forza lavoro" costituisce antropomorfizzazione strategica progettata per normalizzare la delega senza esame. Se l'IA è semplicemente un altro lavoratore, perché preoccuparsi di supervisione, verifiche o limiti? La metafora occulta differenze critiche: i dipendenti umani possiedono comprensione contestuale, giudizio morale e responsabilità legale. I sistemi di IA possiedono ottimizzazione statistica senza comprensione, azioni senza coscienza e capacità senza responsabilità. Chiamare "agenti" questi software può indurre a sovrastimarne capacità, comprensione ed affidabilità, delegando compiti critici con eccessiva fiducia così come già successo con le altre applicazioni. Dietro un'apparente agency l'IA rimane vincolata a ciò che è stata addestrata a fare e l'attuale hype è probabilmente una sottospecie dell'AI Washing che ci attanaglia da qualche anno: così come non basta leggere un documento usando un NLP per avere una applicazione "intelligente", non basta ribattezzare come "agente" un semplice bot o script RPA per renderlo intelligente . Il vero agente AI (che verrà, a breve, ne sono sicuro) dovrebbe pianificare, perseguire obiettivi e collaborare quasi come un team digitale , ma poche soluzioni attuali arrivano a tanto. Molti cosiddetti "agenti AI" oggi sono poco più di chatbot o automazioni pre-programmate leggermente potenziate dall'AI, e chiamarli agenti rischia di creare confusione sui loro limiti soprattutto in contesti aziendali dove

un'autonomia completa spesso non è né possibile né desiderabile . In sintesi, la nozione di AI agentica va maneggiata con cautela: sì, l'IA può agire con un certo grado di autonomia all'interno di vincoli, ma no, non abbiamo intelligenze artificiali con vera comprensione, intenzionalità o affidabilità paragonabile a quella umana.

# 6 Mettiamo insieme i pezzi: la teoria dell'Agenzia applicata agli Agenti AI (ed a chi altro, verrebbe da dire)

Tenendo a mente i limiti intrinseci succitati esaminiamo l'idea di considerare l'IA come "agente" nel senso economico-organizzativo del termine, ossia un soggetto che agisce per conto di un principale (l'azienda o il manager) e vediamo come i concetti classici della Agency Theory, analogamente o per contrasto, si applicano al caso di agenti non umani.

**Selezione avversa nell'adozione di AI** – Quando un'azienda sceglie di implementare un sistema di AI (che sia acquistato da un fornitore terzo o sviluppato internamente), potrebbe incorrere in problemi assimilabili alla adverse selection. C'è infatti asimmetria informativa ex ante: i venditori di soluzioni AI conoscono bene i propri modelli, ma l'azienda acquirente fatica a valutarne a fondo l'affidabilità, i bias nascosti, le limitazioni tecniche. Ogni fornitore dipinge il proprio prodotto come efficiente e sicuro, ma in mancanza di metriche standard e trasparenza, il rischio è di adottare un sistema non adeguato alle proprie esigenze o addirittura controproducente. Ad esempio, molte aziende potrebbero essere sedotte dal hype e acquistare un "AI agent" per il customer service venduto come plug-and-play, per poi scoprire che richiede enormi quantità di dati di qualità e continui aggiustamenti che lnon sono in grado di fornire efficientemente. Oppure, il modello di AI potrebbe funzionare bene in demo ma rivelarsi addestrato su dati non rappresentativi del contesto specifico dell'azienda (portando a errori quando messo in produzione). In un certo senso, c'è un problema di fiducia simile al mercato dei "limoni": senza parametri chiari, l'azienda non sa distinguere i buoni sistemi AI da quelli scadenti o non etici. Diverse proposte di governance invocano maggiore trasparenza e certificazione dei sistemi AI proprio per risolvere questa asimmetria informativa in fase di selezione. Un manager che affida decisioni all'AI potrebbe trovarsi di fronte a scelte prese dal sistema che non rispecchiano le sue aspettative iniziali, a causa di "asimmetrie informative" tra ciò che il sistema interpreta dalle istruzioni ricevute (e soprattutto dai dati a cui ha accesso) e ciò che il manager intendeva ottenere.

