Limiti e opportunità dell'AI Act: due spunti di riflessione in tema di definizioni e approccio *by design*[1]

Salvatore Sapienza
Dipartimento di Scienze Giuridiche, CIRSFID-ALMA AI Università di Bologna. Mail: salvatore.sapienza@unibo.it

Monica Palmirani
Dipartimento di Scienze Giuridiche, CIRSFID-ALMA AI Università di Bologna. Mail monica.palmirani@unibo.it

1. Cenni introduttivi

Attraverso due esemplificazioni, una dedicata alla definizione di "sistema di IA", una dedicata al rapporto tra *design* dei sistemi e ristoro, il seguente studio valuta criticamente opportunità e limiti dell'approccio normativo del Regolamento 2024/1689 ("AI Act"). La strategia regolatoria consiste non tanto nel tentativo di "sanare" gli effetti di un'innovazione già in atto come l'IA, ma di guidarla fin dall'inizio, creando tecnologie che siano intrinsecamente responsabili, allineate con i valori della società, rispettose di diritti e principi fondamentali. Tale strategia normativa, generalmente indicata come New Legislative Framework (NLF), affianca alcune definizioni ad obblighi generici (integrati da standard tecnici), meccanismi di accertamento della *compliance* e obblighi di notifica, lasciando alle autorità competenti il compito di monitorare il mantenimento della conformità attraverso procedure di controllo e *audit*.

Il NLF è ampiamente utilizzato nella regolamentazione di prodotti che presentano rischi intrinseci alla loro conformazione o al loro settore di utilizzo, in genere verso salute o diretti a categorie di soggetti particolarmente vulnerabili. La difficoltà di regolamentare l'IA secondo questo schema risiede a) nell'intrinseca mutevolezza di una tecnologia *in fieri* come l'IA, b) nella portata molto ampia di diritti e libertà fondamentali potenzialmente lesi da un sistema IA e c) dall'utilizzo pervasivo di sistemi automatizzati in contesti ampi e fortemente eterogenei
I due casi di studio presentati di seguito (sez. 2 e 3) evidenziano possibili criticità in due aree vitali, ossia la definizione dell'oggetto di normazione e la predisposizione di meccanismi di ristoro. La sezione conclusiva riassume l'elaborato e identifica potenziali soluzioni.

2. "Inferire" o "Dedurre"? Il nodo semantico nella definizione di "sistema di IA"

La definizione di sistema di IA – ossia *un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output* (Art. 3(1)) – presenta alcuni elementi degni di menzione. Da un lato, il sistema di IA dovrebbe essere funzionalmente autonomo, ossia non necessitare di un *input* umano costante nello svolgimento del compito assegnato; dall'altro, il sistema potrebbe presentare forme di adattabilità strutturale, ossia la capacità di modificare, anche in tempo reale, la sua architettura in risposta a nuovi dati o ambienti di *deployment*. Entrambe le caratteristiche - autonomia funzionale e adattabilità strutturale -

[1] Questo contributo è finanziato dal Grant Erasmus+ Jean Monnet Legal Design and Data Science For Explicable AI in Legal Domain (LEDS 4 XAIL) n. 101085576.

certamente attengono allo stato dell'arte dei moderni sistemi di IA e, pertanto, la loro inclusione nella definizione appare coerente con lo scopo oggettivo dell'AI Act identificato dal legislatore europeo.

Preferendo all'interpretazione teleologica quella letterale, un elemento critico di ambiguità emerge dall'impiego di un verbo atto a descrivere le capacità di un sistema di IA che compare nella definizione italiana. È interessante notare una discrepanza tra la versione italiana e quella inglese del testo normativo. Mentre per la prima il sistema di IA *deduce* […] *come generare output*, per la versione inglese il sistema *infers* […] *how to generate outputs*. Vale la pena evidenziare che la distinzione tra "inferire" e "dedurre", sebbene sottile nel linguaggio comune, assume una rilevanza cruciale nell'interpretazione dell'Act. In estrema sintesi, *inferire* è il processo cognitivo e logico più ampio attraverso il quale si giunge a una conclusione partendo da un insieme di premesse o evidenze, mentre *dedurre* si riferisce a una specifica modalità di inferenza, quella deduttiva, caratterizzata da conclusioni logicamente necessarie, e contrapposta a induzione e abduzione.

L'uso del verbo "dedurre" nella versione italiana potrebbe, in un'interpretazione restrittiva, limitare la definizione di sistema di IA a quei sistemi che operano attraverso processi logico-deduttivi, dove le conclusioni discendono con certezza dalle premesse/*input*. Questo tipo di ragionamento è tipico dei sistemi esperti basati su regole (*rule-based expert systems*), la forma più tradizionale di IA che non pone gli stessi rischi per diritti e libertà fondamentali, non essendo, in genere, dotata di ampi livelli di autonomia funzionale e, soprattutto, mancando di adattabilità a nuovi dati e ambienti di *deployment*. La stessa scelta è stata adottata nelle versioni francese ("déduit") e, parzialmente, tedesca ("ableitet")

Al contrario, l'uso del verbo "infers" nella versione inglese - presente anche in quella spagnola ("infiere") e portoghese ("infere") – è significativamente più inclusivo. Il termine "inferenza" abbraccia non solo la deduzione, ma anche l'induzione e l'abduzione, che sono alla base del funzionamento della maggior parte dei moderni sistemi di *machine learning* e *deep learning*. Questi sistemi, difatti, non "deducono" in senso stretto, ma "inferiscono" modelli, correlazioni e previsioni a partire da grandi quantità di dati.

