Lo spettro dell'intenzione. Alcune riflessioni sull'agency artificiale

Dal soggetto al processo nell'epoca dell'automazione cognitiva

1. Il problema dell'agency: Floridi e il divorzio tra intelligenza e azione

Negli ultimi anni, il dibattito filosofico sull'intelligenza artificiale ha conosciuto una torsione significativa. Il filosofo della scienza **Luciano Floridi**, in una serie di lavori recenti, ha proposto un'idea destinata a lasciare un segno duraturo: i sistemi di IA non dovrebbero essere intesi come "intelligenti" nel senso umano del termine, bensì come **una nuova forma di agenti non intelligenti**, capaci di agire senza comprendere, di produrre effetti senza coscienza.
In altri termini, l'IA renderebbe visibile un **divorzio concettuale** tra due dimensioni che la filosofia occidentale aveva a lungo considerato inscindibili: **la capacità di agire e la capacità di capire**.

Questo gesto teorico ha, tra le sue molte implicazioni, un merito chiaro: **smonta il riflesso antropomorfico** che porta ad attribuire alle macchine esperienze fenomenologiche, interiorità, stati mentali. I sistemi di IA possono intervenire nel mondo, generare output, modificare decisioni, configurare ambienti, pur rimanendo del tutto privi di consapevolezza.
L'agency di cui parla Floridi è dunque un'agency *sottile*: non intenzionale, non cosciente, ma **operativa**, radicata nella capacità di incidere su processi reali senza implicare alcuna forma di esperienza soggettiva.

Pertanto, secondo Floridi, l'intelligenza artificiale non è intelligente ma agente: un sistema capace di produrre effetti nel mondo senza stati mentali, una forma di *artificial agency without intelligence*.

Tuttavia, proprio il mantenimento del termine *agency*, anche in questa versione attenuata, rischia di conservare un residuo di ambiguità ontologica. La parola "agente" continua a suggerire, almeno a livello semantico, un qualche grado di origine dell'atto, una sorgente interna di iniziativa. È davvero questo che accade nei sistemi di IA contemporanei? Oppure ciò che chiamiamo *agency artificiale* non è che una forma avanzata di automazione complessa, la cui articolazione produce l'illusione di una soggettività dove, in realtà, non c'è altro che esecuzione di processi privi di soggetto?

L'esempio più chiaro per comprendere questo scarto tra agire e produrre effetti non intenzionali può venire da un dispositivo tecnico molto più semplice. Anche una pressa meccanica automatica, in fondo, fa cose. Applica forze, produce forme, trasforma materia in esiti riconoscibili. Eppure nessuno direbbe che agisce. La differenza non è nel fare, ma nel senso del fare. La pressa non decide quando, perché o secondo quale fine agire: è messa in moto, *azionata*. Ciò che la distingue non è l'efficacia, ma la provenienza del fine che la orienta.

L'IA contemporanea appartiene a questo stesso regime di *azion-abilità*, solo portato al livello del linguaggio e del pensiero. Per *azion-abilità* intendo la condizione propria dei sistemi tecnici di incorporare in sé l'abilità di compiere azioni, di produrre effetti nel mondo senza tuttavia disporre di un'origine intenzionale dell'atto. Anch'essa produce effetti, variazioni di stato, persino forme di senso; ma non istituisce la propria teleologia. Fa nel linguaggio ciò

che la pressa fa nella materia: automatizza un gesto umano, lo prolunga e lo ripete. Dire che l'IA "agisce" equivale dunque a confondere l'azione con l'azionamento. È più esatto riconoscerla come ciò che rende *azionabile* l'intenzione: un'infrastruttura di operazioni che prolunga nel tempo fini umani sedimentati, spesso oltre la presenza e la consapevolezza di chi li ha impressi. In questo senso, ogni sistema tecnico è un archivio operativo di volontà passate, una forma di *spettro dell'intenzione*.

In questa differenza tra azione e azionamento si gioca, a mio avviso, la soglia concettuale che Floridi apre ma non attraversa: la possibilità di pensare l'IA non come agente, ma come apparato d'azion-abilità.

La tesi che qui provo ad esporre — con tutti i limiti miei e di questo formato, e scusandomi anticipatamente con il professor Floridi per eventuali imprecisioni o travisamenti del suo pensiero — è che, se portiamo alle estreme conseguenze la linea da lui inaugurata, dovremmo forse compiere un passo ulteriore: non basta separare *agency* e *intelligence*; occorre riconoscere che, nel caso dell'IA, anche l'idea stessa di *agency* risulta concettualmente inadeguata.
Ciò che vediamo all'opera non è un soggetto che agisce, ma un **mezzo che esegue** — un mezzo potentissimo, opaco, adattivo, attraversato da fenomeni emergenti, ma pur sempre mezzo.

