GPAI e AI Act: tra innovazione regolatoria e fragilità applicative

Giovanni De Gregorio
Chair in Law and Technology, Católica Global School of Law, Universidade Católica Portuguesa. Mail: gdegregorio@ucp.pt

Giuseppe Muto
PhD Student in Legal Studies nell'Università Commerciale «LuigiBocconi». Mail: giuseppe.muto@unibocconi.it

Giovanni Sartor
Professore ordinario, Dipartimento di Scienze Giuridiche, Università di Bologna. Mail: giovanni.sartor@unibo.it

1. Rischi ed evoluzione della GPAI

L'evoluzione dei sistemi di intelligenza artificiale *general-purpose* (GPAI), basati su modelli fondazionali (enormi reti neurali preaddestrate), pone in evidenza un duplice profilo. Da un lato, tali sistemi sono capaci di operare trasversalmente, con prestazioni elevate, in domini diversi, dall'analisi del linguaggio umano, alla programmazione, alla generazione multimodale, fino a svolgere compiti specialistici in ambiti scientifici e professionali. Dall'altro, gli stessi rimangono inaffidabili, poiché spesso producono risultati incoerenti, errori inattesi, contenuti falsi (allucinazioni), illegali o viziati da pregiudizi. Benché diverse soluzioni tecnologiche siano state adottate per ridurre questi inconvenienti, l'opacità del funzionamento interno dei sistemi fondazionali impedisce di individuare in anticipo i loro difetti e di porvi rimedio in modo sistematico. In tale contesto, i principali rischi derivanti dalle GPAI possono essere distinti in quattro principali categorie: (i) i rischi da uso malevolo, che includono la produzione di contenuti sintetici a fini di disinformazione, frode o attacco cibernetico; (ii) i rischi da malfunzionamento, che concernono la generazione di contenuti erronei in contesti critici e le distorsioni derivanti da dati squilibrati; (iii) i rischi sistemici, che investono il tessuto socio-economico nel suo complesso, dalle trasformazioni del mercato del lavoro, alla concentrazione oligopolistica della capacità di sviluppo, dall'ampliamento del divario digitale globale, ai costi ambientali e alle implicazioni per privacy e diritto d'autore; (iv) infine, il rischio da perdita di controllo del sistema di IA, con particolare riguardo agli agenti autonomi, che, combinando le proprie capacità di pianificazione con le risorse cognitive fornite dai modelli fondazionali, possono adottare comportamenti che vanno contro gli interessi dei loro utilizzatori e della stessa comunità.

Le strategie di mitigazione tecnica dei rischi derivanti dalle GPAI finora sviluppate si articolano in diversi approcci di gestione del rischio e ingegneria della sicurezza, ai fini della costruzione di modelli più robusti e allineati alle intenzioni dei progettisti. Accanto alla prevenzione o, almeno, alla riduzione della generazione di contenuti illegali, iniqui o socialmente pericolosi, si pone la tutela della privacy lungo l'intero ciclo di vita dei modelli e la rimozione mirata di strumenti ad alto rischio. A tali pratiche si affiancano strumenti di monitoraggio *post-deployment*, destinati a rilevare anomalie e tentativi di attacco e a predisporre garanzie (purtroppo con risultati limitati) contro i rischi prospettati. Tuttavia, tali strategie non appaiono pienamente soddisfacenti. Parimenti, le necessarie metodologie di valutazione e *auditing* —dai *benchmark* (parametri di riferimento) standardizzati all'effettuazione di attacchi informatici simulati (*red-teaming),* fino agli *audit* indipendenti su dati e gli *output*— si rivelano strutturalmente incomplete, sia per la difficoltà di accesso ai dati e ai modelli, sia per

l'insufficienza di approcci interdisciplinari nella stima degli impatti sociotecnici *downstream*.

2. La GPAI: profili definitori e la regolazione dell'AI Act

Questa cornice di rischi, mitigazioni e limiti metodologici fornisce il presupposto per comprendere l'intervento regolatorio dell'Unione europea. L'Artificial Intelligence Act (AIA), infatti, si colloca precisamente all'intersezione tra la rapida crescita delle capacità delle GPAI e l'urgenza di predisporre un quadro giuridico capace di governarne i rischi, con particolare attenzione ai modelli suscettibili di produrre effetti sistemici.