L'uso di "inferire" allinea la definizione normativa allo stato dell'arte delle tecnologie di IA, che è prevalentemente orientato a modelli induttivi e abduttivi (specie in quelli più avanzati) e certamente probabilistici. Una lettura letterale di "deduce" potrebbe invece escludere, o creare ambiguità, sull'inclusione di molti sistemi di *machine learning*. Il termine "deduzione" implica un grado di certezza e spiegabilità dell'output che raramente si riscontra nei modelli di IA più complessi (c.d. "black box"). L'inferenza, invece, accoglie la natura non deterministica degli *output* generati da tali sistemi da cui discendono le caratteristiche di autonomia funzionale e adattabilità strutturale che, in ultima *ratio*, costituiscono il presupposto normativo delle criticità che spingono alla regolamentazione dell'IA.

3. Regolamentazione *by design* e meccanismi di ristoro

La regolamentazione *by design* rappresenta uno degli elementi centrali dell'approccio europeo alla regolamentazione dell'IA e, in generale, alla normazione sulle tecnologie del digitale. Non si tratta di una novità assoluta (era già presente nel GDPR), ma assume un ruolo fondamentale nell'AI Act per garantire che i requisiti di sicurezza, controllabilità, spiegabilità, trasparenza, *fairness*, tutela dei dati personali, ecc. siano integrati direttamente nel tessuto dei sistemi di IA,

specie quelli ad alto rischio, sia dalle primissime fasi di sviluppo, che per l'intero ciclo di vita del sistema. I requisiti includono misure di *data governance* e qualità del dato, trasparenza (attraverso la documentazione), misure di sicurezza e resilienza, supervisione umana (collegata alla spiegabilità dell'*output* dei sistemi), misure di privacy e protezione dei dati personali.

L'Act non si limita a imporre obblighi *ex ante*, ma include anche disposizioni *ex post* per affrontare i rischi che possono emergere dopo che un sistema è stato immesso sul mercato. Tali disposizioni sono particolarmente pertinenti per i sistemi di IA che si evolvono e apprendono continuamente dopo la messa in servizio, e che rendono, pertanto, difficile prevedere tutti i rischi al momento dello sviluppo. L'articolo 72 del Regolamento impone ai fornitori di sistemi di IA ad alto rischio di raccogliere e revisionare periodicamente le esperienze acquisite dall'uso dei loro sistemi, documentando eventuali rischi emergenti nelle fasi di *deployment*. Le misure di monitoraggio creano un ciclo di *feedback* in cui le lezioni apprese dalla pratica consentono di apportare miglioramenti tecnici al sistema, rendendo l'approccio *by design* un processo continuo e non soltanto una misura di prevenzione iniziale.

L'iniziale strategia dell'UE per la responsabilità civile in materia di IA era basata su un approccio complementare. A supporto degli obiettivi di tutela posti dall'AI Act, la Commissione aveva proposto una Direttiva sulla responsabilità per l'IA (AILD). Le caratteristiche intrinseche dell'IA, come l'opacità dei suoi algoritmi, la complessità e la continua adattabilità, nonché un esteso ciclo di vita, rendono difficile o, in taluni casi, impossibile per le vittime identificare e provare la colpa di una parte potenzialmente responsabile o il nesso di causalità tra tale colpa e il danno subito. L'obiettivo dell'AILD era quello di colmare le difficoltà nell'attribuzione di responsabilità per i danni causati dai sistemi di IA. Essa sarebbe andata incontro a questa esigenza assicurando meccanismi presuntivi a beneficio degli utenti di prodotti di IA e facilitando, quindi, l'attivazione di meccanismi di ristoro a loro beneficio.

Il ritiro dell'AILD è stato deciso a causa della prevedibile mancanza di un accordo politico sulla normativa. Al di là delle ragionevoli critiche al ritiro della proposta, la revoca dell'AILD, pur mirando a semplificare il quadro normativo e ridurre gli oneri per i fornitori di sistemi e prodotti di IA, rischia di creare una lacuna cruciale nella strategia legislativa complessiva dell'Unione. L'AI Act, a causa della sua vocazione preventiva e *by design*, non prevede disposizioni esplicite sulla responsabilità per le richieste di risarcimento danni connessi a prodotti, fatto salvo il diritto di presentare un reclamo a un'autorità di vigilanza del mercato.

4. La necessità di categorie concettuali appropriate

La sfida di normare i fenomeni tecnologici in rapida evoluzione non risiede nella generalmente menzionata (e spesso presunta) incapacità del diritto di adattarsi alle nuove tecnologie, ma nella necessità di aggiornare le sue categorie concettuali. I due temi discussi in questo breve scritto – le difficoltà definitorie e l'approccio *by design* – costituiscono due esempi di un cambiamento di paradigma che porta con sé benefici e limitazioni tipiche di ogni transizione. Tra i primi, va sicuramente annoverata l'audacia del primo tentativo organico di disciplinare l'IA a livello mondiale e la corretta identificazione della tensione tra due obiettivi strategici, ossia la tutela dei diritti fondamentali e la promozione di un mercato unico digitale competitivo. Due obiettivi che, in genere, vengono posti in antitesi ma che, quanto meno nelle premesse dell'Act, trovano una sintesi nella normazione *by design*. Allo

stesso tempo, tuttavia, la difficoltà di cristallizzare fenomeni tecnologici in definizioni normative e la difficoltà di coniugare l'ancoraggio allo stato dell'arte con gli strumenti canonici della normatività, che includono la possibilità di ottenere un ristoro, spingono a ritenere che la fase di transizione non sia del tutto matura. Potrebbe, forse, essere necessario realizzare una miglior sintesi informatico-giuridica per rendere operativo questo modello normativo. Le categorie concettuali da essa proposte superano il divario tra il testo normativo e gli elementi informatici, rendendo la regolamentazione dell'IA non un freno, ma un catalizzatore per un'innovazione responsabile ed efficace.