Intelligenza artificiale costituzionalmente orientata: sfide e potenzialità

1. Introduzione

1.1 La trasformazione digitale nel settore giuridico pubblico.

L'intelligenza artificiale (IA) sta trasformando, tra tanti, anche il settore giuridico, promettendo miglioramenti di efficienza e vantaggi sbalorditivi[1], non trovando alcun ostacolo se non l'usuale curva di apprendimento. Il mercato farà sicuramente buon uso, anche in questo campo, dei vantaggi concessi dai nuovi paradigmi computazionali; tuttavia, il loro avvicinamento al diritto pubblico e alla Pubblica Amministrazione (PA) solleva doverose perplessità. La PA è giustamente tentata dall'adozione di questi sistemi per tenere il passo con l'irrefrenabile accelerazione delle dinamiche sociali, prospettando strumenti *data-driven* per migliorare l'efficienza del processo decisionale[2], ma questa spinta innovativa crea inevitabilmente una tensione con la necessità di preservare i valori dello stato di diritto, e non solo quelli più strettamente legati all'azione pubblica.

La dipendenza da fornitori privati, la blanda presenza normativa, l'opacità delle pratiche di addestramento sono solo alcune delle congiunzioni che rischiano di compromettere il concetto stesso di sovranità popolare, disallineando gli obiettivi democratici con logiche di mercato.

In questo intricato nodo tecnologico e culturale, il diritto come scienza deve prendere delle posizioni chiare e dimostrarsi aperto alle influenze della tecnica, onde evitare un eccessivo svantaggio rispetto alle altre discipline, senza però aver paura di far valere le proprie fondamenta teoriche. Questo processo costringerà l'intera comunità dei praticanti ad una rivalutazione delle proprie categorie.

Questo contributo si propone di sorvolare le criticità tecniche e sociali dell'uso dell'IA nel settore pubblico, valutando le opportunità esistenti ed indicando possibili accorgimenti che possano portare l'IA giuridica ad assistere e garantire un'amministrazione democratica e costituzionalmente sostenibile della Cosa Pubblica.

1.2 Il buon andamento, l'andamento efficace, l'andamento equo.

Ciò che la dottrina novecentesca non ha mancato di celebrare come culmine dell'esperienza giuridica continentale è lo schema costituzionale rigido [3], il quale considera la Costituzione come regolazione, limite e fonte dell'ordinamento sociale, sia giuridico che politico. Ciò significa che la Carta Costituzionale disciplina non solo le strutture e le modalità dell'agire pubblico, ma imprime anche una specifica direzione programmatica di stampo ideologico [4], che solo successivamente può (e deve) piegarsi alla dialettica pluralista, in una fase esecutiva. Questa idea informava i Costituenti al

---

[1] Galetta, D. U., & Corvalán, J. G. (2019) *Intelligenza Artificiale per una Pubblica Amministrazione 4.0? Potenzialità, rischi e sfide della rivoluzione tecnologica in atto*, analizzano in un'ottica comparata le potenzialità dell'IA per la PA, inclusi i benefici in termini di efficienza, riduzione di tempi e costi, miglioramento delle performance, citando rischi e sfide.

[2] Ad esempio attraverso l'analisi documentale e la previsione dei rischi, vd. par. 2

[3] Quello che ha permesso la nascita, tra molte e forse per prima, della Repubblica Italiana. Per una analisi chiara della posizione che si intende dare alla Carta Costtuzionale, confrontare con Mortati, C. (1976). *Istituzioni di diritto pubblico* (Vol. II, IX ed.). CEDAM. (In particolare le sezioni dedicate alla natura e alla rigidità della Costituzione). Nelle sue "Istituzioni", Mortati analizza approfonditamente la struttura e il significato della Costituzione italiana, inclusa la sua rigidità come garanzia dell'ordinamento e limite al potere politico. La sua concezione della "Costituzione in senso materiale" e la difesa del valore normativo dei principi costituzionali sono fondamentali per comprendere la tecnicità del tema.

[4] Dichiaratamente determinata a discostarsi il più possibile dall'infelice allineamento precedente.

momento della redazione dell'art. 97 e permea tutt'oggi il significato di "buon andamento", non lasciando alcun dubbio sul fatto che, per considerarsi tale, serva non solo adempiere efficacemente alle disposizioni del potere esecutivo, ma anche tutelare attivamente i diritti e gli interessi costituzionali [5] che emergono dai casi concreti, prima di ogni considerazione politica.

Pertanto, ogni azione di efficientamento dovrebbe essere inevitabilmente accompagnata da una rilevabile estensione delle tutele fondamentali, non potendosi considerare efficace una riforma che amplifichi solo quantitativamente i processi e che sacrifichi, anche minimamente, quel nucleo di diritti su cui si basa il sistema sociale.

Questo obbliga ogni innovazione che voglia interfacciarsi con l'azione amministrativa a dover rispettare l'insieme di tutte le tutele nella misura maggiore possibile: ciò porta con sé delle difficoltà che fanno sembrare più appetibile il mercato e le logiche privatistiche, dove il dovere di tutela è sostituito da quello di non-lesione dei diritti altrui, beatamente e legittimamente animato dalla crescita dell'efficienza e del relativo profitto.

È proprio in questi differenti approcci che si annida il "dilemma" dell'IA nella PA: i sistemi algoritmici, ontologicamente ottimizzati per l'efficienza quantitativa e sviluppati secondo logiche di mercato, **rischiano di privilegiare un "andamento efficace" (misurabile e scalabile) a scapito di quell'"andamento equo"** che costituisce il nucleo irrinunciabile del "buon andamento" costituzionalmente inteso. La sfida diventa quindi quella di verificare se e come sia possibile piegare la potenza di calcolo non solo al miglioramento delle *performance*, ma anche e soprattutto alla garanzia attiva e sostanziale dei diritti fondamentali.

Prendendo atto delle radici puramente fisiche e matematiche del fenomeno, il dilemma pratico risiede nella ricerca di una funzione capace di estrarre non solo i *pattern* più efficienti o ricorrenti, ma tutti quelli richiesti dall'insieme delle necessità giuridiche, talvolta traducendo in linguaggio-macchina degli elementi valoriali e di principio che sfuggono alla tradizionale rappresentazione numerica[6].

Qualche anno fa questa premessa sarebbe stata sufficiente ad escludere la fattibilità di una simile impresa, ma queste nuove tecniche, tali da sfidare non il fisico ma l'intelletto dell'uomo, si stanno dimostrando idonee a rendere possibile una tale prospettiva.

