

***Human Oversight* come presidio di responsabilità: linee guida emerse dal secondo incontro della Cabina di Regia sull'Intelligenza Artificiale**

EXECUTIVE SUMMARY

Il secondo incontro della Cabina di Regia sull'Intelligenza Artificiale, promosso da SPES Academy, ha rappresentato una tappa di consolidamento di un percorso già avviato di coordinamento tra istituzioni, imprese e mondo della ricerca, finalizzato a rafforzare il contributo dell'Italia nell'attuazione del Regolamento (UE) 2024/1689. L'iniziativa conferma la volontà di costruire una governance nazionale dell'intelligenza artificiale ispirata ai principi della Costituzione digitale europea, capace di coniugare innovazione tecnologica, tutela dei diritti fondamentali e sicurezza sistemica. Il tema affrontato è la *human oversight* (supervisione umana). Quest'ultima, pur essendo uno dei principi fondanti dell'AI Act, è disciplinata peraltro nell'Art. 14, rischia, tuttavia, di restare un concetto astratto se non viene resa operativa all'interno dei processi decisionali e produttivi di aziende, pubbliche amministrazioni e istituzioni. Il confronto ha quindi mirato a individuare modalità concrete per trasformarla in un meccanismo reale di controllo e di responsabilità condivisa da inserire efficacemente nella *governance* dell'AI, analizzando le sfide concrete che la supervisione umana comporta.

Dalla discussione è emersa una convergenza sostanziale tra istituzioni e imprese sull'esigenza di trasformare la supervisione umana da principio dichiarativo a criterio operativo della governance dell'intelligenza artificiale. Dopo una sintesi delle risultanze e delle proposte discusse nel corso della riunione, il *position paper* e le linee guida di seguito presentati valorizzano questa sintesi, delineando un doppio orizzonte: da un lato, la responsabilità pubblica nella definizione del quadro regolatorio e nella tutela dei diritti fondamentali; dall'altro, la responsabilità privata nel tradurre tali principi in pratiche aziendali, strumenti di controllo e cultura organizzativa.

INDICE DEI CONTENUTI

<i>Executive summary</i>	1
Risultanze del secondo incontro	2
<i>Position Paper</i>	4
Human oversight: dal principio alla prassi	4
Sfide trasversali	5
Linee guida proposte.....	5

Risultanze del secondo incontro

La discussione ha beneficiato, come nel precedente incontro, della pluralità dei soggetti coinvolti (autorità, grandi aziende, accademici, esperti legali), permettendo di mappare le sfide in un quadro realistico.

Il Prof. Oreste Pollicino ha aperto la discussione sottolineando come la supervisione umana rappresenti una forma contemporanea di controllo di costituzionalità applicato all'AI: una barriera contro l'automatismo cieco e una garanzia del rispetto dei valori fondamentali. La funzione di *oversight*, ha osservato, deve essere sostanziale e non simbolica, capace di assicurare una presenza umana informata lungo l'intero ciclo di vita dell'algoritmo. Su questa base si colloca la proposta di un Framework nazionale per la supervisione umana, articolato su cinque pilastri, *spiegazione, progettazione, formazione, controllo effettivo e differenziazione*, in coerenza con l'impianto dell'AI Act e della recente legge nazionale n. 132 del 2025.

Centrale è stato l'intervento del *prefetto Bruno Frattasi, Direttore Generale dell'Agenzia per la Cybersicurezza Nazionale*, che ha richiamato l'importanza della nuova legge nazionale, che anticipa l'attuazione del regolamento europeo, quale strumento di policy making volto a stabilire principi uniformi ma adattabili alle specificità dei diversi settori. La prospettiva dell'Agenzia è quella di superare l'antagonismo, spesso percepito, tra cybersecurity e protezione dei dati personali, riconoscendone la comune missione di tutela della libertà e della sicurezza. La sfida, ha evidenziato, è garantire un equilibrio tra sicurezza del dato e rispetto della dignità della persona, in particolare nel contesto lavorativo, dove l'uso dell'intelligenza artificiale deve sempre essere compatibile con i diritti dei lavoratori e con il principio di non discriminazione. Frattasi ha inoltre richiamato l'esigenza di considerare la supervisione umana anche come presidio di sicurezza delle persone, non soltanto dei sistemi, citando gli incidenti potenzialmente connessi a dispositivi indossabili e tecnologie di monitoraggio.

