AIC
Associazione Italiana
dei Costituzionalisti

Osservatorio AIC

Bimestrale di attualità costituzionale

Anno 2023/ Fascicolo V

L'Associazione Italiana dei Costituzionalisti è iscritta al Registro degli Operatori della Comunicazione a far data dal 09.10.2013 con n. 23897.

La rivista Osservatorio costituzionale, inclusa tra le riviste scientifiche dell'Area 12 - Scienze giuridiche, dal Fascicolo 1/2016, è ivi registrata ai sensi dell'art. 16 della legge n. 62 del 2001.

Codice ISSN: 2283-7515.

Le Linee Guida per la pubblicazione sulla rivista sono reperibili al sito www.osservatorioaic.it.

OSSERVATORIO COSTITUZIONALE

Indice del Fascicolo 5 del 2023

Attualità

L. Gori, *I paradossi della democraticità interna ai partiti politici. Le c.d. elezioni primarie del Partito democratico*.....5

M. Calamo Specchia, *L'autonomia differenziata e la solidarietà interterritoriale: spunti di riflessione per una riforma costituzionalmente sostenibile*.....31

S. Sapienza, *ChatGPT, dati e identità personale: rischi, bilanciamenti, regolazione*.....47

D. Lanfranco, *Riorganizzazione del MEF: una riforma silenziosa*.....72

Giurisprudenza

A. Gerardi Virgili, *Stato psicofisico e cosciente partecipazione processuale: il nuovo alveo dell'art. 72 bis c.p.p. Nota a Corte Cost., sent., 23/02/2023 (dep. 7/04/2023), n. 65*.....97

R. Pinardi, *Una pronuncia a rime "possibili", ma anche "parziali". Nota alla sent. n. 40 del 2023 della Corte costituzionale*.....119

C. Forte, M. Pieroni, *Quantificazione e copertura degli oneri continuativi correnti delle leggi di spesa: recenti orientamenti della Corte costituzionale di cui alle sentenze nn. 48, 57, 82, 84 e 110 del 2023*.....134

C. Napoli, *Accesso alle cariche elettive in condizioni di uguaglianza e disuguaglianze di fatto. Riflessioni a margine della sentenza n. 60/2023 della Corte costituzionale*.....160

C. Gigante, *La Corte costituzionale in materia di libertà vigilata conseguente a liberazione condizionale. Nota a Corte cost., sent. n. 66 del 2023*.....183

E. Di Carpegna Brivio, *Le vie costituzionali della solidarietà comunale. Nota a Corte cost. 71/2023*.....198

D. Loprieno, *Sulla erosione dell'automatismo espulsivo in materia di rinnovo del permesso di soggiorno. Nota a sentenza n. 88/2023 della Corte costituzionale*.....214

G. Di Genio, *Beni staggiti e beni collettivi nella "dottrina notarile" offerta alla Corte costituzionale*.....236

A. Rosanò, *L La sentenza Inspecţia Judiciară della Corte di giustizia dell'Unione europea: considerazioni alla luce della crisi dello Stato di diritto in Romania e nell'Unione europea*.....245

C.A. d'Alessandro, *La pronuncia del Conseil constitutionnel a conclusione della "prova da sforzo" del sistema costituzionale francese per l'approvazione della recente riforma delle pensioni*.....267

AIC
Associazione Italiana dei Costituzionalisti
Osservatorio costituzionale

OSSERVATORIO COSTITUZIONALE
Codice ISSN: 2283-7515
Fasc. 5/2023
Data: 3 ottobre 2023

ChatGPT, dati e identità personale: rischi, bilanciamenti, regolazione[*]

di **Salvatore Sapienza** – *Ricercatore a Tempo Determinato in Informatica Giuridica presso il Dipartimento di Scienze Giuridiche dell'Alma Mater Studiorum – Università di Bologna*