---

[5] Intesi come quei diritti fondamentali, ovvero "fondativi", della Repubblica, tanto da essere integrati nella sua Costituzione all'art. 2. Di ispirazione in questo tema gli scritti di Stefano Rodotà, in particolare Rodotà, S. (2012). *Il diritto di avere diritti*. Le sue opere incitano la necessità di una tutela effettiva e non meramente formale, con un ruolo attivo dello Stato nel garantire la dignità e lo sviluppo della persona umana.

[6] Krakovsky, M. (2022). *Formalizing Fairness*. Communications of the ACM, 65(8), 11–13. Questo articolo è tra i primi a palesare la necessità di formalizzare la fairness, riconoscendo che le definizioni puramente quantitative possono non cogliere appieno il contesto etico e giuridico. Si evidenzia come diverse definizioni di fairness (es. group fairness) portino a differenti metriche statistiche e come la scelta di una definizione sia essa stessa una questione valoriale.

2. Definizioni, capacità, traiettorie.

2.1 Macchine e tecniche

Il termine "Intelligenza Artificiale" aggrega in maniera suggestiva un insieme eterogeneo di tecnologie capaci di simulare aspetti dell'intelligenza umana. L'interesse crescente, sostenuto da strategie sovranazionali, testimonia come i recenti progressi computazionali abbiano finalmente catturato l'attenzione istituzionale, spostando l'IA da dominio speculativo a potenziale strumento di governo[7]. Per disvelare le sfide costituzionali che ne derivano, è tuttavia preliminare circoscrivere l'IA nel contesto giuridico-amministrativo e citarne le capacità tecnologiche essenziali.

Sebbene manchi una definizione tecnica univoca, nel quadro europeo assumono rilievo le caratterizzazioni normative che ne enfatizzano l'autonomia operativa e la capacità intrinseca di influenzare l'ambiente fisico o virtuale tramite output quali previsioni, decisioni o contenuti[8]. Parimenti cruciale è il richiamo alla necessità che tali sistemi operino per il conseguimento di obiettivi definiti dall'uomo[9], un monito essenziale per preservare l'imprescindibile ancoraggio dell'azione pubblica alle dovute responsabilità politiche e giuridiche.

Più che una singola tecnologia, l'IA per la PA rappresenta una *suite* di capacità interconnesse.

Fondamentale è **l'Apprendimento Automatico**[10] **(Machine Learning - ML)**, che abilita i sistemi a migliorare le proprie prestazioni su compiti specifici apprendendo autonomamente da grandi volumi di dati, senza essere esplicitamente programmati per ogni scenario; esso è cruciale per estrarre *insights*, identificare *pattern* complessi e formulare previsioni probabilistiche, basandosi su paradigmi come l'apprendimento supervisionato[11], non supervisionato[12] o per rinforzo[13].

Ad esso, si affianca **l'Elaborazione del Linguaggio Naturale**[14] **(Natural Language Processing - NLP)**, che permette la essenziale capacità di processare e interpretare la materia prima del diritto – il

---

[7] La Commissione Europea, con il *Libro bianco sull'intelligenza artificiale - Un approccio europeo all'eccellenza e alla fiducia* (2020, 19 febbraio) COM(2020)65 ha posto una solida base per ulteriori valutazioni istituzionali come quella della Camera del Senato e di Confindustria. Esso non solo promuove la ricerca e l'innovazione, ma delinea chiaramente le opzioni politiche per governare lo sviluppo e l'applicazione dell'IA, con un focus sulla creazione di un "ecosistema di fiducia" e sulla gestione dei rischi, specialmente per le applicazioni ad alto impatto, inclusi i servizi pubblici.

[8] Questa è la strada scelta per la definizione di "Sistemi di IA" di cui all'articolo 3 par. 1 Regolamento (UE) 2024/167 del Parlamento europeo e del Consiglio, del 13 marzo 2024, che stabilisce norme armonizzate sull'intelligenza artificiale e modifica taluni atti legislativi dell'Unione

[9] OECD. (2019). *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449. La definizione stessa di sistema IA adottata dall'OCSE (derivata dalla proposta originale dell'AI Act) include la nozione di sistemi che operano per obiettivi, e il contesto generale dei principi chiarisce che questi obiettivi devono essere allineati con i valori umani e definiti in modo responsabile.

[10] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer, è considerato un manuale universitario classico, da qui sono riprese la maggior parte delle definizioni utilizzate nel contributo.

[11] IBM. *Cosa è l'apprendimento supervisionato?*. (Consultato l'11 maggio 2025) può essere una lettura interessante e accessibile per capirne le caratteristiche fondamentali.

[12] IBM. *Cos'è l'unsupervised learning?*. (Consultato l'11 maggio 2025). *Ibid.* nota 11.

[13] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press, è il testo canonico sull'apprendimento per rinforzo. L'agente non viene istruito su quali azioni intraprendere, ma deve scoprire quali azioni producono la maggiore ricompensa provandole. Cruciali sono i concetti di agente, ambiente, stato, azione, ricompensa e policy.

[14] Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed. draft). Prentice Hall è un testo molto citato dal quale sono tratte le nozioni utilizzate nel contributo.

linguaggio – abilitando l'estrazione automatica di informazioni[15], la classificazione semantica, la sintesi, la traduzione e persino la generazione controllata di testi, fungendo da interfaccia computazionale per il dominio linguistico-giuridico.

Non vanno poi trascurati i **Sistemi Esperti** e gli approcci basati su regole esplicite, volti a codificare la conoscenza e i processi inferenziali di esperti umani (spesso tramite logiche IF-THEN o alberi decisionali) per replicarne il ragionamento all'interno di domini di conoscenza specifici e ben strutturati, supportando l'applicazione coerente e sistematica di istruzioni e procedure prestabilite, notoriamente favorite dal diritto[16].

Infine, la recente **IA Generativa (IAG)**, spesso basata su architetture multilivello[17] capaci di predire accuratamente *pattern* molto complessi, introduce nuove capacità creative; partendo da dataset di considerevoli dimensioni[18], è in grado di produrre contenuti inediti – come testi articolati, codice software funzionante o dati sintetici realistici – che mimano la creatività umana con notevole plausibilità e coerenza.