L'art. 3(63) AIA definisce *general-purpose AI model* un modello di AI **«**addestrato con grandi quantità di dati utilizzando l'autosupervisione su larga scala, che sia caratterizzato da una generalità significativa e sia in grado di svolgere con competenza un'ampia gamma di compiti distinti, indipendentemente dalle modalità con cui il modello è immesso sul mercato, e che può essere integrato in una varietà di sistemi o applicazioni a valle**»**.

In relazione ai sistemi di GPAI, l'AI Act introduce un quadro di obblighi specifici che riflette la crescente consapevolezza del loro impatto trasversale sull'ordinamento giuridico e sul tessuto sociale. I fornitori sono tenuti, in primo luogo, a predisporre e mantenere aggiornata la documentazione tecnica del modello, indicando le fasi di addestramento e di test, nonché i risultati delle valutazioni secondo i criteri fissati dall'allegato XI (art. 53(1)(a) AIA). In secondo luogo, gli stessi debbono implementare politiche idonee a garantire il rispetto della normativa unionale in materia di diritto d'autore e diritti connessi (art. 53(1)(c) AIA). Tali obblighi non si estendono ai modelli rilasciati in modalità *open source*, purché i relativi parametri siano resi pubblici (con l'eccezione dei modelli generativi suscettibili di produrre un rischio sistemico) (art. 53(2) AIA). Per i fornitori stabiliti fuori dall'Unione, l'immissione sul mercato europeo è subordinata alla nomina di un rappresentante autorizzato, incaricato di fungere da interlocutore con le autorità europee e nazionali e di garantire la conservazione e la disponibilità della documentazione per un periodo decennale (art. 54(1) AIA).

Particolarmente stringente è la disciplina dei modelli qualificati come "a rischio sistemico", in ragione delle loro capacità ad alto impatto: in tali casi, ai fornitori si impone non solo la notifica alla Commissione europea entro due settimane dal superamento delle soglie di calcolo previste, ma anche l'effettuazione di test standardizzati (incluso l'*adversarial testing,* cioè l'esposizione del sistema ad attacchi o, comunque, a messaggi ingannevoli intesi a evidenziarne le debolezze) (art. 52(1) AIA e art. 55(1) AIA), l'adozione di misure di mitigazione dei rischi a livello unionale (art. 52(1)(b) AIA), la comunicazione tempestiva di incidenti gravi (art. 55(1)(c) AIA) e la predisposizione di adeguate garanzie di cybersicurezza (art. 55(1)(d) AIA).

Accanto agli obblighi diretti, l'AI Act prevede strumenti di co-regolazione tesi a rafforzare la robustezza dei sistemi e a prevenire e mitigare i rischi, in particolare attraverso i codici di condotta (art. 56 AIA) e gli standard armonizzati (art. 40 AIA). I primi, elaborati con il coinvolgimento di attori pubblici e privati, consentono ai fornitori di GPAI, inclusi quelli a rischio sistemico, di basarsi sugli stessi al fine di dimostrare la conformità ai requisiti normativi (fino alla pubblicazione delle norme armonizzate).

Questi ultimi, invece, traducono i principi generali del regolamento in specifiche tecniche (*testing*, validazione, gestione dei dati), che sembrerebbero offrire ai fornitori una presunzione di conformità agli obblighi specifici di cui all'art. 53(1) AIA. Se da un lato tali strumenti assicurano

un adeguamento dinamico alla rapida evoluzione tecnologica, dall'altro sollevano criticità legate alla prevalenza degli interessi industriali nei processi di definizione, che la Commissione cerca di bilanciare attraverso consultazioni pubbliche e meccanismi di vigilanza.

3. Questioni irrisolte e sfide

L'innovatività del quadro normativo fin qui descritto non può oscurarne i limiti. Ad esempio, si può citare la scelta dell'art. 51(2) AIA di identificare i GPAI "a rischio sistemico" sulla base di una soglia computazionale (10^25 FLOPs), giustamente criticata per la sua riduttività. Infatti, un parametro numerico, ancorato al volume di operazioni di calcolo impiegate per l'addestramento, rischia di tradursi in un criterio arbitrario e rapidamente obsoleto, incapace di cogliere fattori altrettanto rilevanti, quali l'architettura del modello, la qualità dei *dataset*, il contesto applicativo o le dinamiche di interazione fra più sistemi. La stessa Commissione è consapevole della necessità di poter modificare tale soglia mediante atti delegati, ma ciò conferma l'instabilità di un approccio fondato su un indice tecnico unico, che rischia tanto di sottovalutare modelli di minore dimensione ma comunque pericolosi, quanto di sovrastimare rischi che non si concretizzano in pratica. Il problema, dunque, non è solo tecnico, ma di metodo regolatorio: un criterio di classificazione così rigido può minare la prevedibilità e l'effettività della disciplina.