Nel corso del dibattito, *Alessandra Fidanzì, Head of Digital Center of Excellence di Eni*, ha illustrato l'esperienza del gruppo nella definizione di un *framework* per l'uso responsabile dell'AI, fondato su procedure di governance, policy trasversali e una forte collaborazione tra funzioni aziendali. La multidisciplinarietà, ha spiegato, è condizione essenziale per comprendere i rischi e costruire strumenti di controllo efficaci. Alla discussione ha, inoltre, contribuito *Carmelo Fontana, in rappresentanza di Google Italia*, il quale ha portato un contributo di natura operativa sul modo in cui la *human review* e la *human oversight* sono già integrate nei servizi digitali e nei prodotti della piattaforma. In particolare, ha spiegato che sia i prompt sia gli output possono essere rivisti da personale umano per garantire il rispetto delle politiche di sicurezza e la qualità dei risultati. Si tratta, ha sottolineato, di un approccio pragmatico che trasforma la supervisione umana in una caratteristica strutturale del prodotto e in un elemento di competitività sul mercato, capace di generare nuovi servizi e maggiore affidabilità nei confronti degli utenti.

Su questo, hanno integrato anche rappresentanti del settore bancario. *Andrea Daly, del gruppo Sella*, ha insistito sull'importanza di definire indicatori di rischio chiari come base per una "spiegabilità" effettiva. La supervisione, ha osservato, deve poter rilevare e prevenire comportamenti discriminatori o distorsioni negli output dei sistemi di AI. *Andrea Cosentini, Head of Data Scientist & AI di Intesa Sanpaolo*, ha posto invece l'attenzione sulla "spiegabilità" come relazione dialogica: l'operatore umano deve poter interrogare il sistema e ricevere risposte comprensibili e contestuali, così da esercitare un giudizio critico sugli output. Dunque, la "spiegabilità", in questa prospettiva, diventa una forma di dialettica cognitiva con la macchina, in cui la decisione algoritmica non è mai definitiva ma sempre soggetta a revisione umana informata.

Pietro Caminiti, Responsabile IT e Digitale di Terna, ha ribadito che la supervisione umana deve essere prevista sin dalla fase di progettazione dei sistemi, superando la logica di un'applicazione ex post. Oggi non è più sufficiente richiamare i principi di privacy by design o security by design, ma occorre passare a un approccio di *security for purpose*, nel quale la sicurezza è pensata in funzione dello scopo del sistema e della sua capacità di proteggere gli interessi collettivi e individuali. Agostino Nuzzolo, General Counsel di TIM, ha evidenziato le difficoltà operative legate alla natura automatica di molti processi, come quelli di controllo di rete, che rendono complesso l'esercizio di un controllo umano effettivo. La questione centrale diventa dunque come strutturare forme di intervento umano che siano realistiche e non puramente dichiarative.

Silvia Castagna, in rappresentanza della Presidenza del Consiglio dei Ministri, ha sottolineato che la percezione pubblica dell'intelligenza artificiale in Italia resta in larga parte difensiva, alimentata da timori di sostituzione o di perdita di controllo. È quindi necessario promuovere una narrazione istituzionale più equilibrata, che rafforzi la fiducia collettiva e la consapevolezza del valore sociale dell'AI. Da parte del settore dei *media*, la *dott.ssa Francesca De Rosa della RAI* ha richiamato l'attenzione sulla tutela del diritto d'autore e sulla necessità di preservare il contributo creativo umano. L'AI può alleggerire attività ripetitive, ma il controllo finale sui contenuti e sulle decisioni editoriali deve rimanere saldamente in capo alle persone. In tale prospettiva, la RAI sta avviando una propria cabina di regia interna per il monitoraggio dei progetti di AI e per la formazione su tematiche quali la disinformazione e l'etica dei contenuti generati artificialmente.