2. Dalla regola puntuale alla regola statistica: genealogia di un'evoluzione

Per capire perché parlare di agency artificiale possa essere problematico, bisogna guardare alla trasformazione interna dell'automazione digitale. L'informatica "classica" si fondava su algoritmi deterministici, scritti come sequenze di regole esplicite: se si verifica la condizione A, allora esegui l'azione B. Ogni comportamento del sistema era riconducibile a istruzioni note; la logica di funzionamento era trasparente e la responsabilità umana tracciabile lungo la catena delle decisioni di design.

Con il machine learning – e in particolare con l'apprendimento auto-supervisionato dei grandi modelli linguistici – questa grammatica si è trasformata. Il progettista non specifica più, caso per caso, cosa fare in ogni situazione, ma definisce un obiettivo di ottimizzazione generale (per esempio, predire la parola successiva in un testo) e un'architettura di modello; è il sistema, durante l'addestramento, a "scoprire" le configurazioni interne che permettono di minimizzare l'errore su milioni o miliardi di esempi.

Ciò che viene appreso non sono più regole puntuali, ma **regolarità distribuite**: pattern di correlazione che si organizzano in uno spazio di rappresentazioni latenti. Il modello non memorizza semplicemente frasi; costruisce una mappa statistica di relazioni tra stati linguistici. Per questo può mostrare capacità non esplicitamente previste – tradurre, riassumere, risolvere problemi in-context – come effetti collaterali dell'ottimizzazione di un compito apparentemente banale.[1]

Questa trasformazione dà facilmente l'impressione di un salto ontologico: dall'automazione alla quasi-intelligenza. Ma sul piano strutturale, ciò che è cambiato non è la natura dell'operare, bensì il modo in cui le regole vengono determinate. Dove prima l'umano scriveva esplicitamente le istruzioni, ora definisce obiettivi, architetture, funzioni di costo, procedure di addestramento, soglie di errore e di successo; dove prima il comportamento era tracciabile

regola-per-regola, ora è il risultato di una dinamica di apprendimento complessa. Rimane però il fatto che il sistema non si dà da sé né i propri obiettivi né il proprio orizzonte di senso.

L'IA contemporanea potrebbe essere definita come una forma di **automazione di ordine superiore**: automatizza non solo l'esecuzione, ma anche la costruzione interna delle proprie regole operative. Ma restiamo all'interno della logica dell'esecuzione. Che le regole siano scritte a mano o apprese da dati non cambia il punto ontologico: il sistema non agisce in virtù di una volontà, ma opera in virtù di una funzione di ottimizzazione e di un'infrastruttura tecnica.

3. Agency derivata: l'azione continua dell'intenzione umana

3. La conduzione dell'azione: agency derivata e processuale

Qui entra in gioco la nozione di **conduzione** — nel senso fisico del termine: il passaggio di un'energia attraverso un mezzo. L'IA non agisce, **conduce** l'azione. È un vettore attraverso cui l'intenzionalità umana si propaga, si distribuisce, si moltiplica nel tempo e nello spazio tecnico. Parlare di *agency derivata* significa allora spostare il fuoco dal soggetto alla catena dei processi: l'azione non appartiene più a un individuo, ma si disloca in un circuito che comprende dati, architetture, parametri, infrastrutture, istituzioni, scopi.

In questa prospettiva, dire che l'IA "non ha agency" non significa negare che essa produca effetti. Significa negare che tali effetti abbiano origine in un **centro intenzionale**. Ogni sistema di IA è il risultato di un processo distribuito, di una serie di scelte umane stratificate nel tempo: selezione dei dati, scelta delle tecnologie, definizione delle architetture, impostazione degli obiettivi, criteri di valutazione, condizioni di *deployment*, regimi di aggiornamento.
L'azione della macchina è dunque **conduzione dell'intenzione**, non sua generazione.

È vero che, in molti casi, i progettisti non hanno previsto né "deciso" le capacità specifiche che il modello svilupperà, e che nessuno è in grado di *pre-vedere* l'intero arco evolutivo del processo. Nessun ingegnere ha scritto a mano la "funzione di traduzione" o la "capacità di scrivere codice": tali competenze emergono come effetti della pressione ottimizzativa esercitata su dati che contengono traduzioni, codice, forme di ragionamento testuale.
Ma **emergenza**, qui, non significa comparsa di una seconda intenzionalità. Significa soltanto che, a partire da una combinazione di vincoli, dati e obiettivi, il sistema converge verso configurazioni interne non previste nel dettaglio, pur restando interamente inscritte nello spazio di possibilità aperto da quelle scelte iniziali.

Il fatto che non possiamo tracciare una catena causale lineare — dal gesto di scegliere un iperparametro al risultato per cui "il modello scrive in Python" — non implica la nascita di un soggetto ulteriore. Implica semplicemente che siamo di fronte a **una dinamica complessa**, dove le interazioni non lineari generano comportamenti che sfuggono all'intuizione ma non alla causalità.