Talvolta questi processi possono richiedere una grande quantità di potenza di calcolo e risorse hardware dedicate, come GPU e TPU, per elaborare grandi quantità di dati e generare i modelli [19]. Questo costosissimo calcolo genera dei numeri, anzi, delle matrici di numeri, univoci e specifici: i pesi (weights) del modello, che influenzeranno deterministicamente ogni output. Al netto di ogni addestramento, i pesi possono essere eseguiti su hardware più modesti, persino smartphone o dispositivi *embedded*, con un consumo di risorse molto inferiore[20]. Questo perché l'inferenza, ossia l'applicazione del modello addestrato a nuovi dati, richiede operazioni di calcolo molto meno intense rispetto all'addestramento.

Queste diverse capacità tecnologiche convergono nel delineare specifiche traiettorie applicative all'interno della PA.

---

[15] Si veda a questo proposito ISTAT, *Digitalizzazione, Interoperabilità e Intelligenza Artificiale, Dritto delle Nuove Tecnologie* (2024), Par. 4.

[16] Non a caso l'interesse delle scienze giuridiche è da tempo fervido in questo senso, rintracciando in questi sistemi il nucleo della "Giurimetria" classica. Si veda già McCarty, L. T. (1977). *Reflections on TAXMAN: An Experiment in Artificial Intelligence and Legal Reasoning*. Harvard Law Review, 90(5), 837-893.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is All You Need*. Advances in neural information processing systems, 30, ha introdotto l'architettura Transformer, fondamentale per molti modelli linguistici di grandi dimensioni (LLM) e altre applicazioni di IAG.

[18] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). *Scaling Laws for Neural Language Models*. Questo studio di OpenAI ha dimostrato empiricamente come le prestazioni dei modelli linguistici di grandi dimensioni scalino in funzione della dimensione del modello, della dimensione del dataset e della quantità di calcolo utilizzata per l'addestramento. Sottolinea l'importanza di dataset molto grandi, ma la qualità implicita è necessaria per ottenere buone performance.

[19] Strubell, E., Ganesh, A., & McCallum, A. (2019). *Energy and Policy Considerations for Deep Learning in NLP*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics hanno per primi affrontato rigorosamente le prospettive ambientali ed energetiche del fenomeno.

[20] Sono state infatti scoperte tecniche che permettono di "comprimere" (in gergo, quantizzare) il modello addestrato per poter funzionare in dispositivi meno performanti con perdite di qualità insignificanti, una rassegna delle tecniche principali è estensivamente illustrata in Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2017). *A Survey of Model Compression and Acceleration for Deep Neural Networks.*

2.2. Principali aree applicative dell'IA nella PA.

Le potenzialità dell'IA nella PA si dispiegano principalmente lungo tre assi, ognuno dei quali presenta delle criticità.

Un primo asse è **l'automazione dei processi**, dove Sistemi Esperti (*in primis*), NLP e ML possono essere impiegati per efficientare compiti standardizzati e ripetitivi, privi di discrezionalità[21], mirando alla riduzione di tempi e costi[22]. Sebbene allettante e a prima vista innocua, questa traiettoria impone una vigilanza accurata della **trasparenza algoritmica**[23] – affinché l'automazione non si traduca in un annullamento delle garanzie procedurali. In questi casi risulta più palese come l'algoritmo *sia* il procedimento; pertanto, non potrà che essere una diretta "traduzione" di specifiche previsioni legislative in linguaggio-macchina. La giurisprudenza, auspicabilmente con l'aiuto della dottrina, dovrebbe mettersi alla ricerca di metodi e sistemi volti a validare questa equivalenza, ovvero assicurarsi che le istruzioni informatiche non deviino dal dettato normativo[24].

Un secondo, e forse più critico, asse applicativo concerne il supporto decisionale. In questo campo, modelli computazionali possono elaborare dati complessi per informare l'azione pubblica[25], promettendo decisioni più tempestive ed *evidence-based*. È proprio in questa arena che si manifestano i rischi più acuti per lo stato di diritto: l'opacità dei modelli, la discriminazione algoritmica insita nei dati o nella logica del sistema, e la latente deresponsabilizzazione del decisore umano mettono a dura prova i confini dell'imparzialità, della proporzionalità e della ragionevolezza. Inserendosi nell'attività discrezionale, questi problemi rischiano di essere annichiliti dalla volontà politica contingente, alimentando indebitamente percorsi potenzialmente dannosi.

Infine, l'IA può essere impiegata per **l'interazione con cittadini e imprese**, migliorando comunicazione e accesso ai servizi tramite chatbot, assistenti virtuali e percorsi personalizzati. Questo aumenta accessibilità e reattività, ma pone sfide legate alla spersonalizzazione del rapporto e alla necessità di garantire sempre un supporto umano qualificato come alternativa o escalation[26].

2.3. I driver dell'adozione: efficienza e FOMO geopolitica.

La spinta verso l'adozione dell'IA nella PA è alimentata da un insieme di fattori convergenti: la ricerca di efficienza operativa e riduzione dei costi; l'obiettivo di migliorare qualità e accessibilità dei servizi

---

[21] Ad esempio la gestione documentale, i data entry e smistamento istanze.

[22] Ibid. nota 1.

[23] Come sarà approfondito successivamente in 3.1, può essere utile ricordare il Consiglio d'Europa, Comitato dei Ministri. (2020). *Raccomandazione CM/Rec(2020) del Comitato dei Ministri agli Stati membri sull'impatto per i diritti umani dei sistemi algoritmici.*

[24] Contissa, G. (2019). *The automation of legal reasoning. A study on the logic of law and the technologies of code*. Springer, ha lavorato ampiamente sulle sfide della formalizzazione e codificazione del diritto, evidenziando come la "traduzione" di norme giuridiche in codice non sia un processo neutro ma implichi interpretazione e scelte discrezionali.

[25] Attraverso, ad esempio, l'analisi predittiva dei rischi, la valutazione ex-ante di impatto normativo o ottimizzazione allocativa delle risorse. Pellegrin, J, Colnot, L & Delponte, L 2021, Research for REGI Committee – *Artificial Intelligence and Urban Development*, European Parliament, Policy Department for Structural and Cohesion Policies, Brussels offre spunti autorevoli.