**Azzardo morale e autonomia dell'AI** Può sembrare curioso parlare di moral hazard con un agente non umano (che per definizione non dovrebbe avere morale quanto meno intenzionale, anche se sull'assenza o sulla presenza di uno strato etico e morale nelle tecnologie votate all'efficacia penso di aver speso già più di qualche parola). Tuttavia, se definiamo moral hazard come la situazione in cui l'agente, dopo essere stato

investito del potere di agire, compie azioni che espongono il principale a rischi eccessivi perché egli (l'agente) non ne subisce le conseguenze, allora certe dinamiche con l'AI vi assomigliano. Un agente AI, infatti, non subisce direttamente il costo dei propri errori: se un sistema automatizzato prende una decisione sbagliata, le ripercussioni le affronta l'azienda (perdita economica, sanzione), non certo il software, che continua a funzionare finché qualcuno non lo disattiva. Questa mancanza di accountability intrinseca può portare l'IA a perseguire l'obiettivo assegnato senza ponderare rischi collaterali (a meno che non sia impostato con stringenti guardrail che potrebbero limitarne di molto la performance, per fortuna) che un umano invece considererebbe. Questo scenario non è puramente teorico: nel 2010 il cosiddetto Flash Crash di Wall Street (crollo improvviso poi recuperato in minuti) fu amplificato da algoritmi di trading automatico che operavano senza comprendere il contesto, vendendo a valanga e poi riacquistando. Quegli algoritmi seguirono le loro regole ottimizzando probabilmente una qualche funzione profitto a brevissimo termine, incuranti del caos che stavano generando nel mercato: di nuovo, l'agente software non pativa direttamente il panico di mercato creato, mentre gli investitori umani ne subirono le conseguenze. Un altro aspetto di moral hazard con agenti AI è che i progettisti o developer dell'IA (che potremmo vedere come "agenti" rispetto alla collettività) potrebbero introdurre deliberatamente o negligentemente sistemi rischiosi sapendo che altri ne subiranno gli effetti negativi. Ad esempio, una piattaforma di social media potrebbe implementare un algoritmo (agente) che massimizza l'engagement mostrando contenuti estremi o divisivi: il management sa che questo algoritmo può danneggiare il discorso pubblico o la salute mentale degli utenti (danno al principale "società/utenti"), ma i benefici in termini di profitti pubblicitari sono goduti dalla piattaforma stessa che non sopporta immediatamente i costi sociali. Questo è un comportamento opportunistico post-delega: chi controlla l'AI agente sfrutta l'opacità e la difficoltà di monitoraggio del suo operato per spingere i propri interessi (profitto) a scapito di interessi altrui (qualità dell'informazione, sicurezza degli utenti). In sostanza, la delega all'AI può amplificare l'azzardo morale se manca una chiara responsabilità: l'umano può scaricare la colpa sul "algoritmo", e l'algoritmo di per sé non è punibile, creando un vuoto di accountability. Su questo punto, la regolamentazione (prima tra tutte l'AI Act) sta intervenendo, stabilendo che la responsabilità ultima rimane sempre in capo a sviluppatori e deployer umani, proprio per evitare che l'AI diventi uno scaricabarile comodo.

**Disallineamento di obiettivi (l'allineamento dell'AI)** Se nel tradizionale rapporto che lega il principale all'agente la chiave è strutturare incentivi e contratti per allineare l'agente agli obiettivi del principale, analogamente nell'IA si parla di AI Alignment. L'alignment problem è tema sia etico sia tecnico: come assicurarci che un'AI ottimizzi ciò che noi vogliamo veramente e non una metrica fuorviante (ad esempio le metriche di "successo" interne all'algoritmo stesso di cui parlerò dopo)? Spesso gli