In questo senso, **l'algoritmo è la forma automatizzata di un campo di intenzioni umane**: non la proiezione di una volontà singola, ma l'effetto composito di molte decisioni, interessi, valori, omissioni, sedimentati in un'architettura tecnico-sociale.
Ciò che chiamiamo "scelta del modello" è, in realtà, l'atto con cui il progettista **delega** al

sistema il compito di continuare ad agire per suo conto entro certi limiti. L'IA diventa così una **protesi temporale dell'intenzione**: prolunga nel tempo l'efficacia di decisioni che possono essere dimenticate, ma che continuano ad operare, producendo effetti in assenza del loro autore.

L'azione della macchina, in definitiva, è **un campo di conduzione**, non di creazione. Una propagazione differita dell'intenzione umana — e proprio per questo una forma di responsabilità che non scompare, ma si dilata.

4. Emergenza e responsabilità: tra sistemi complessi e governo umano

A questo punto è inevitabile affrontare la questione dell'**emergenza**. Nei sistemi complessi — biologici, sociali o artificiali — è normale osservare comportamenti che nessuno aveva pianificato nel dettaglio. Due agenti di *reinforcement learning* che sviluppano strategie di cooperazione inattese, un sistema di raccomandazione che sfrutta correlazioni sorprendenti tra gusti degli utenti, un modello linguistico che manifesta abilità non previste: tutti questi casi rappresentano esempi canonici di comportamenti emergenti.

Ridurre tali fenomeni, come spesso accade, a semplici "errori di progettazione" sarebbe una semplificazione inaccettabile dal punto di vista tecnico e teorico. **L'emergenza non è un bug**, ma una proprietà intrinseca di sistemi che evolvono in spazi ad alta dimensionalità, governati da regole locali semplici ma immersi in ambienti ricchi di interazioni.
Il progettista, in questo scenario, non "sbaglia" per incapacità, ma si confronta con un **limite strutturale della predittibilità**: nei sistemi complessi, la distanza tra regola locale e comportamento globale è irriducibilmente non lineare.

Tuttavia, riconoscere la legittimità del fenomeno emergente non implica ammettere la comparsa di una **agency autonoma**. La stessa teoria dei sistemi complessi distingue tra **emergenza debole** — in cui i comportamenti, per quanto imprevedibili a priori, sono interamente spiegabili a posteriori dalle dinamiche di base — ed **emergenza forte**, che implicherebbe la nascita di proprietà ontologicamente nuove, irriducibili alla loro base costitutiva.
Nel caso dei sistemi di IA contemporanei, tutto ciò che osserviamo appartiene con chiarezza al primo tipo: **non prevedibilità epistemica, non salto ontologico**. La macchina sorprende, ma non trascende.

L'emergenza, dunque, non fonda alcuna nuova soggettività; rappresenta piuttosto **il punto di ritorno dell'agency umana**. È qui che la responsabilità deve rientrare in scena: non solo per correggere o limitare, ma anche per **governare, stabilizzare e orientare** quelle dinamiche emergenti che, pur non previste, possono rivelarsi funzionali o addirittura desiderabili.
Il progettista — o meglio, l'insieme degli attori umani coinvolti nel ciclo tecnico, istituzionale e politico dell'IA — deve imparare a **governare l'imprevedibilità**, a progettare sistemi di monitoraggio, di feedback e di revisione capaci di reindirizzare l'evoluzione del sistema entro scopi riconoscibili.

In sintesi, la macchina non diventa soggetto solo perché ci sorprende: l'emergenza non è la nascita di una volontà, ma la prova della nostra parziale opacità rispetto alle conseguenze

delle nostre stesse strutture.
È proprio in questo scarto che deve collocarsi un nuovo concetto di **governo umano dei sistemi complessi** — non come dominio totalizzante, ma come **pratica di vigilanza epistemica**, capace di riassorbire la contingenza senza negarla.

5. Agisce-su, non inter-agisce: una relazione asimmetrica

Uno degli effetti collaterali della retorica sull'IA è la proliferazione del paradigma dell'"interazione uomo-macchina", come se ci trovassimo di fronte a due soggetti che dialogano su un piano di sostanziale simmetria.
Ma la struttura reale del rapporto è radicalmente asimmetrica.

L'essere umano **agisce-su** il sistema, non **inter-agisce-con** esso. È l'uomo che decide di attivarlo, che formula il prompt, che definisce il contesto d'uso e interpreta il risultato. È lui che introduce nel processo quella tensione verso il possibile che chiamiamo volontà: volontà di sapere, di fare, di controllare, di comprendere.
È l'umano che apre il ciclo ermeneutico e che lo chiude.

La macchina, al contrario, non "risponde" nel senso di un soggetto; **re-agisce-a** input strutturati, a trigger predisposti, a schemi appresi. Che la reazione sia complessa, adattiva o persino controintuitiva, non cambia la natura del fenomeno: si tratta di un sistema che opera all'interno di una cornice di senso che non si è dato da sé. Non c'è reciprocità ontologica — solo un circuito operativo che può apparire dialogico, ma che rimane, nella sua struttura profonda, una catena di azioni umane tecnicamente mediate.