Infine, *Antonio Luca Donato, allievo della SPES Academy*, ha ricordato che i dati non sono mai neutrali: la presenza umana è indispensabile per riconoscere e correggere i *bias* che attraversano tanto i dataset quanto le logiche decisionali. Il professor Pollicino ha ripreso questo punto sottolineando come la supervisione umana rappresenti, in ultima istanza, un presidio di giustizia e di equilibrio sistemico, volto a garantire che la tecnologia resti strumento e non soggetto autonomo di decisione.

La discussione, presieduta e conclusa dal *Pres. De Luca*, ha dato voce a riflessioni e suggestioni che convergono nel *position paper* e nelle linee guida che sono riportate di seguito. Il principio di fondo resta però questo: integrare la supervisione umana nella progettazione dei sistemi, rafforzare la formazione multidisciplinare e costruire un modello di controllo effettivo basato su indicatori di rischio e criteri di differenziazione settoriale.

POSITION PAPER

Il presente position paper nasce dall'elaborazione collettiva delle riflessioni emerse nel corso del secondo incontro e sopra richiamate. Le posizioni espresse dai rappresentanti delle istituzioni, del settore privato e del mondo accademico hanno contribuito, con approcci differenti ma complementari, a delineare una visione condivisa della supervisione umana come presidio di responsabilità e garanzia costituzionale nel governo dell'intelligenza artificiale. Sulla base di tale confronto, è stato costruito il seguente *position paper*, che mira a offrire una lettura integrata delle diverse prospettive e a formulare linee di indirizzo utili al consolidamento di un modello nazionale di *human oversight* coerente con il quadro europeo e con la recente legge italiana n. 132 del 2025.

Human oversight: dal principio alla prassi

La supervisione umana rappresenta oggi uno dei cardini del modello europeo di governance dell'intelligenza artificiale. Il principio di *human-in-the-loop* (HITL) è la forma contemporanea del "controllo di costituzionalità" applicata all'AI: una barriera umana contro l'automatismo cieco. La supervisione, difatti, non è un principio generico, ma un presidio obbligatorio solo in tutte le situazioni in cui le decisioni o i risultati generati dai sistemi di AI possono produrre conseguenze rilevanti sulla vita, sui diritti o sulla dignità delle persone. In questi casi, come viene anche esplicitato dall'AI Act, l'intervento umano non può essere solo formale o successivo, ma effettivo, informato e capace di incidere realmente sul processo decisionale, garantendo che la tecnologia resti uno strumento e non un sostituto della responsabilità umana. E qui si apre la questione molto pratica: come si dimostra?

Pertanto, perché tale principio non resti un enunciato astratto, occorre tradurlo in una funzione sostanziale di *governance*, in grado di attribuire all'essere umano poteri effettivi di intervento, di comprensione (*explainability utile*) e di arresto dei processi automatizzati, creando una filiera di responsabilità lungo l'intero ciclo di vita dell'AI. In questo senso, il percorso delineato dalla Cabina di Regia si articola lungo cinque direttrici che costituiscono i pilastri di una *governance* matura della supervisione umana:

1. Spiegazione – *explainability*

Non basta "comprendere l'algoritmo", la vera sfida della "spiegabilità" è fornire all'operatore umano gli strumenti per **decidere meglio**. L'obiettivo non è solo rendere i sistemi intelligibili, ma costruire strumenti che consentano all'operatore di esercitare un giudizio informato e critico. Prioritariamente, la *governance* dell'AI di ogni ente deve evitare e reagire al fenomeno dell'*over-reliance*, ossia la tendenza dell'uomo a fidarsi eccessivamente della "macchina" o delle sue spiegazioni. La "spiegabilità" deve quindi essere utile, non decorativa: un dialogo cognitivo tra sistema e supervisore, non una mera restituzione di informazioni tecniche.