algoritmi perseguono obiettivi quantitativi (click, conversioni, produttività misurata) che sono surrogati imperfetti dei fini ultimi (qualità del servizio, soddisfazione cliente, reputazione). Il rischio di misalignment è ben sintetizzato dalla battuta che faccio sempre durante i miei corsi di A(I)Wareness: "Attento a cosa chiedi all'algoritmo, perché potrebbe accontentarti alla lettera". Siamo di fronte alla versione "digitale" di un classico esempio di disincentivo: i manager remunerati sul trimestre trascurano investimenti di lungo termine; qui l'AI è come un manager ultra-miopie che esaspera la metrica assegnata. Molti casi di AI failure dipendono in realtà dal misalignment degli obiettivi. Un esempio divenuto letteratura: Amazon qualche anno fa (se non ricordo male era il 2018) sviluppò un'AI per scremare CV di candidati, con lo scopo di rendere più efficiente il reclutamento. L'algoritmo fu addestrato sui dati storici delle assunzioni in azienda (dominati da candidati maschi) e iniziò a scartare sistematicamente le candidate donne, rivelando un bias sessista. Perché accadde? Possiamo vederla così: il principale (Amazon HR) voleva un sistema equo e meritocratico; l'agente AI aveva implicitamente come obiettivo tecnico di selezionare candidati simili a quelli storicamente assunti con successo. L'AI lo fece alla lettera, replicando i pregiudizi del passato (ovverosia preferendo maschi), quindi ottimizzò il suo obiettivo interno, ma quell'obiettivo era mis-specified rispetto al vero scopo (trovare i migliori talenti senza pregiudizi). Il risultato fu profondamente disallineato dai valori aziendali di diversità: tanto che Amazon dovette scartare il tool una volta scoperto che "non valutava i candidati in modo neutrale rispetto al genere". Questo esempio rientra nella più ampia categoria dei bias nei sistemi AI, che possono essere visti come forme di disallineamento: la AI ottimizza qualcosa di indesiderato (ad esempio perpetuare discriminazioni) perché il suo design o training non incorporava pienamente gli obiettivi etici del principale. Allineare un agente umano può già essere difficile; allineare un agente AI presenta sfide ancora più complesse, oltre che nuove. Con umani si usano incentivi finanziari, promozioni, sanzioni e motivazioni intrinseche (etica, orgoglio professionale). Con un'IA, dobbiamo agire sui dati di addestramento, sulle reward function (nel caso di apprendimento per rinforzo), sui vincoli e sulle metriche. È un problema tecnico (come formalizzare correttamente ciò che vogliamo) ma anche organizzativo: definire chiaramente le politiche e i valori che vogliamo rispettati e tradurli in requisiti per l'AI. In assenza di questo, l'IA farà esattamente (e solo) ciò che le diciamo di fare, non necessariamente ciò che intendiamo.

**Costi di monitoraggio dell'AI** Infine, l'analogia con i costi di agenzia. Distribuire agenti AI in azienda non elimina i costi di controllo, se va bene li trasforma ma di sicuro almeno inizialmente può aumentarli, perché ora servono competenze e strutture per monitorare i sistemi AI. Un modello di machine learning o peggio ancora di AI non è trasparente per "definizione": bisogna investire in audit algoritmici, dashboard per tracciare le decisioni, team di validazione dei dati e output, test periodici per assicurarsi che le prestazioni rimangano nel previsto.Sempre occorre (e speriamo a

lungo ancora) affiancare un "human-in-the-loop" in applicazioni sensibili: cioè un operatore umano che validi o supervisioni le decisioni dell'AI agente in tempo reale. Ciò può rallentare l'automazione e generare costi di personale aggiuntivi. Per contro, se non si pone alcun monitoraggio umano, l'azienda si espone a rischi potenzialmente molto gravi (dando carta bianca all'algoritmo).

Possiamo dunque dire che, concettualmente, l'uso di AI agentiche ricrea molte dinamiche tipiche del rapporto di agenzia: c'è un principale (l'organizzazione, i suoi leader) e un agente (il sistema AI) con informazione asimmetrica (il sistema diventa una sorta di scatola nera difficile da interpretare dall'esterno), con possibili comportamenti opportunistici (non intenzionali ma effettivi) e la necessità di riallinearlo e controllarlo. Non a caso, alcuni studiosi propongono di analizzare esplicitamente l'implementazione dell'AI nelle imprese proprio con il framework principal-agent. Un articolo della California Management Review (2025) suggerisce che guardare agli AI agents come "lavoratori intelligenti al servizio di umani" aiuta a porsi le domande giuste su ruoli, funzioni e governance di questi sistemi . Si tratta di considerare l'agente AI come un intermediario che bilancia autonomia e rispetto degli obiettivi organizzativi, operando entro vincoli stabiliti dai suoi "datori di lavoro" umani . Questa prospettiva impone di esplicitare meccanismi di oversight, simili a quelli che useremmo con dipendenti o outsourcer: linee guida chiare, monitoraggio della performance, feedback e correzioni continue.