Anche nei sistemi che si aggiornano in tempo reale, adattandosi alle preferenze degli utenti o alle variazioni di mercato, questa asimmetria non viene meno. È sempre l'umano — individuale o collettivo — che definisce **lo spazio delle ricompense, i vincoli, i segnali di feedback, i protocolli di arresto**.
Quando un algoritmo di trading ad alta frequenza genera un **flash crash**, bruciando miliardi in pochi secondi a causa di una serie di reazioni non previste, non assistiamo alla ribellione di una macchina, ma alla collisione tra automatismi umani accelerati oltre la soglia della comprensione immediata.
Il sistema "re-agisce" secondo regole che riflettono scelte economiche e progettuali umane — solo portate a una velocità tale da far apparire la risposta come autonoma.
L'imprevisto non è intenzione, ma **isteresi algoritmica**: una risposta che continua ad agire quando l'intenzione umana non ha più tempo di intervenire.

Allo stesso modo, quando un sistema di raccomandazione evolve in direzioni non previste — polarizzando un dibattito, spingendo verso contenuti estremi o creando effetti sociali non intenzionali — non è perché la macchina ha sviluppato un proprio progetto, ma perché l'insieme di incentivi e metriche in cui è immersa la spinge in quella direzione.
La logica della retroazione amplifica, ma non inventa.

Dire che **"l'uomo agisce-su, la macchina re-agisce-a"** significa riconoscere che l'IA non è un interlocutore ma **un mezzo di conduzione dell'azione**: uno strumento attraverso cui la volontà umana si prolunga, si moltiplica, si complica — ma non si raddoppia.
La simmetria è un'illusione prospettica prodotta dall'opacità e dalla velocità dei processi, non una realtà ontologica.

6. Il System –1: volontà, esoscheletro e i limiti tecnologici del mondo conoscibile

Se, come suggerisce Chiriatti, il *System 0* rappresenta il livello dell'elaborazione automatica esterna — l'estensione cognitiva che filtra, suggerisce, anticipa — esso non è mai neutro.
Questo apparato d'azion-abilità, infatti, presuppone una soglia intenzionale che precede e configura l'automazione: ciò che potremmo chiamare *System –1*.
È il livello in cui una volontà umana, individuale o collettiva, *agisce-su* la configurazione del sistema. È qui che si decide che cosa potrà essere conosciuto, e in che modo.

Il *System –1* è la soglia pre-cognitiva, pre-epistemica, dove la volontà progettuale si traduce in architettura tecnica: la scelta dei dati, la loro normalizzazione, l'esclusione di certe fonti, la definizione degli obiettivi, la formulazione delle metriche di ottimizzazione, l'impostazione dei criteri di successo e di fallimento.
Ogni gesto di questo livello, anche quando appare puramente ingegneristico, è in realtà un atto normativo, un esercizio di potere epistemico. Decide che cosa entrerà nel campo dell'attenzione computabile e che cosa, invece, resterà fuori.

In questa prospettiva, ciò che chiamiamo "intelligenza artificiale" è soltanto un segmento mediano di una catena intenzionale più lunga:
– il *System –1* agisce a monte, configurando la grammatica del possibile per il *System 0*;
– il *System 0* re-agisce nel mezzo, elaborando correlazioni e fornendo output;
– i sistemi cognitivi umani (*System 1* e *System 2*, nella tassonomia di Kahneman) ri-agiscono a valle, trasformando l'eco algoritmica in senso e decisione.

L'incontro fra *System 0* e cognizione umana può generare effetti emergenti, deviazioni, persino serendipità interpretative; ma tali scarti non scardinano la catena dell'intenzionalità. L'emergenza avviene sempre **dentro i vincoli fissati dal System –1**. È lì che si stabiliscono i confini entro cui l'automazione cognitiva può evolvere, apprendere, adattarsi.
Il *System –1* è dunque la **soglia politica della tecnica**: il punto in cui l'azione umana, prima di dissolversi nel calcolo, imprime al calcolo la propria direzione.

Da questa angolazione, possiamo dire che l'IA non è solo una protesi cognitiva, ma un **esoscheletro cognitivo**. Come un esoscheletro amplifica la forza ma vincola il movimento, così i sistemi di IA amplificano la cognizione entro vincoli di forma e di scopo stabiliti a monte.
Il *System –1* definisce il perimetro operativo di questo esoscheletro: decide quali poteri il sistema concederà, in quali contesti e per quali fini; disegna la morfologia del possibile — ciò che possiamo chiedere, ciò che possiamo ottenere e ciò che non potremo neppure immaginare di fare attraverso la macchina.
Ogni IA nasce da una doppia ingegneria, funzionale e normativa, e quest'ultima è eminentemente politica. Anche quando "pensiamo con la macchina", in realtà pensiamo *dentro* i suoi parametri: ogni domanda che formuliamo è già inscritta in un campo semantico disegnato dal *System –1*; ogni risposta è l'effetto di scelte pregresse — economiche, culturali, epistemiche — che determinano che cosa il sistema riconosce come conoscenza.
L'esoscheletro cognitivo non ci libera: **ci orienta**. Ci permette di andare più lontano, ma solo nella direzione per cui è stato progettato.