[26] Importante in questo senso il consolidamento della dottrina che orbita intorno all'interpretazione dell'art 22 Regolamento (UE) 2016/679 del Parlamento Europeo e del Consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (*General Data Protection Regulation* - GDPR), rubricato "*Processo decisionale automatizzato relativo alle persone fisiche, compresa la profilazione"*.

pubblici; la necessità di gestire la crescente complessità socioeconomica tramite decisioni *data-driven*; l'allineamento a strategie europee ed interazionali di innovazione. Questi *driver*, pur legittimi, accentuano la tensione evidenziata nell'introduzione: l'avanzata verso l'efficienza e la modernizzazione tecnologica rischia di porre in secondo piano le complesse esigenze di tutela dei principi costituzionali che devono informare l'azione amministrativa, soprattutto se nell'attuazione si palesa una "rincorsa" competitiva a realtà che non hanno mai prestato la dovuta attenzione a questi temi.

3. Rompere la scatola nera.

3.1. Materia grigia, materia opaca

Il principio di trasparenza assume nuova centralità e complessità con l'introduzione di sistemi informatici nei processi decisionali pubblici. L'opacità algoritmica, ovvero la difficoltà di comprendere i meccanismi decisionali interni di algoritmi automatici, minaccia direttamente questo principio. Tale opacità può rendere arduo per i cittadini comprendere le ragioni di decisioni che impattano sulla loro vita, limitando la loro capacità di esercitare i propri diritti e di contestare tali decisioni. L'intersezione tra digitalizzazione dei procedimenti e trasparenza diviene un nodo cruciale per la tenuta del sistema democratico, poiché la conoscibilità dei meccanismi decisionali fonda il controllo diffuso sull'operato dei pubblici poteri.

La giurisprudenza amministrativa italiana, in particolare il Consiglio di Stato, ha affrontato questo tema, facendo emergere per la prima volta il concetto di "legalità algoritmica", rafforzando la necessità di un pieno sindacato giurisdizionale sulle decisioni guidate da sistemi informatici. Sentenze quali la n. 2270/2019 e la n. 8472/2019[27] hanno stabilito principi fondamentali, riparametrando la trasparenza in termini di "conoscibilità" e "comprensibilità" dell'algoritmo, talvolta qualificando quest'ultimo come "atto amministrativo informatico" per assoggettarlo alle medesime garanzie [28]: esso deve quindi essere noto non solo tecnicamente (codice sorgente), ma anche nei suoi elementi logici e funzionali (autori, processo elaborativo, meccanismo decisionale, priorità, dati rilevanti)[29].

---

[27] Entrambe fonti di ampia scolastica e di una buona dozzina di massime, tra le quali è opportuno evidenziare un estratto della sentenza n. 8472/2019: "*In caso di ricorso ad algoritmi per l'adozione di provvedimenti amministrativi, la "caratterizzazione multidisciplinare" dell'algoritmo (costruzione che certo non richiede solo competenze giuridiche, ma tecniche, informatiche, statistiche, amministrative) non esime dalla necessità che la "formula tecnica", che di fatto rappresenta l'algoritmo, sia corredata da spiegazioni che la traducano nella "regola giuridica" ad essa sottesa e che la rendano leggibile e comprensibile. Con le già individuate conseguenze in termini di conoscenza e di sindacabilità. In senso contrario non può assumere rilievo la riservatezza delle imprese produttrici dei meccanismi informatici utilizzati i quali, ponendo al servizio del potere autoritativo tali strumenti, all'evidenza ne accettano le relative conseguenze in termini di necessaria trasparenza. (Conferma T.A.R. Lazio Roma, Sez. III, n. 9230/2018.)*".

[28] Lapidaria la sentenza n. 2270/2019: "*L'utilizzo di procedure "robotizzate" di decisione della P.A., tramite algoritmi, per quanto legittimo, non può essere motivo di elusione dei princìpi che conformano il nostro ordinamento e che regolano lo svolgersi dell'attività amministrativa. La regola tecnica che governa ciascun algoritmo resta pur sempre una regola amministrativa generale, costruita dall'uomo e non dalla macchina, per essere poi (solo) applicata da quest'ultima, anche se ciò avviene in via esclusiva. L'algoritmo, ossia il software, deve essere considerato a tutti gli effetti come un "atto amministrativo informatico" […]*". Si veda anche un recente contributo, Carullo, G. (2021). *Decisione algoritmica e intelligenza artificiale*, Diritto dell'informazione e dell'informatica, fasc. 3, 2021, p. 431-461. Federalismi.

[29] Ovvero rispettare la dottrina relativa al provvedimento, al procedimento e la sua motivazione (cfr. art. 3 l. 241/1990), ricordando però che, a detta del Consiglio di Stato: "*Non può ritenersi applicabile all'attività amministrativa algoritmica tutta la legge sul procedimento, concepita in un'epoca nella quale l'amministrazione non era investita della rivoluzione tecnologica. Il tema dei pericoli connessi allo strumento non è ovviato dalla rigida e meccanica applicazione di tutte le minute regole procedimentali della L. n. 241 del 1990, dovendosi invece ritenere che la fondamentale esigenza di tutela posta dall'utilizzazione dello strumento*

In questa nuova ottica, può diventare rilevante non confondere "inspiegabilità" e "incalcolabilità".

L'inspiegabilità è la difficoltà pratica, spesso dovuta alla complessità intrinseca di un sistema, che incontra un osservatore umano nel fornire una giustificazione chiara, concisa e semanticamente ricca del percorso logico o dei meccanismi interni che conducono quel sistema a un determinato output. Questo non implica che il processo sia caotico, non deterministico o incalcolabile: la limitatezza della nostra mente non può indurci a rifiutare ciò che va oltre il nostro intuito.

L'incalcolabilità, nel contesto della teoria della computazione, denota invece l'impossibilità effettiva di determinare l'output di una funzione o di risolvere un problema per ogni possibile input mediante un algoritmo eseguibile in un tempo finito. Questo implica che non può esistere un procedimento meccanico generale per trovare la soluzione. Non è questo ciò che accade negli algoritmi di nostro interesse, nemmeno i più complessi.

Tuttavia, emerge la necessità di un'ulteriore, cruciale distinzione. È opportuno discernere tra un'**opacità intrinseca**, o "complessità inerente", e un'**opacità indotta**, o "confidenzialità strategica".