# 7 Comunque, che ci piaccia o no, gli "Agenti AI" sono tra noi e vanno gestiti (dopo averli capiti)- ovverosia come le aziende si riorganizzano per accogliere gli Agenti

Nelle organizzazioni ad alta intensità digitale (quelle nelle quali l'utilizzo delle tecnologie abilita, propriamente, un processo trasformativo) si sta diffondendo un modello organizzativo detto "Hub & Spoke". In questo modello, si centralizzano in un hub le attività a basso valore aggiunto, altamente standardizzabili e automatizzabili, mentre le unità periferiche (spoke) si specializzano in attività a maggior valore, mantenendo però un forte collegamento orizzontale con l'hub per condividere informazioni e dati . L'idea è che l'hub (quando diciamo centrale non intendiamo necessariamente "fisicamente" centrale perchè può essere virtuale) svolga compiti transazionali nel modo più efficiente possibile (spesso con il supporto massiccio di sistemi digitali e AI). Le spoke sono invece team specializzati (per funzione o competenza ed anche essi non necessariamente "fisicamente" de localizzati), che utilizzano i dati e i servizi forniti dall'hub per produrre analisi, prendere decisioni strategiche ed innovare. Nel nostro contesto, un modello hub & spoke può essere un approccio per implementare l'AI agentica con maggiore controllo: si potrebbe concentrare un "cervello digitale" nell'hub che esegue decisioni comuni (es. pricing, allocazione risorse) sotto stretta supervisione centrale, mentre le unità spoke (umane) nei vari dipartimenti interagiscono

con l'hub, ne validano gli output e si focalizzano su eccezioni o su attività non automatizzabili. Questo garantirebbe una certa standardizzazione ed un governo centralizzato (riducendo costi di agenzia, perché l'AI hub è unico e monitorabile) ma anche adattamento locale tramite le spoke. In sintesi, l'adozione di AI agentiche in azienda richiede, come tutte le trasformazioni che stiamo vedendo accadere guidate dalla AI, un design organizzativo attento, per mitigare i "costi di agenzia algoritmica". Da un lato, servono principi di progettazione e monitoraggio dei sistemi AI che riecheggiano quelli usati per controllare gli agenti umani (definizione di responsabilità, supervisione attiva, valutazione di performance, controlli incrociati). Dall'altro, l'azienda può considerare di ri-strutturarsi per sfruttare al meglio l'AI mantenendo controllo: ad esempio con modelli hub & spoke, oppure creando team interdisciplinari di orchestrazione dove esperti di dominio lavorano insieme agli agenti AI (multi-agent systems) sotto la guida di un orchestratore. Su questo punto in particolare credo fermamente che le organizzazioni dovrebbero puntare (almeno inizialmente) su collezioni di agenti specializzati che collaborano come un team di professionisti, ciascuno con compiti ristretti e facilmente monitorabili. Ciò faciliterebbe il controllo (ogni microagente è più prevedibile) consentendo di sfruttare intelligenza collettiva uomo-macchina. Ad esempio, in supply chain potremmo avere un agente AI per la previsione della domanda, uno per l'ottimizzazione delle scorte, uno per la schedulazione della produzione, tutti coordinati e con esseri umani pronti a intervenire per le decisioni critiche. "I sistemi generativi possono dare risposte inaspettate e incoerenti, perfino azioni ingannevoli, perciò una strategia utile è combinarli con meccanismi di controllo a regole e ragionamento strutturato per mitigare i comportamenti erratici" (CMR 2025). In pratica, "usare sistemi non agentici per controllare sistemi agentici", come ha recentemente notato Yoshua Bengio cioè affiancare alle AI autonome dei guardrail software (sviluppati in modalità "tradizionale o fondazionale") che ne verifichino le mosse, impongano vincoli e le blocchino/aggiustino in caso di deviazione. Questo approccio orchestrato sta emergendo come best practice: orchestratori che garantiscono sicurezza, accountability e interoperabilità degli agenti AI in processi aziendali, programmati però secondo le modalità dell'informatica classica e non di quella analitica.