2. Progettazione – *design*

L'*oversight* non può essere un'aggiunta successiva, ma deve essere predisposta in fase di design. È necessario costruire circuiti che definiscano chiari segnali di allarme e indicatori per monitorare il funzionamento del sistema (ad esempio, per rilevare condotte discriminatorie). La supervisione umana è da intendersi come caratteristica fondante del prodotto, essenziale per la trasparenza e la competitività. Questo approccio segna un cambio di paradigma: si va oltre i concetti di *privacy-by-design* e *security-by-design* verso la definizione di *security-for-purpose*. Ciò implica la definizione di circuiti interni e di indicatori capaci di segnalare tempestivamente anomalie o comportamenti discriminatori, trasformando la supervisione in una componente strutturale del prodotto.

3. Formazione – *training*

La supervisione umana perde di senso se *l'operatore non conosce ciò che sta controllando*. Si ribadisce l'importanza della formazione (o *AI Literacy*), argomento affrontato durante il I incontro della cabina di regia, e di cui si richiama il *position paper*. Si deve includere consapevolezza sui propri *bias* cognitivi, l'uomo deve imparare a istruire correttamente la macchina affinché questa dia gli output desiderati. Inoltre, dato che l'AI introduce funzioni obiettivo generiche e non più lineari, la formazione non può essere solo tematica ma *multidisciplinare*, per lo sviluppo dello spirito critico. La formazione deve essere, dunque, multidisciplinare e critica: non soltanto tematica o tecnica, ma orientata a far maturare la competenza di giudizio che consente all'uomo di intervenire con cognizione di causa e di comprendere quando, e come, l'AI sbaglia.

4. Controllo effettivo

Il quarto pilastro riguarda il **controllo effettivo**, ossia la possibilità reale di interrompere o modificare l'azione automatizzata. Questo requisito si lega essenzialmente al precedente: solo comprendendo, il potere di "stop" è in grado di divenire il punto di verifica della genuinità della supervisione. La possibilità di arrestare un processo, tuttavia, non è soltanto una facoltà tecnica: implica un equilibrio organizzativo, poiché chi detiene tale potere deve disporre delle conoscenze e dell'autonomia necessarie per esercitarlo responsabilmente. Il controllo umano, per essere credibile, deve essere reso praticabile da procedure chiare, da ruoli definiti e da strumenti di conoscenza adeguati.

5. Casistica – proporzionalità

Il dibattito ha affrontato la questione se adottare criteri unici per l'intera organizzazione o, al contrario, profilare i livelli di *human oversight* in base alla classe di rischio o al contesto d'uso. Se da un lato l'AI è una tecnologia trasversale che richiede principi di carattere generale, dall'altro è emersa la necessità di diversificare l'approccio a seconda dei settori. Ciò è in linea con la recente l. 132/2025 e trova riscontro anche nell'AI Act europeo, dove si interviene proprio sulla **catalogazione del rischio**. La differenziazione è stata ritenuta fondamentale anche a livello operativo e formativo: l'uso dell'AI in un ruolo di *back office*, ad esempio, non può essere equiparato a quello in un ruolo critico umano, ma, al più, devono potersi compensare.

Sfide trasversali

Le sfide trasversali emerse nel dibattito confermano la necessità di una visione integrata tra cybersecurity e protezione dei dati personali, a beneficio di un migliore e maggiormente efficace controllo umano. In questo senso, l'antinomia fra le due discipline dovrebbe essere definitivamente superata: entrambe sono componenti essenziali della libertà e della sicurezza digitale, e non ambiti contrapposti. Laddove la cybersecurity protegge il dato, nel rispetto della privacy e della protezione dei dati, dall'altro, quest'ultima frena eventuali "abusi" e sostiene la tutela della **dignità della persona umana**, che si deve applicare in ogni settore. Proprio in nome di tale principio, è stato ribadito il divieto di forme di sorveglianza massiva o intrusiva, in particolare nei luoghi di lavoro, e la necessità di una vigilanza continua sul rispetto delle libertà fondamentali. La percezione pubblica dell'intelligenza artificiale, spesso improntata a sfiducia o timore, costituisce un ulteriore elemento di criticità. Il rafforzamento della supervisione umana è anche uno strumento di fiducia democratica: serve a dimostrare che l'innovazione tecnologica non sostituisce la responsabilità, ma la redistribuisce. La *human oversight*, nella prospettiva delineata dalla Cabina di Regia, è dunque il punto di incontro tra etica, diritto e tecnica: un presidio di garanzia costituzionale che rende effettiva la promessa europea di un'intelligenza artificiale antropocentrica, affidabile e orientata al bene comune.