La prima scaturisce dalla natura stessa di modelli computazionali estremamente complessi, caratterizzati da milioni o miliardi di parametri interconnessi[30]. In tali scenari, ricostruire una spiegazione lineare e integralmente comprensibile per l'intelletto umano del percorso che conduce dall'input all'output rappresenta una sfida genuina. Aspetti come la rappresentazione distribuita dei features – ossia le modalità con cui le caratteristiche salienti dell'input vengono codificate e ripartite tra i neuroni – o le dinamiche di apprendimento che emergono in spazi vettoriali ad elevata dimensionalità, pongono ostacoli significativi a un'interpretabilità completa e intuitiva. Questa forma di inspiegabilità è legata ai limiti attuali delle nostre capacità analitiche e metodologiche di fronte a sistemi di tale portata e non ai sistemi stessi, che sono e restano valutabili e riproducibili[31].

Diverso è il caso dell'opacità indotta, la quale si configura quando la mancanza di trasparenza non è una conseguenza inevitabile delle caratteristiche tecnologiche del sistema, bensì il risultato di scelte deliberate. Gli sviluppatori, o i soggetti che detengono il controllo del sistema algoritmico, possono infatti decidere di limitare la divulgazione di informazioni cruciali per una pluralità di ragioni prettamente volte a preservare un vantaggio competitivo. Talvolta, motivazioni di sicurezza, come la prevenzione di tentativi di *reverse engineering* finalizzati a usi malevoli quali gli *adversarial attacks*,

---

informatico c.d. algoritmico sia la trasparenza, da intendersi sia per la stessa P.A. titolare del potere per il cui esercizio viene previsto il ricorso allo strumento dell'algoritmo, sia per i soggetti incisi e coinvolti dal potere stesso.".

[30] Tipici delle Reti Neurali Profonde (*Deep Neural Network – DNN*) alla base dell'IAG, introdotti con Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language models are few-shot learners*. Advances in neural information processing systems, 33, 1877-1901. Questo paper introduce il modello GPT-3 ed è un esempio emblematico della scala raggiunta dalle DNN. GPT-3 ha 175 miliardi di parametri. Il paper stesso discute la dimensione del modello come fattore chiave per le sue capacità.

[31] Particolarmente stimolanti sono gli studi interattivi come quello di Olah, C., Mordvintsev, A., & Schubert, L. (2017). *Feature Visualization*. Distill.pub, 2(11), e7. Il progetto Distill.pub esplora tecniche per illustrare cosa "vedono" i neuroni nelle reti neurali. Anche se si possono identificare pattern che attivano specifici neuroni, l'articolo evidenzia come le rappresentazioni siano spesso complesse e come la funzione di un neurone sia definita dalle sue interazioni con molti altri. Non c'è spesso un singolo neurone "del concetto X", ma piuttosto il concetto X emerge da un pattern di attivazione distribuito.

o la volontà di limitare l'esposizione a responsabilità legali legate ad acquisizioni illecite di dati, contribuiscono a questa forma di riserbo.[32]

È pertanto essenziale riconoscere come una porzione significativa della percezione comune di una "scatola nera" invalicabile possa discendere più da questa opacità indotta che da una incalcolabilità intrinseca e insormontabile. Sebbene alcuni meccanismi interni delle reti neurali più sofisticate, come l'attribuzione precisa del contributo di ogni singolo neurone a una decisione, possano effettivamente rimanere complessi da verbalizzare in termini semplici e diretti, l'adozione di una **metodologia rigorosa nello sviluppo, una documentazione esaustiva e un accesso controllato ma significativo ai dati** possono ridurre drasticamente il livello di opacità generale. Infatti, la piena conoscibilità della struttura del modello, delle caratteristiche dei dati impiegati per il suo addestramento, dei processi di validazione adottati e delle metriche di *performance* utilizzate, sebbene non garantisca una spiegazione atomistica di ogni singolo calcolo, abilita comunque operazioni fondamentali. Tra queste, la possibilità concreta di effettuare un *auditing* per verificare la conformità del sistema a requisiti normativi, l'identificazione e la mitigazione di *bias* discriminatori, l'esecuzione di analisi di sensibilità[33] per comprendere come le variazioni negli input influenzino gli output, e la diagnosi di errori o delle cosiddette "allucinazioni" nei modelli generativi, risalendo alle cause probabili di output anomali o errati mediante tecniche proprie dell'*explainable AI* (XAI)[34], quali SHAP[35], LIME[36] o l'analisi dei gradienti e delle attivazioni neuronali [37].

Di conseguenza, pur ammettendo che una certa misura di inspiegabilità possa essere connaturata alle frontiere più avanzate della ricerca sull'IA, l'obiettivo di pervenire a un sistema tendenzialmente trasparente è realisticamente perseguibile attraverso la rimozione degli strati di opacità indotta e l'adozione di pratiche di progettazione, sviluppo e divulgazione improntate a una maggiore apertura, specialmente nel contesto dell'azione amministrativa, assunto il ruolo irrinunciabile della trasparenza e non ignorando la necessità di una tutela dei diritti privati coinvolti[38].

---

[32] La letteratura giuridica sul bilanciamento tra trasparenza e segreto industriale nel contesto dell'IA è in crescita. Si veda anche il dibattito intorno all'art. 22 GDPR e il diritto a "informazioni significative sulla logica utilizzata", che spesso si scontra con la protezione del know-how algoritmico. Un parere dell'Avvocato Generale della CGUE (causa C-203/22) ha toccato questo bilanciamento nel contesto del *credit scoring*, riconoscendo l'importanza dei segreti commerciali ma sottolineando che non possono essere uno scudo onnicomprensivo contro gli obblighi di trasparenza del GDPR.

[33] Per una maggiore comprensione di questa tecnica è consigliata la lettura di Vincenzo Calabrò. *Nuovi modelli di IA: Explainable AI e Hybrid AI,* ICT Security Magazine (2025).

[34] Adadi, A., & Berrada, M. (2018). *Peeking inside the black-box: A survey on explainable artificial intelligence (XAI).* IEEE Access, è una survey molto citata che fornisce una tassonomia delle tecniche XAI, discutendone i pro e i contro.

[35] *SHapley Additive exPlanations*, introdotto da Lundberg, S. M., & Lee, S. I. (2017). *A unified approach to interpreting model predictions.* Advances in neural information processing systems, 30. Questa tecnica assegna a ciascuna *feature* un valore di importanza per una particolare predizione, basandosi sui valori di Shapley della teoria dei giochi cooperativi. Il metodo garantisce proprietà desiderabili come l'efficienza (la somma dei valori SHAP eguaglia la differenza tra la predizione e la predizione media) e la consistenza.