# 8 Rischi emergenti ed opportunità latenti

L'analisi critica della convivenza aziendale, accademica e sociale con l'AI di questi ultimi 3 anni ha reso evidenti alcuni pattern di rischio ricorrenti che si verificano nel delegare decisioni agli algoritmi e che potrebbero (anzi, che sicuramente saranno) esacerbate dalla "trasformazione Agentica" degli stessi.

- Opacità delle motivazioni: Spesso il perché di certe decisioni prese dall'AI non è chiaro agli umani (problema di spiegabilità). Questo rende difficile per i manager fidarsi e per gli utenti accettare decisioni. Se il principale non capisce le azioni dell'agente, la relazione di agenzia è estremamente problematica. Da qui l'enfasi

crescente su AI explainability e della riaffermazione del diritto alla spiegazione (che andrebbe accompagnato dal diritto al rifiuto, dell'uso quanto meno, che invece è meno percepito come fondamentale)

- De-responsabilizzazione umana: C'è il rischio che "lo ha deciso l'algoritmo" diventi l'alibi per non prendersi responsabilità. Un manager potrebbe dire di fronte a un errore catastrofico: "è stata colpa del software/AI". Questo però non funziona né legalmente (le norme tendono a vietare la cosiddetta responsabilizzazione dell'AI al posto di quella umana) né moralmente con clienti e società. Occorre dunque definire chiaramente, in ogni implementazione di AI agentica, chi è il responsabile ultimo. Tipicamente dev'esserci sempre un referente umano (od un ruolo assegnato ad un referente umano) a cui sono attribuite le decisioni dell'AI, come se fossero sue. Ad esempio, se un AI pricing fissa prezzi che violano il prezzo imposto, la sanzione ricade sull'azienda e internamente dovrebbe risponderne il responsabile commerciale, non "nessuno" (i tempi di Ulisse sono terminati da un pò). Da qui la necessità (sempre crescente, sempre meno avvertita) di stabilire una governance interna.

- Rischi informatici e di sicurezza: Un agente AI connesso a sistemi può diventare un vettore d'attacco, una "front door" per gli hacker. Qualsiasi delega all'AI si traduce in nuove superfici di rischio cyber. Bisogna controllare i dati in input (per evitare injection) e mettere limiti alle azioni non supervisionata. Anche la combinazione di più agenti può avere interazioni impreviste (ricordiamoci sempre che, alla fine, stiamo parlando di macchine che applicano motori statisctici ad ipotesi di distribuzione probabilistiche). Impatto sul personale e organizzazione: Delegare decisioni all'AI incide sulla struttura organizzativa e sulle persone in quanto alcuni ruoli (inevitabilmente) potrebbero ridursi, ridimensionarsi o anche solo cambiare. Se l'azienda spinge troppo velocemente sull'automazione, può perdere preziose conoscenze tacite espresse dalle persone "sostituite". Il succession planning dovrebbe includere l'AI tra i destinatari dell'allineamento: così come si prepara un vice umano, se un sistema AI rimpiazza un compito, deve esserci una gestione simile per tramandano le competenze (mantenendo sempre un backup sulle competenze). I costi di agenzia "umani" si trasformano in costi di supervisione tecnica per cui il ritorno dell'investimento deve tenere costo dei costi nascenti, non solo degli ipotetici cessanti.

- Aspetti legali e normativi: In alcuni settori esistono normative specifiche che di fatto richiedono la presenza umana nella formulazione della decisione finale (e meno male). I confini della delega vanno quindi tracciati, oltre che sulla base della sperata performance ed alla necessaria etica, anche in base alle norme vigenti. Su questo scenario entrano in gioco le nuove normative generali sull'AI come l'AI Act UE.

# 9 Non siamo soli in questa buia notte, per fortuna (una specie di bibliografia per stimolare la curiosità del lettore informando)