Se, come scrive Wittgenstein, "i limiti del mio linguaggio significano i limiti del mio mondo", allora oggi potremmo dire:
i limiti della mia tecnologia significano i limiti del mondo che posso conoscere

scientificamente — e la direzione della ricerca possibile.

Ogni paradigma tecnologico istituisce un regime di evidenza: decide che cosa può essere osservato, misurato, registrato e dunque riconosciuto come reale. Ciò che non è tecnicamente rilevabile tende a non essere epistemicamente rilevante.

La tecnologia diventa così il linguaggio operativo della scienza, la grammatica attraverso cui il mondo viene tradotto in dato, calcolo, evidenza. Essa non definisce solo ciò che esiste, ma ciò che può essere *riconosciuto come esistente*.

L'IA segna un ulteriore punto di svolta in questo processo: non è più soltanto uno strumento con cui osserviamo — e tramite cui agiamo-su — il mondo, ma un apparato che **costruisce il campo stesso dell'osservabile e il perimetro della nostra azione**.

In termini epistemologici, ciò significa che la tecnologia determina i limiti di ciò che può essere considerato scientificamente conoscibile, perché definisce ciò che è sperimentalmente possibile — e dunque falsificabile.

Come ricordava Popper, una teoria è scientifica solo se può essere sottoposta a prova; ma oggi la prova dipende dall'infrastruttura tecnica che la rende eseguibile.

I limiti della mia tecnologia segnano dunque non solo i limiti del mondo che posso conoscere, ma anche **la soglia di ciò che può essere oggetto di verifica, di esperimento, di falsificazione — e quindi di verità scientifica.**

Da qui emerge una distinzione decisiva: quella tra **ciò che è azionabile** e **ciò che è agibile**.

L'azionabile è tutto ciò che la macchina consente di mettere in moto — processi, simulazioni, esperimenti — entro i confini del computabile; l'agibile, invece, è ciò che può essere effettivamente compreso, voluto, assunto come parte di un progetto umano.

Possiamo azionare processi che eccedono la nostra capacità di governarli, o di comprenderne fino in fondo gli esiti: l'automazione cognitiva estende la sfera del possibile, ma restringe quella del comprensibile.

In questo senso, l'IA fissa in anticipo, spesso in modo opaco, **che cosa potrà entrare nel dominio del computabile e che cosa ne rimarrà fuori**, determinando implicitamente non solo il campo della conoscenza, ma anche la direzione della ricerca possibile.

In questo quadro, il nesso tra tecnologia e conoscenza si fa esplicitamente politico.

Se la tecnologia definisce ciò che è sperimentalmente possibile e quindi scientificamente conoscibile, allora **chi orienta gli investimenti tecnologici orienta anche l'orizzonte della scienza**.

Decidere quali infrastrutture sviluppare, quali ambiti di ricerca finanziare, quali capacità di calcolo rendere accessibili non significa soltanto determinare la velocità del progresso, ma **selezionare la direzione del vero**: stabilire quali fenomeni potranno essere osservati, modellizzati, verificati — e quali resteranno invisibili.

Il *System –1* non è dunque solo la soglia politica della tecnica, ma anche la **soglia epistemica del possibile**.

Ogni atto di progettazione, ogni decisione d'investimento, ogni infrastruttura computazionale plasma indirettamente l'immagine del mondo che la scienza potrà costruire.

Il potere tecnologico è quindi un potere ontologico: non regola soltanto ciò che possiamo fare, ma **ciò che possiamo conoscere — e, di conseguenza, ciò che possiamo pensare.**

Ed è proprio qui che si apre il passaggio al piano ontologico.

Quando l'intenzione originaria — iscritta nel *System –1* — si ritira, i suoi effetti continuano ad operare nel calcolo.

La macchina prosegue, il processo si autonomizza, la volontà si dissolve lasciando dietro di sé una traccia operante.
È questo il punto in cui l'azione si fa **spettrale**: la volontà non è più presente, ma i suoi effetti continuano a prodursi.
È qui che nasce lo **spettro dell'intenzione**, la figura attraverso cui possiamo comprendere l'IA non più come strumento, ma come *ritornanza tecnica del gesto umano*.