[36] *Local Interpretable Model-agnostic Explanations,* introdotto da Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. LIME è una tecnica che spiega le predizioni di qualsiasi classificatore (o regressore) in modo interpretabile e fedele, apprendendo un modello interpretabile (es. lineare) localmente intorno alla predizione che si vuole spiegare.

[37] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). *Deep inside convolutional networks: Visualising image classification models and saliency maps* è uno dei primi e più noti lavori sull'uso delle mappe di salienza (calcolate tramite il gradiente della classe di output rispetto all'input) per visualizzare quali pixel di un'immagine sono più influenti per la classificazione.

[38] La linea europea è espressa da Mantelero, A. (2018). *AI and Big Data: A blueprint for a human rights, democracy and rule of law framework*. Council of Europe, che affronta la necessità di bilanciare la trasparenza con altri interessi legittimi, inclusa la protezione del know-how commerciale, nel contesto dell'IA.

3.2 Il pre-concetto algoritmico: *bias* e ideologia

L'analisi dell'opacità algoritmica conduce inevitabilmente al tema del *bias*, termine la cui polisemia richiede una rigorosa contestualizzazione sia tecnica sia giuridica.

Nel discorso comune e, talvolta, in quello giuridico meno avvertito, il *bias* è percepito quasi esclusivamente come un'anomalia, una deviazione patologica da un ideale di neutralità assoluta, e quindi come un difetto da emendare o eliminare. Tuttavia, una simile concezione, seppur animata da condivisibili istanze di equità, rischia di interpretare erroneamente la natura stessa dei processi di apprendimento automatico e, per converso, di rendere incomprensibile la reale portata della sfida sociale.

È fondamentale, per chiarezza, distinguere almeno due accezioni del termine rilevanti in ambito informatico. Il *bias* comunemente inteso (identificabile come *bias statistico*) è sostanzialmente un errore sistematico o una distorsione che porta il modello a produrre risultati iniqui o non veritieri, allontanandosi da un ideale di correttezza. Questa è l'accezione che immediatamente solleva preoccupazioni di natura etico-giuridica. Esiste, tuttavia, un'altra accezione, quella di *bias induttivo* (o di apprendimento), che si riferisce all'insieme di assunzioni a priori, vincoli o preferenze intrinseche che l'algoritmo di *Machine Learning* utilizza per poter apprendere dai dati e generalizzare situazioni nuove[39].

Il *bias* non è, quindi, un intruso accidentale, bensì un elemento costitutivo e, in una certa misura, indispensabile del processo di addestramento di un modello di ML.

Un modello, per "imparare" e non limitarsi a memorizzare, deve possedere dei criteri guida, delle "preferenze" su come interpretare i dati e quali *pattern* considerare significativi. Un ipotetico modello totalmente privo di *bias induttivo* si troverebbe in difficoltà: potrebbe non riuscire a discernere alcuna struttura utile nei dati, generando output casuali e privi di valore; oppure, se fosse estremamente flessibile nel tentativo di adattarsi perfettamente ai dati di addestramento senza alcuna guida, potrebbe sccadere nel cosiddetto *overfitting*[40]. Quest'ultimo fenomeno si verifica quando il modello apprende pedissequamente i dati di addestramento, inclusi il rumore e le specificità casuali, risultando poi incapace di generalizzare correttamente dati nuovi e inediti; l'*overfitting* è spesso associato a un'eccessiva "varianza" del modello, ovvero una sua elevata sensibilità a piccole fluttuazioni nei dati di addestramento[41].

La sfida ingegneristica risiede nel trovare un equilibrio: un *bias induttivo* troppo forte o errato può portare il modello a semplificare eccessivamente la realtà, ignorando informazioni cruciali e

---

[39] Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Rutgers CS tech report CBM-TR-117. Questo paper definisce il bias (induttivo) come qualsiasi base per scegliere una generalizzazione rispetto a un'altra, che non sia la stretta coerenza con le istanze di training osservate. Mitchell argomenta che i bias sono necessari per il "salto induttivo" che permette di generalizzare a nuove situazioni.

[40] Hellstrom T., Dignum V. and Bensch S. (2020, 20 settembre). *Bias in Machine Learning - What is it Good for?.* Affrontano l'argomento in maniera pragmatica ed esaustiva.

[41] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer pp 29-37. Definisce la varianza come la quantità di cui la stima di una funzione f cambierebbe se fosse stimata usando un diverso set di dati di addestramento. Idealmente, la stima di f non dovrebbe variare molto tra i set di addestramento; tuttavia, se un metodo ha alta varianza, allora piccole variazioni nel set di addestramento possono risultare in grandi cambiamenti nella stima di f.

producendo errori sistematici (cioè un *bias statistico* elevato). D'altro canto, un *bias induttivo* troppo debole può rendere il modello instabile e incline all'*overfitting* (alta varianza), annullando l'utilità. Questa interdipendenza è nota in letteratura come il "trade-off bias-varianza"[42]. Pertanto, l'obiettivo non è eliminare il *bias induttivo* in sé, ma orientarlo correttamente e gestirne gli effetti per minimizzare il *bias inteso come errore sistematico dannoso*.

In questa prospettiva, il *bias induttivo* può essere inteso come l'insieme dei "pre-concetti" del modello, la sua "ideologia" computazionale. Questi pre-concetti non sono necessariamente negativi; sono le lenti attraverso cui il modello interpreta la realtà rappresentata dai dati. Essi derivano da molteplici fonti: le caratteristiche intrinseche del *dataset* di addestramento, le scelte operate nell'ingegnerizzazione delle *features*, l'architettura stessa e la funzione obiettivo che si cerca di ottimizzare. L'algoritmo, dunque, non è una *tabula rasa*, ma un sistema che apprende e opera sulla base di un orientamento implicito o esplicito. Particolarmente insidiosa, in questo contesto, può rivelarsi l'introduzione di *bias* attraverso i meccanismi di *Reinforcement Learning from Human Feedback* (RLHF)[43], impiegati per affinare modelli linguistici di grandi dimensioni o altri sistemi di IA. In tali processi, i valutatori umani, nel fornire riscontri (*feedbacks)* per guidare l'apprendimento del modello, possono inconsciamente o consciamente proiettare i propri pregiudizi, valori e interpretazioni della realtà, che vengono così codificati nel comportamento finale del sistema. Un esempio emblematico di come i *bias* possano radicarsi e produrre effetti distorsivi significativi è il caso del software COMPAS[44], utilizzato negli Stati Uniti per la valutazione predittiva del rischio di recidiva. Nonostante le intenzioni, è emerso che il sistema tendeva ad assegnare punteggi di rischio più elevati agli individui afroamericani rispetto ai bianchi, a parità di altre condizioni, perpetuando così forme di discriminazione sistemica[45].