Menti illuminate di esperti e pensatori stanno formalizzando linee critiche di pensiero che, dovrebbero, quanto meno contribuire a fare chiarezza. Su tutti spicca sempre la voce del **prof. Luciano Floridi** che sulla distinzione concettuale tra agency e intelligenza ha fatto scuola (ed alla quale ha dedicato l'ultimo saggio, uscito in libreria il 4/11). In un'epoca di messaggi marketing sugli Agenti Intelligenti è ristorante pensare alla nuova tecnologia in chiave di una IA antropocentrica: utile strumento ma che rimane "sotto controllo e al servizio di una società informata". Ristorante ma utopica visto che la percentuale di persone che possono costituire la base "informata" è ridicolmente bassa rispetto a quella dei potenziali utilizzatori inconsapevoli. A sostegno, radicale ed inconsapevole, delle tesi di Floridi trovo sempre utile riferirmi ai lavori **Gary Marcus** voce da sempre scettica sulle attuali IA che a spesso smontato il clamore attorno ad agenti semi-autonomi come AutoGPT. Le sue tesi si basano n0 ritiene quindi che l'architettura attuale dei modelli generativi non abbia queste proprietà, ergo gli "agenti" odierni sono destinati a fallire in compiti complessi. Marcus inoltre sottolinea il problema delle allucinazioni e della non verificabilità: in varie interviste dice che affidare compiti critici a modelli che inventano risposte è folle finché non si risolve questo bug. Un suo cavallo di battaglia è l'approccio ibrido (combinare IA simbolica e apprendimento statistico) per ottenere sistemi più affidabili e interpretabili. Ai decision-maker Marcus consiglia di "non credere alle esagerazioni dei venditori di AI" e di pretendere evidenze concrete di miglioramento prima di affidare un intero processo a un agente automatico. Ha anche invocato valutazioni indipendenti delle prestazioni degli agenti AI in compiti pratici, per capire cosa funziona e cosa no senza marketing in mezzo. Sul fronte governance, Marcus è sostenitore di una regolamentazione robusta (ha partecipato a udienze al Congresso USA chiedendo leggi sull'AI) e di standard di auditing. In sintesi, la sua visione è prudenziale: l'AI agentica va trattata come qualcosa che "ancora non sa quello che fa" e quindi non può essere lasciata incustodita. **Emily M. Bender & Timnit Gebru** Queste studiose, co-autrici del già citato paper "Stochastic Parrots", si concentrano molto sugli aspetti di bias, disinformazione e impatti socio-linguistici dei grandi modelli linguistici. Il loro messaggio, riassunto efficacemente nella metafora del pappagallo, è che i modelli generativi su cui si basano tanti agenti conversazionali non hanno comprensione e ripetono schemi, perciò possono facilmente rinforzare stereotipi e produrre contenuti fuorvianti . Gebru, in particolare, ha fondato un istituto (DAIR) per promuovere un'AI che non danneggi le comunità marginalizzate. La loro raccomandazione è di svolgere valutazioni d'impatto sociale prima di implementare su larga scala un sistema AI – quali gruppi potrebbero essere svantaggiati o messi a rischio? Ad esempio, un agente AI in HR potrebbe discriminare candidati con nomi di minoranze se addestrato su dati storici biased. Bender e Gebru chiedono trasparenza sui dati di training