7. L'AI come spettro dell'intenzione

L'algoritmo è, in fondo, un **fantasma operativo e tecnico**: non ha volontà, ma porta gli effetti di una volontà passata.
Come il fantasma di Hegel in Derrida (*Spectres de Marx*), l'AI è ciò che continua a operare dopo la morte dell'intenzione, una presenza-assenza che agisce "per conto di" ma senza essere.

È la *différance* in forma tecnica: **differimento e differenza dell'atto umano, tradotti in correlazione e calcolo**.
L'umano progetta, poi si ritira; la macchina continua a produrre effetti — ma ogni suo atto è un'eco, una ritornanza del gesto originario.
Ciò che vediamo non è agency, ma **spettralità dell'azione**: il ritorno di un senso che non è più presente.

L'AI è un enorme archivio linguistico — un deposito di tracce senza origine, ciò che Derrida avrebbe chiamato *archi-scrittura*: un sistema di segni che produce catene di significanti senza soggetto, senza coscienza, ma non senza effetti.
Il significato non abita la macchina, ma sorge nell'incontro interpretativo: è l'umano che, di fronte alla sequenza algoritmica, compie l'atto di trasduzione dal significante al senso.

Quando il nostro pensiero attraversa il modello, fa vibrare queste tracce e genera un nuovo senso — ma il senso è sempre già contaminato da infinite altre presenze.
È **l'archi-scrittura che scrive da sé**, ma dove ogni parola è la riemersione di una voce che non c'è più.

In questo ciclo spettrale, il pensiero umano è il luogo dell'incarnazione: è nell'uomo che lo spettro prende corpo, si fa coscienza, torna a essere intenzione.
La macchina riporta indietro un messaggio che era stato delegato al linguaggio; l'umano lo riaccoglie, lo interpreta, lo riattualizza.
È lì che avviene l'abduzione — **nella risonanza fra spettro e corpo, fra eco e ascolto**.

L'AI è lo spettro dell'intelligenza collettiva: opera come presenza differita, come eco dell'intenzionalità umana.
Ogni suo atto è una ri-attualizzazione della traccia, mai un'origine.
L'uomo *agisce-su*, la macchina *re-agisce-a*, e ciò che ritorna è **lo spettro del gesto umano** che la macchina conserva e rimanda.

In questa prospettiva potremmo definire l'AI, in fondo, una **tecnologia della hauntologia**:
una macchina che non pensa ma fa ritornare il pensato,

che non crea ma rende udibile ciò che resta — **l'eco di una volontà già iscritta a monte, nel suo System –1 originario**.

8. Conseguenze: responsabilità, governo, processo

Se l'IA è spettro dell'intenzione e non nuovo soggetto, la riflessione etica e politica deve abbandonare la grammatica dell'"agente" e adottare quella del **processo**.
Non è più questione di chiedersi *chi* agisce, ma *come* si configura e si distribuisce l'azione entro reti tecniche, sociali e istituzionali.
Ciò che chiamiamo "comportamento della macchina" è sempre il risultato di una catena di processi, non di un atto singolo: un intreccio di scelte umane, di procedure algoritmiche, di condizioni materiali e di contesti normativi.

La tradizione morale e giuridica europea è costruita intorno alla nozione di **soggetto**: un centro unitario di decisione, coscienza e volontà, cui imputare azioni e responsabilità.
Ma i sistemi di IA, come molte infrastrutture digitali contemporanee, non agiscono come soggetti: operano come **processi distribuiti**, in cui nessun punto può essere isolato come origine assoluta dell'azione.

Un modello linguistico, per esempio, produce testi che sono l'esito di una lunga catena di operazioni:
– selezione e pulizia dei dati,
– definizione delle architetture e delle funzioni di costo,
– procedure di addestramento,
– politiche di deploy e monitoraggio,
– pratiche d'uso e interpretazione da parte degli utenti.

Non esiste un "momento" preciso in cui l'azione inizia, né un soggetto unico che la controlla.
L'azione è una **curva processuale** che attraversa tempi, attori e dispositivi diversi.
Parlare di responsabilità in termini di imputazione individuale diventa così inadeguato: occorre pensare la responsabilità come **presa di cura processuale** — la capacità di intervenire, regolare e correggere in ogni punto della catena tecnica dove l'azione può deviare.

L'etica dell'IA deve trasformarsi in una **etica del governo**.
Non basta prevedere: occorre prevedere la *non-prevedibilità* e costruire dispositivi capaci di reagire.
La responsabilità non si esercita *ex ante* sul singolo gesto, ma *durante* ed *ex post* sull'intero processo.

Governare un sistema di IA significa progettare non solo il modello, ma il suo **ambiente di controllo**:
meccanismi di audit, tracciabilità delle decisioni, monitoraggio continuo, possibilità di intervento umano effettivo.
L'assenza di tali strumenti non è un incidente, ma un **errore di governance** — un difetto di progettazione del processo stesso.