La criticità, quindi, non risiede nell'esistenza del *bias* in sé – che, come si è visto, è in parte ineludibile e funzionale – quanto nella sua qualità, direzione e conformità ai principi giuridici fondamentali. Il problema sorge quando i "pre-concetti" del modello incorporano o amplificano distorsioni socialmente ingiuste, discriminazioni vietate dall'ordinamento (ad esempio, basate su sesso, razza, origine etnica, religione, come sancito dall'art. 3 della Costituzione e dalle normative antidiscriminatorie), o quando conducono a risultati che violano i canoni di ragionevolezza, proporzionalità e imparzialità che devono governare l'azione amministrativa.

L'obiettivo non può più essere l'utopica eliminazione di ogni forma di *bias* – che equivarrebbe a pretendere un modello privo di capacità predittiva – bensì l'identificazione, la misurazione, la mitigazione dei *bias* dannosi e, specularmente, la promozione di *bias* "virtuosi", ossia di orientamenti

---

[42] *Ibid* nota 41, pagina 33, nonché Geman, S., Bienenstock, E., & Doursat, R. (1992). *Neural networks and the bias/variance dilemma*. Neural computation, 4(1), 1-58.

[43] Oltre al fondativo Ziegler, Daniel M.; Stiennon, Nisan; Wu, Jeffrey; Brown, Tom B.; Radford, Alec; Amodei, Dario; Christiano, Paul; Irving, Geoffrey (2019). "Fine-Tuning Language Models from Human Preferences"; Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., ... & Hadfield-Menell, D. (2023). *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback* discute in modo approfondito i problemi aperti e le limitazioni fondamentali dell'RLHF, inclusa la questione di come i *bias*, i valori e le preferenze dei valutatori umani vengano inevitabilmente "distillati" nel modello, e le difficoltà nel garantire che tali preferenze siano rappresentative o desiderabili su larga scala.

[44] Ovvero "Correctional Offender Management Profiling for Alternative Sanctions tool".

[45] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," ProPublica, May 23, 2016 e J. Angwin, J. Larson, S. Mattu, and L. Kirchner "How We Analyzed the COMPAS Recidivism Algorithm" ProPublica, May 23, 2016.

del modello che siano allineati con i valori costituzionali e gli obiettivi di equità perseguiti dal legislatore.

Trovare la "giusta combinazione", quel bilanciamento che renda il modello performante dal punto di vista tecnico e al contempo accettabile e legittimo dal punto di vista giuridico e sociale, non è un compito meramente tecnico risolvibile con formule matematiche universali. È, piuttosto, un processo dinamico e complesso che richiede un dialogo costante tra diverse competenze – informatiche, statistiche, giuridiche, etiche – e un coinvolgimento attivo della "comunità dei praticanti" intesa in senso lato, inclusi i cittadini e i loro rappresentanti. È una questione di *governance* del *bias*, che implica scelte discrezionali su quali *trade-off* siano accettabili, quali definizioni di equità (*fairness*) adottare tra le molteplici disponibili e come tali scelte debbano essere documentate, giustificate e sottoposte a scrutinio, se non a voto.

In conclusione, il "pre-concetto algoritmico", sebbene tecnicamente inevitabile, diviene giuridicamente rilevante e potenzialmente problematico quando si traduce in decisioni che ledono diritti o principi fondamentali. La sfida per un'IA costituzionalmente orientata non è negare l'esistenza del *bias*, ma comprenderne la genesi, governarne gli effetti e assicurare che l'"ideologia" del modello sia compatibile con i precetti inderogabili dell'ordinamento giuridico.

4. Verso un'IA costituzionalmente orientata: dove siamo e dove possiamo arrivare

4.1. Protocolli e intese

Si può concludere che non sia mai superfluo ribadire come al centro di ogni riflessione debba permanere la supremazia della Costituzione e dei diritti fondamentali. Questi non possono essere considerati variabili negoziabili sull'altare dell'innovazione tecnologica, ma costituiscono il perimetro invalicabile entro cui l'IA deve operare. Da ciò discende la necessità di ancorare saldamente l'impiego dell'IA ai principi cardine dell'azione amministrativa: legalità, trasparenza, *accountability* (intesa come piena rendicontabilità e attribuibilità delle responsabilità) e non discriminazione.

Per raggiungere tale obiettivo, è imprescindibile un progressivo abbattimento del confine concettuale e operativo tra la materia informatica e quella giuridica. Un punto di partenza può essere il concetto di "protocollo", inteso in senso ampio.

Come l'informatica si basa su protocolli di comunicazione[46] e di elaborazione [47] rigorosamente definiti per garantire interoperabilità e correttezza funzionale, così il diritto si fonda su protocolli procedurali e sostanziali.

L'idea di "protocollo" condivide in entrambi i campi la nozione di un insieme strutturato di regole e convenzioni che governano l'interazione e il comportamento degli attori (siano essi sistemi software o soggetti giuridici) per il raggiungimento di un fine specifico. Trovare un linguaggio e metodologie comuni, che traducano i requisiti giuridici – come il diritto al contraddittorio o il principio di

[46] Ad esempio i protocolli TCP/IP per la trasmissione dei dati in rete.
[47] Come le specifiche di un formato di file o le interfacce di programmazione delle applicazioni – API.

proporzionalità – in specifiche tecniche verificabili e implementabili nei sistemi IA, e, viceversa, che rendano comprensibili ai giuristi le logiche e i limiti degli algoritmi, è un passo fondamentale.

Questo potrebbe tradursi nello sviluppo di "protocolli di conformità costituzionale" per i sistemi IA, ovvero insiemi di requisiti tecnici e procedurali che un sistema deve soddisfare per essere considerato rispettoso dei principi fondamentali, facilitando così il dialogo interdisciplinare e la verifica *ex ante* ed *ex post* della sua legittimità.

4.2. Accorgimenti normativi e regolatori

Il quadro normativo esistente, incluso il Regolamento europeo sull'IA (Regolamento (UE) 2024/167), fornisce una base importante, ma appare ad oggi più un monito che una direzione. È opportuno definire requisiti più stringenti e dettagliati per l'IA impiegata dalla PA, considerando la sua diretta incidenza sui diritti e sui principi pubblicistici.