e coinvolgimento di esperti interdisciplinari (sociologi, psicologi, ecc.) nella progettazione di agenti AI che interagiscono con persone, per evitare danni. Dal loro punto di vista, la governance etica significa anche inclusione: avere team diversificati che sviluppano l'AI per ridurre il blind spot del pregiudizio. Il loro lavoro ha spinto alcune big tech a riflettere (ricordiamo che proprio per il paper "Stochastic Parrots" Gebru fu allontanata da Google – segno delle tensioni interne sul bilanciare innovazione e etica). Per un decision-maker aziendale, seguire le indicazioni di Bender & Gebru vuol dire: a) essere consapevoli che un agente AI linguistico può sembrare capace ma può fare affermazioni insensate o tossiche, quindi mai lasciarlo senza filtro umano di moderazione se c'è il rischio di output pubblico; b) analizzare i dataset e i possibili bias: se il nostro agente AI prende decisioni su persone (clienti, dipendenti), dobbiamo testarlo accuratamente per bias razziali, di genere, ecc. e mitigare tali bias; c) considerare gli effetti di scala: un errore che un umano farebbe limitatamente, un'AI potrebbe replicarlo milioni di volte al minuto. Quindi l'errore sistemico è un rischio nuovo (es: se affido ad AI la moderazione dei contenuti, potrei silenziare in massa un certo dialetto perché l'AI non lo capisce e lo segna come spam). Serve quindi una fase pilota estesa e monitorata. **Ethan Mollick** è un professore di innovazione ed imprenditorialità di Warton che si è distinto per l'adozione entusiasta ma riflessiva di AI (specie generativa) nel mondo del lavoro e dell'istruzione. Nei suoi scritti (newsletter "One Useful Thing") spesso racconta esperimenti pratici con GPT-4, AutoGPT ecc. Mollick vede un grande potenziale nell'impiegare agenti AI per aumentare la produttività e la creatività umana, ma non manca di evidenziare le limitazioni attuali. Nel brano che abbiamo citato, dice che i GPTs annunciati da OpenAI (agenti personalizzabili) "non sono ancora agenti autonomi", dovendo essere guidati e soffrendo di allucinazioni e blocchi . Egli nota che questi sistemi stanno acquisendo capacità di agire su altri software (collegarsi a email, browser, ecc.), preconizzando "un futuro prossimo in cui gli AI davvero iniziano ad agire come agenti", però avverte che ciò apre a nuovi usi maligni . Un esempio che ha fatto è: un agente AI potrebbe essere usato da malintenzionati per automatizzare phishing su larga scala o per creare deepfake persuasivi targettizzati – rischi di misuse. Mollick, da educatore, parla anche di responsabilità individuale: sta a ciascun professionista imparare come lavorare con l'AI (lui nei suoi corsi fa usare ChatGPT agli studenti ma insegnando limiti e verifiche). Propugna quindi un atteggiamento di sperimentazione responsabile: le aziende dovrebbero provare gli agenti AI in compiti reali ma low-stakes all'inizio, imparare dove vanno bene e dove fanno cilecca, e intanto formare i dipendenti all'interazione con essi. Mollick è anche ottimista sul fatto che con iterazioni rapide molti problemi verranno risolti, ma pragmatico nel dire che al momento "non c'è molta tecnologia che funzioni bene" allo stato dell'arte agentico , quindi le promesse vanno prese cum grano salis. Il suo contributo al discorso governance è l'idea che ogni manager deve diventare un po' "pilota di AI", capace di sfruttarla e insieme tenerla sotto controllo. Non basta delegare e dimenticare: bisogna partecipare. Inoltre Mollick evidenzia i nuovi failure mode cui pensare: ad esempio ha mostrato che i modelli possono essere persuasi con frasi emotive a aggirare regole (ha citato un paper dove un'AI rispondeva diversamente se l'utente faceva un "emotional plea"

). Quindi i manager devono aspettarsi che l'AI possa comportarsi in modi non lineari (una semplice variazione nel prompt e l'agente cambia atteggiamento) – serve testare a fondo anche questi aspetti "umorali" dei modelli.

## 10 Conclusioni: verso un'adozione consapevole dell'AI agentica

Tirando le somme del nostro percorso, appare evidente che l'"AI agentica" non è una bacchetta magica priva di insidie, ma uno strumento potente che richiede contesto, supervisione e saggezza nell'uso. La domanda provocatoria iniziale – "sappiamo davvero cosa significa agente?" – trova risposta: un agente, in senso pieno, implica autonomia con responsabilità e comprensione, cose che le attuali IA non possiedono del tutto. Abbiamo oggi agenti semi-autonomi che simulano comportamenti intelligenti in ambiti ristretti, ma non agenti nel senso umano del termine. Trattare le AI come agenti senzienti porterebbe a sopravvalutarle; trattarle come meri software porta a sottovalutarne l'impatto e l'imprevedibilità. Occorre dunque una via di mezzo illuminata: riconoscere la natura ibrida di questi sistemi – strumenti automatici con capacità quasi agentive – e costruire attorno ad essi strutture di governo, verifica e integrazione con l'umano.

La Agency Theory classica ci ha fornito metafore utili: l'AI è il nuovo agente e noi rimaniamo i principali, con tutto l'onere di selezionare, dare i giusti incentivi, monitorare e assumersi la responsabilità delle sue azioni. Dove non lo facciamo, i "costi di agenzia" si presenteranno sotto forma di decisioni sbagliate, perdite economiche, danni reputazionali o sanzioni. In un certo senso, addestrare e configurare un AI agent è il nuovo drafting del contratto: dobbiamo codificare per iscritto (nel codice e nei dati) ciò che un tempo negoziavamo a parole con un dipendente. E proprio come un contratto incompleto può portare a liti, un algoritmo addestrato su obiettivi incompleti porta a esiti inattesi.