Sul versante dell'*enforcement*, invece, il quadro delineato dall'AI Act offre un evidente disallineamento. Da un lato, vige un modello decentrato, fondato sull'art. 70 AIA, che attribuisce ampi poteri investigativi e sanzionatori alle autorità nazionali competenti per i sistemi vietati ad alto rischio e a rischio limitato. A ciascuno Stato membro spetta, altresì, la definizione delle sanzioni, purché "efficaci, proporzionate e dissuasive", con il rischio concreto di generare fenomeni di *forum shopping* e disomogeneità applicative (art. 99(1) AIA). Dall'altro lato, i GPAI a rischio sistemico sono collocati in un modello centralizzato: il controllo su di essi compete all'AI Office istituito presso la Commissione (DG CNECT) dotato di compiti di vigilanza, valutazione dei modelli, indagine sulle violazioni e sviluppo di strumenti di *soft law* (art. 64 AIA), affiancato dall'*European Artificial Intelligence Board*, organismo di coordinamento tra le autorità nazionali ma privo di poteri vincolanti. Questo assetto riflette l'esigenza di un controllo forte e unitario per un numero ristretto di attori globali, ma al tempo stesso rischia di produrre un dualismo istituzionale non privo di tensioni, soprattutto nella definizione delle rispettive competenze e nella disponibilità di risorse adeguate a garantire l'effettività dei controlli.

Un ulteriore nodo critico riguarda il coordinamento del regime dei GPAI con le altre fonti del diritto unionale e nazionale. L'AI Act non opera in un *vacuum*, ma si innesta su normative già consolidate, come il GDPR e il diritto d'autore. In materia di protezione dei dati, il caso della sospensione di ChatGPT da parte del Garante italiano ha mostrato con chiarezza le difficoltà di conciliare l'addestramento dei modelli su vasti *dataset* con i principi di liceità, correttezza e trasparenza del trattamento. Sul versante del *copyright*, la sentenza *Kneschke c. LAION* (Corte Regionale di Amburgo, 27 settembre 2024) ha ribadito come l'uso di *dataset* protetti possa trovare legittimazione solo entro i limiti dell'eccezione di *text and data mining* per scopi di ricerca scientifica non commerciale, portando alla luce l'inevitabile intreccio fra AI Act e disciplina nazionale del diritto d'autore. A ciò si aggiunge il rapporto con il Digital Services Act, che disciplina i rischi sistemici delle grandi piattaforme online (VLOP e VLOSE), ma lascia irrisolto il nodo dei GPAI in

quanto tali. Si profila, così, un rischio di sovrapposizione normativa: da un lato, l'AI Act presume il rispetto delle norme del DSA per la gestione dei rischi di disinformazione; dall'altro, esso individua rischi specifici dei GPAI non coperti dal DSA, con conseguente incertezza applicativa e possibile duplicazione di obblighi.

Questi profili rendono evidente come la disciplina dei GPAI sia insieme ambiziosa e fragile. Ambiziosa, perché per la prima volta l'Unione europea ha tentato di costruire un quadro orizzontale di regole su misura per i modelli fondazionali, riconoscendone la centralità nell'ecosistema digitale e i rischi trasversali che essi comportano. Fragile, perché fondata su criteri tecnici discutibili, su un sistema di *enforcement* complesso e frammentato e su un intreccio con altre fonti normative che rischia di produrre conflitti e incertezza.

Una valutazione critica della disciplina dei GPAI non può dunque che condurre a un duplice ordine di considerazioni: da un lato, l'AI Act costituisce un progresso normativo senza precedenti, che eleva la disciplina dei GPAI a questione di interesse costituzionale europeo; dall'altro, il successo della sua applicazione risulta legato all'effettività dei meccanismi di *enforcement* e alla capacità delle istituzioni europee di coordinare questo nuovo strumento con il più ampio quadro delle garanzie sovranazionali.