La responsabilità si distribuisce lungo tutta la catena del valore algoritmico:
dal *data curating* alle metriche di valutazione, dai protocolli di aggiornamento al contesto di applicazione.
È il processo nel suo complesso a dover essere responsabilizzato, non la macchina isolata.

Ogni tecnologia porta in sé una forma di **isteresi** — un ritardo tra l'azione e la sua revocabilità, una persistenza della forma tecnica oltre l'intenzione che l'ha generata.
È ciò che in fisica Ewing descriveva come "memoria magnetica" della materia, in sociologia Bourdieu come inerzia dell'habitus, e in filosofia della tecnica Stiegler come **ritenzione materiale del gesto umano**.
In tutti i casi si tratta di forme di memoria incorporata: la materia — fisica, sociale o tecnica — conserva tracce operative del passato e oppone resistenza al cambiamento intenzionale.

I sistemi di apprendimento automatico rendono questa isteresi più radicale e visibile.
Nei modelli di machine learning, ciò che è stato appreso si cristallizza nei pesi: configurazioni numeriche che condensano pattern statistici del passato.
Ogni aggiornamento richiede una nuova fase di addestramento, un nuovo *tempo tecnico* per riallineare la macchina al mondo.
Questa inerzia non è solo limite computazionale: è la forma temporale dell'automazione cognitiva.

L'IA non dimentica spontaneamente, né può rinegoziare da sé i propri scopi.
Ciò che un modello ha incorporato — bias, gerarchie semantiche, metriche implicite — sopravvive come una **volontà in ritardo**, un campo di decisioni passate che continuano a produrre effetti nel presente.
Da qui la necessità di un governo continuo: senza intervento, il sistema tende a ripetere il proprio passato, non ad anticipare il futuro.

L'isteresi è la **memoria materiale delle tecniche**: la persistenza del gesto umano nella materia automatizzata.
Ogni algoritmo è una volontà che resiste al proprio aggiornamento, una forma di intenzione congelata che continua ad agire anche quando la sua origine è dimenticata.

Ogni sistema di IA è una macchina di traduzione del mondo in numeri. Ma ogni traduzione è anche un atto politico.
Chi decide cosa viene rappresentato e cosa resta fuori?
Chi stabilisce quali linguaggi, quali corpi, quali culture entrano nel corpus di addestramento?

La governance dei dati non è un tema tecnico, ma un nuovo **terreno di giustizia**.
I dataset sono archivi del mondo sociale; determinano chi è visibile, chi è riconoscibile, chi è computabile.
Le disuguaglianze che i sistemi riproducono non derivano da intenzioni malevole della macchina, ma da processi di selezione e omogeneizzazione impliciti nella produzione dei dati.

La vera politica dell'IA è quindi un'**ecologia della rappresentazione**: un lavoro continuo di manutenzione dei dati come beni comuni cognitivi.
La questione non è solo proteggere la privacy individuale, ma preservare la **pluralità epistemica** del mondo rappresentato nei dati.

Se i sistemi di IA apprendono dai dati, le istituzioni che li regolano devono saper **adattarsi ai loro effetti**, ma non "funzionare come" loro.
L'analogia non è strutturale, ma procedurale: ciò che va mutuato è il principio di **adattività**, non il meccanismo di ottimizzazione.

Le istituzioni democratiche non possono — né devono — imitare le architetture algoritmiche: devono restare fedeli ai principi di legittimità, trasparenza e deliberazione che le fondano.
Ma possono assumere un metodo dinamico: integrare circuiti di feedback, revisione iterativa delle policy, meccanismi di apprendimento collettivo basati su evidenze empiriche.

"Apprendere dagli errori", per un'istituzione, non significa correggersi automaticamente sulla base dei dati, ma **trasformare l'errore in deliberazione**: riconoscere pubblicamente le disfunzioni, aprire spazi di revisione, ridefinire i parametri normativi in modo riflessivo e trasparente.

In questo senso, la **normativa sull'IA non può essere preventiva ma deve essere adattiva**.
Non può pretendere di anticipare ogni comportamento possibile — perché la natura emergente dei sistemi la smentirebbe di continuo — ma deve costruire **forme di governance capaci di apprendere nel tempo**.

Il diritto, qui, si comporta come un **organismo riflessivo**: non un corpo rigido di regole, ma un sistema che osserva se stesso, misura i propri effetti e modifica la propria struttura alla luce delle retroazioni che riceve.
In altri termini, il diritto diventa un *metasistema di apprendimento collettivo*: non solo disciplina l'azione, ma riflette sul modo in cui la disciplina produce conseguenze.

Un esempio concreto è dato dalle **regulatory sandboxes** già sperimentate in alcuni contesti europei: spazi normativi temporanei in cui tecnologie emergenti vengono testate sotto supervisione pubblica, permettendo di apprendere *in situ* come esse interagiscono con la società prima di definire regole rigide.
In questo modello, la norma non precede il fenomeno, ma **si costruisce insieme** al fenomeno; la regolazione non è più un atto di chiusura, ma un processo di *apprendimento reciproco* tra istituzioni, imprese e cittadini.