Linee guida proposte

Dalla discussione sono emersi indirizzi chiari per rendere lo *human oversight* un principio operativo e non solo una dichiarazione astratta. Tanto gli attori istituzionali quanto quelli economici condividono l'obiettivo di tradurre il principio di *human oversight* in una funzione effettiva di governo dell'intelligenza artificiale, capace di coniugare tutela dei diritti fondamentali, sicurezza sistemica e innovazione.

Nel settore pubblico, la supervisione umana assume un valore costituzionale e democratico: garantisce che l'uso dell'AI rimanga coerente con l'interesse generale, con i principi della trasparenza amministrativa e con la protezione dei diritti dei cittadini. Le amministrazioni centrali e le autorità indipendenti sono chiamate a sviluppare modelli di controllo coerenti con l'AI Act e con la legge n. 132/2025, attraverso la conduzione di valutazioni di impatto sui diritti fondamentali e la formazione mirata dei funzionari.

Nel settore privato, la supervisione umana diventa invece una leva di governance e competitività. Le imprese che sviluppano o utilizzano sistemi di intelligenza artificiale ad alto rischio sono tenute a integrare la supervisione nei propri processi di progettazione, validazione e monitoraggio, passando da un approccio formale di conformità a un modello di *responsible AI* effettivo. La *human oversight* è così riconosciuta come componente di qualità del prodotto, ma anche come strumento di fiducia per consumatori e mercati.

La Cabina di Regia ha evidenziato la necessità di una cooperazione strutturata tra pubblico e privato, fondata su un principio di co-governance: le istituzioni fissano gli standard e garantiscono la tutela dei diritti, mentre le imprese sperimentano modelli applicativi e sviluppano pratiche di audit e controllo basate su criteri condivisi:

- **Governance pubblica e governance privata:** il settore pubblico garantisce il quadro regolatorio, il monitoraggio e la tutela dei diritti; il settore privato traduce tali principi in modelli di gestione, controlli interni e sistemi di audit indipendenti.
- **Oversight-by-design:** la supervisione umana deve essere integrata fin dalla fase di progettazione dei sistemi, come caratteristica strutturale e non come adempimento accessorio. Diviene cruciale adottare un approccio *for-purpose* e progettare i sistemi AI considerando altresì la *security-for-purpose*, riconoscendo la specificità di questa tecnologia rispetto alle precedenti: definire indicatori di rischio, costruire segnali di allarme e indicatori chiari che monitorino il funzionamento del sistema.
- **Formazione e competenze:** l'alfabetizzazione all'intelligenza artificiale deve essere considerata requisito minimo per chi esercita funzioni di controllo, con percorsi formativi differenziati per ruolo e responsabilità. La formazione deve passare per lo sviluppo dello spirito critico, focalizzare la formazione sull'acquisizione di uno spiccato senso critico, necessario per contestualizzare i segnali della macchina e avere prontezza e coscienza di intervento.
- **Spiegabilità utile e decisione consapevole:** la c.d. *explainability* deve essere orientata alla decisione, permettendo all'operatore umano di comprendere il rischio, valutare la coerenza degli output e intervenire in modo informato.
- **Proporzionalità e differenziazione del controllo:** la supervisione deve adattarsi al livello di rischio, alla natura del sistema e al contesto d'uso, valutando eventualmente l'adozione di criteri di *oversight* differenziati a seconda del contesto, in linea con i principi di necessità

e proporzionalità. In questo senso, occorre definire la catena delle responsabilità in maniera chiara, individuando chi ha il potere reale di “stop” sul sistema.

- **Partecipazione e fiducia pubblica:** la supervisione umana deve favorire la trasparenza delle decisioni automatizzate e promuovere processi partecipativi che rafforzino la fiducia collettiva nell’uso dell’intelligenza artificiale.

*Hanno contribuito alla redazione del presente documento i dottori Federica Paolucci,
Davide Angelini, Alessia Dorigoni.*