Diviene cruciale l'introduzione di Valutazioni d'Impatto Algoritmico (AIA) obbligatorie, focalizzate non solo sui rischi tecnici, ma specificamente sull'impatto sui diritti fondamentali e sui principi costituzionali, prima dell'adozione e durante l'intero ciclo di vita dei sistemi IA. Tali valutazioni dovrebbero essere pubbliche e soggette a consultazione.[48]

Parallelamente, la vigilanza continua, l'*audit* indipendente e, per i sistemi a più alto rischio, meccanismi di certificazione basati su standard robusti e condivisi, non possono essere episodici, ma devono configurarsi come un processo dinamico di scrutinio, affidato a organismi dotati di elevate e comprovate competenze sia tecniche che giuridiche, capaci di comprendere l'evoluzione dei sistemi e dei rischi associati. [49]

Le procedure di approvvigionamento pubblico rappresentano un ulteriore snodo cruciale: esse devono evolvere da una logica di mero acquisto di tecnologia a un processo di co-progettazione di soluzioni che incorporino *ab origine* i valori pubblici. Trasparenza, equità, robustezza e verificabilità devono diventare criteri di aggiudicazione preponderanti, capaci di orientare il mercato verso un'offerta di IA realmente "etica by design".[50]

Infine, per sostanziare la trasparenza attiva e abilitare una forma di "sovranità tecnologica" partecipata, è necessario considerare seriamente l'implementazione di piattaforme pubbliche per l'accesso controllato al codice sorgente dei sistemi sviluppati o commissionati da enti pubblici affiancate da sistemi formalizzati e certificati per la segnalazione e la gestione collaborativa di vulnerabilità, *bias* e proposte di miglioramento[51]. Questi meccanismi, ispirati alla logica delle *pull*

---

[48] AlgorithmWatch è una ONG che ha svolto un ruolo importante nel promuovere la necessità di valutazioni d'impatto algoritmico, proponendo anche framework e metodologie per la loro conduzione, con un focus sui rischi per i diritti fondamentali e la democrazia. Vedi AlgorithmWatch. (2020). *Automating Society Report 2020* (e lavori successivi).

[49] Sono in lavorazione standard ISO/IEC relativi all'IA, inclusi quelli sulla gestione del rischio (ISO/IEC 23894), sulla *trustworthiness* (ISO/IEC TR 24028), e *framework* per la valutazione della conformità e l'audit dei sistemi IA (es. ISO/IEC 42001 sui sistemi di gestione dell'IA).

[50] Significativo in questo senso il contributo della Commissione Europea. (2021). *Ethics guidelines for trustworthy AI*, documento del Gruppo di esperti di Alto Livello sull'IA.

[51] Magari non dissimili da popolari *software* di *version control* come *Git*,

request ma calati in un contesto di garanzie pubblicistiche, potrebbero trasformare il controllo da mero esercizio ispettivo a dialogo costruttivo e continuo.

4.3. Strumenti tecnici e organizzativi

La mera elencazione di strumenti tecnici e organizzativi, per quanto necessari, rischia di rimanere sterile se non inserita in una visione strategica che ne potenzi le sinergie e ne orienti l'applicazione verso la piena compatibilità costituzionale. Non si tratta di adottare singole soluzioni, ma di coltivare un ecosistema in cui tecnologia, competenze e partecipazione civica concorrano a un fine comune.

La promozione della ricerca su tecniche di Explainable AI (XAI), ad esempio, deve andare oltre la semplice richiesta di "affidabilità e comprensibilità". È necessario che tali tecniche siano co-progettate con giuristi ed esperti di dominio per assicurare che le "spiegazioni" prodotte siano non solo tecnicamente accurate, ma anche giuridicamente significative e processualmente utilizzabili.

Analogamente, le strategie di *data governance* pubblica non possono limitarsi a garantire la conformità formale al GDPR. Devono incarnare un principio di responsabilità proattiva nella gestione del dato pubblico come bene infrastrutturale comune.

L'investimento nella formazione e nell'aggiornamento delle competenze all'interno della PA, pur cruciale, deve superare la logica della mera acquisizione di "figure ibride". È necessario un cambiamento culturale sistemico che porti tutti i livelli dell'amministrazione a sviluppare una coscienza critica sull'impiego degli algoritmi[52]. Non si tratta solo di "dialogare con i tecnici", ma di dotare i funzionari pubblici degli strumenti concettuali per essere utenti consapevoli, committenti esigenti e garanti dei principi fondamentali nell'interazione con i sistemi di IA. Questo include la capacità di identificare i rischi, di valutare l'adeguatezza delle soluzioni proposte e di pretendere garanzie di trasparenza e *accountability*.

Infine, il coinvolgimento pubblico e degli *stakeholder* non può ridursi a consultazioni estemporanee. È necessario sperimentare forme più strutturate e istituzionalizzate di deliberazione pubblica e di co-progettazione, specialmente per i sistemi IA ad alto rischio. Meccanismi come i *citizen jury*, i *panel* di consenso o le piattaforme di democrazia partecipativa digitale potrebbero essere impiegati per integrare le preferenze collettive e i valori sociali nella definizione degli obiettivi e dei vincoli dei sistemi algoritmici, trasformando la partecipazione da adempimento formale a reale strumento di legittimazione democratica e allineamento valoriale.

Solo così l'IA potrà essere percepita non come un'imposizione tecnocratica, ma come uno strumento effettivamente al servizio della collettività.

Il percorso sembra esigente ma ineludibile. Richiede visione, impegno e vigilanza critica costante, non per frenare l'innovazione, ma per plasmarla al servizio dei principi fondamentali della nostra convivenza civile.

---

[52] Una Direttiva del Ministro per la Pubblica Amministrazione (Italia). (2025, 14 gennaio). *Direttiva sulla formazione e valorizzazione del personale delle pubbliche amministrazioni*, sottolinea la necessità di promuovere competenze tecniche, trasversali e umanistiche per uno sviluppo etico ed efficace dell'IA, e introduce il concetto di "AI literacy" per diffondere la conoscenza di base sull'IA nella PA.

L'ambizione dovrebbe rimanere quella di trasformare la potenza dell'IA in effettivo strumento di progresso, giustizia e democrazia con la Costituzione come irrinunciabile guida.