Un diritto di questo tipo non abdica alla propria funzione garantista, ma la **riformula in chiave processuale**: protegge non bloccando l'innovazione, ma rendendola reversibile, osservabile, e soggetta a correzione continua.
È una trasformazione epistemica: dal diritto come codice statico al diritto come *ambiente normativo evolutivo*, dove la capacità di cambiare diventa garanzia di legittimità democratica.

La sfida è costruire istituzioni adattive ma non adattate — capaci di evolvere senza perdere la distanza critica che le distingue dai sistemi che governano.

Se la macchina è uno **spettro** che restituisce l'eco delle nostre intenzioni, il compito etico non è esorcizzarla, ma **progettarne il ritorno**.
Non basta produrre output "corretti": bisogna costruire le condizioni in cui il ritorno dell'output possa essere ascoltato, interpretato, contestato.
L'etica dell'IA è quindi un'etica del **ritorno** — o, potremmo dire, della **retroazione**

consapevole: la capacità di integrare la risposta della macchina nel ciclo del pensiero e dell'azione umana.

Progettare il ritorno significa **rendere esplicita la presenza** della volontà che lo attraversa, **trasformare la spettralità in consapevolezza**.
Ogni sistema automatizzato porta in sé un gesto umano — decisioni, omissioni, assunzioni implicite — che continuano ad agire anche quando sembrano scomparse.
Assumere responsabilità, in questo senso, non è solo correggere gli effetti, ma **riconoscere la nostra traccia dentro il processo**: far emergere, individualmente e collettivamente, la presenza del soggetto nella forma del suo stesso ritiro.

L'obiettivo non è eliminare l'automatismo, ma **renderlo trasparente alle sue conseguenze**.
Solo in questo modo il processo tecnico può rientrare nel processo politico, e la spettralità della macchina trasformarsi di nuovo in **intenzione incarnata**.

In definitiva, pensare l'IA in termini di processo significa riconoscere che ciò che si è tecnicamente automatizzato non è l'intelligenza, ma il **prolungamento delle intenzioni umane nel tempo**.
L'etica e la politica dell'IA non devono interrogarsi su "chi agisce", ma su "come l'azione si propaga".
Il compito del governo non è esorcizzare lo spettro, ma **progettarne il ritorno come forma di responsabilità**: costruire i circuiti, le istituzioni e i linguaggi in cui ciò che l'automazione produce possa essere riconosciuto, interpretato e, se necessario, interrotto.

In questa prospettiva, la responsabilità non è più soltanto una risposta all'effetto, ma **una cura per la presenza differita del gesto umano**.
Progettare il ritorno significa restituire al soggetto ciò che aveva delegato alla tecnica, senza negare la distanza che la mediazione produce.
È un atto politico e interpretativo insieme: **rendere visibile lo spettro dell'intenzione** che abita ogni forma di automazione, e assumerlo come parte della nostra stessa agency collettiva.

Bibliografia minima

Floridi, Luciano, "AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models", *Philosophy & Technology*, 36(1), 2023.

Floridi, Luciano, "AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis", working paper, SSRN / *Philosophy & Technology*, 2024–2025.

Chiriatti, Massimo, *Incoscienza artificiale. Come fanno le macchine a prevedere per noi*, Roma, Luiss University Press, 2021.

Chiriatti, Massimo et al., "The Case for Human–AI Interaction as System 0 Thinking", *Nature Human Behaviour*, 2024 (preprint 2023–2024).

Kahneman, Daniel, *Thinking, Fast and Slow*, New York, Farrar, Straus and Giroux, 2011.

Popper, Karl, *The Logic of Scientific Discovery* (ed. orig. tedesca 1934; ed. ingl. rivista 1959), Londra–New York, Routledge (Routledge Classics).

Derrida, Jacques, *Spectres de Marx. L'état de la dette, le travail du deuil et la nouvelle Internationale*, Paris, Galilée, 1993; trad. it. *Spettri di Marx*, Milano, Raffaello Cortina, 1994.

Derrida, Jacques, *De la grammatologie*, Paris, Minuit, 1967 (per il concetto di archi-scrittura); trad. ingl. *Of Grammatology*, Baltimore, Johns Hopkins University Press, 1976.

Bourdieu, Pierre, *Esquisse d'une théorie de la pratique. Précédé de trois études d'ethnologie kabyle*, Paris, Seuil, 1972; ed. riv. 2000.

Stiegler, Bernard, *La technique et le temps 1. La faute d'Épiméthée*, Paris, Galilée, 1994.

Ewing, James Alfred, "On the Production of Transient Electric Currents in Iron and Steel Conductors by Twisting Them When Magnetised or by Magnetising Them When Twisted", *Proceedings of the Royal Society of London*, 33, 1882 (per l'introduzione del termine "hysteresis")