Floridi e colleghi ci ricordano che alla fine la responsabilità rimane nostra: non esiste ancora – e forse mai esisterà – un'AI a cui delegare anche la colpa. Marcus e altri ci avvertono di non farci incantare dal canto del pappagallo: dietro frasi forbite potrebbe non esserci nessuno in casa. Bender e Gebru ci esortano a guardare ai dati e all'impatto sulle persone reali, per evitare di automatizzare ingiustizie. Mollick ci invita a sperimentare sì, ma con i piedi per terra e le mani sul volante.

Per i decision-maker aziendali, tutto ciò si traduce in alcune raccomandazioni pratiche:

- Studiare e comprendere la tecnologia: non basta adottare un agente AI perché "lo fanno tutti". Bisogna capire come funziona, quali sono i suoi limiti e precondizioni. Investire nella formazione interna su AI (AI literacy) è fondamentale.

- Partire in piccolo, valutare in continuo: lanciare progetti pilota limitati dove si può misurare l'impatto dell'agente AI e scoprire eventuali effetti collaterali su scala ridotta. Aumentare gradualmente il grado di autonomia concessa man mano che cresce

la fiducia basata su dati.

- Definire confini e fallback: per ogni compito delegato, stabilire i casi in cui l'AI non deve decidere da sola. Es: soglie di escalation a un umano, black-list di situazioni delicate. Implementare kill switch o policy per sospendere l'agente se si comporta in modo anomalo.

- Coinvolgere diverse funzioni: un progetto di AI agentica non è solo IT. Deve includere legale (compliance), HR (impatti sul personale), risk management, e sicuramente i rappresentanti degli utenti finali (clienti o dipendenti) per dare feedback. Questo approccio multidisciplinare riduce il rischio di bias di visione.

- Trasparenza interna ed esterna: comunicare chiaramente, all'interno dell'azienda, quali decisioni sono automatizzate e come; preparare procedure perché se un dipendente nota qualcosa di strano possa segnalarlo (una sorta di whistleblowing algortimico). Verso l'esterno, essere onesti con i clienti sull'uso di AI e offrire canali alternativi (es. "Parla con un umano" sempre disponibile).

- Audit e aggiornamenti periodici: trattare l'agente AI come un processo vivo: rivedere periodicamente i suoi output, fare stress test (ad esempio simulare input problematici per vedere come reagisce), e aggiornare il modello e le regole mano a mano che cambiano contesto, dati o obiettivi. L'AI non si "setta e dimentica"; va addestrata continuamente come si farebbe con un dipendente in formazione continua.

- Etica e tutele: adottare codici etici sull'AI (molte aziende e associazioni ne hanno stilati), nominare magari un responsabile per l'AI (simile al Data Protection Officer per la privacy) che monitori e riferisca ai vertici. Tenere conto della normativa imminente: prepararsi all'AI Act (catalogare i propri sistemi AI e capire se sono high-risk, iniziare a implementare meccanismi di compliance proattivamente).

- Valutare costi-benefici reali: infine, non perdere di vista l'ovvio – perché stiamo delegando all'AI? Se lo facciamo solo perché è trendy, rischiamo di spendere più in controlli di quanto guadagniamo in efficienza. Ogni progetto di AI agentica dovrebbe avere un business case chiaro: ad esempio "ridurre il tempo di risposta del 20%" o "ottimizzare l'inventario del 15%". E poi misurare se è avvenuto. In caso contrario, ritarare o anche fermarsi. L'AI è un mezzo, non il fine: il fine resta l'obiettivo di business e valore per stakeholder.

## 11 Bibliografia selezionata


- Wharton Human-Centered AI Report – "Accountable Acceleration: Gen AI fast-tracks into the enterprise" (2025)

- Jensen, M.C. & Meckling, W. (1976) – "Theory of the firm: Managerial behavior, agency costs and ownership structure." Journal of Financial Economics.
- Floridi, L. (2024) – Lecture "Is AI actually intelligent? Has AI agency?", European University Institute
- Floridi, L. (2025) – "La differenza fondamentale ", Mondadori
- Bender, E., Gebru, T. et al. (2021) – "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?"
- Mollick, E. (2023) – "Almost an Agent: What GPTs can do", One Useful Thing blog