TITLE: ChatGPT, Data and Personal Identity: risks, balances, regulation

ABSTRACT: Il saggio affronta le criticità legate all'utilizzo e allo sviluppo su vasta scala di Large Language Model generativi addestrati anche su dati personali. L'articolo mette in risalto le sfide giuridiche legate alla natura dei dati personali utilizzati per l'addestramento di questi e ai possibili errori probabilistici dei modelli che interferiscono con l'identità personale degli individui rappresentati nei loro *output*. Alla discussione di alcune premesse tecniche, segue l'individuazione delle criticità giuridiche che da esse scaturiscono, con particolare riferimento al bilanciamento tra tutela dei dati personali, libertà d'espressione e libertà di iniziativa economica. A queste considerazioni, l'articolo affianca la trattazione del Provvedimento d'urgenza adottato dall'Autorità Garante per la Protezione dei Dati Personali il 30 marzo 2023 nei confronti di OpenAI, a cui è conseguita la limitazione del trattamento dei dati personali da parte del fornitore di ChatGPT. Infine, l'articolo esamina la Proposta regolamentare europea per un Artificial Intelligence Act, che mira a minimizzare i rischi derivanti dall'uso dei sistemi di IA, inclusa quella generativa. Le definizioni riguardanti i "Foundation Model" e i "General Purpose AI System" introdotte dal Parlamento Europeo individuano un regime specifico per questi Large Language Model e le applicazioni basate su esse. L'articolo discute questa Proposta parlamentare, soprattutto con

[*] Lavoro sottoposto a referaggio secondo le linee guida della Rivista.

riguardo all'introduzione di principi etici nel *design* dei modelli di base, di un meccanismo di autovalutazione per garantire la tutela di diritti e libertà fondamentali quando i Large Language Model rischiano di comprometterli, e la definizione di nuovi ruoli e obblighi per i fornitori di modelli di base.

The essay addresses the critical issues related to the use and large-scale development of generative Large Language Models also trained on personal data. The article highlights the legal challenges related to the nature of the personal data used for their training and the possible probabilistic errors of the models that interfere with the personal identity of the individuals represented in their outputs. The discussion of some technical premises is followed by the identification of the legal critical issues that arise from them, with reference to the balance between the protection of personal data, freedom of expression and freedom of economic initiative. In addition to these considerations, the article supports the discussion of the decision issued by the Italian Data Protection Authority on 30 March 2023 against OpenAI, which resulted in the limitation of the processing of personal data by the ChatGPT provider. Finally, the article examines the European regulatory proposal for an Artificial Intelligence Act, which aims to minimize the risks deriving from the use of AI systems, including the generative one. The definitions concerning "Foundation Models" and "General Purpose AI Systems" introduced by the European Parliament identify a specific regime for such Large Language Models and the applications based on them. The article discusses this parliamentary proposal, especially regarding the introduction of ethical principles in the design of base models, a self-assessment mechanism to ensure the protection of fundamental rights and freedoms when large language models risk compromising them, and the definition of new roles and obligations for base model providers.

KEYWORDS: large language model; dati personali; identità personale; bilanciamento; large language models; personal data; personal identity; balancing

SOMMARIO: 1. Introduzione e premesse di metodo. – 2. I Large Language Model e ChatGPT: premesse tecniche essenziali. – 3. Il *training* di ChatGPT tra dati personali, inferenze e allucinazioni. – 4. Il provvedimento dell'Autorità Garante per la Protezione dei Dati Personali del

30 marzo 2023. – 5. Le prospettive regolamentari nell'AI Act: l'intervento del Parlamento Europeo e il ritorno dei principi. – 6. Considerazioni finali.

1. Introduzione e premesse di metodo

I recenti sviluppi sul tema dell'Intelligenza Artificiale (IA) pongono importanti questioni in merito ad un ampio novero di questioni giuridiche[1]. In particolare, l'IA di tipo "generativo" e il crescente utilizzo di modelli linguistici in grado di elaborare e generare informazioni testuali ha destato particolare interesse in ordine ai profili relativi alla tutela dei dati personali, della libertà d'espressione e di quella di iniziativa economica. Questo breve saggio intende illustrare le criticità legate al bilanciamento di questi diritti limitatamente al contesto di utilizzo dei Large Language Model e delle applicazioni basate su di esso, come ChatGPT, traendo spunto dalle vicende giuridiche che hanno visto questo applicativo quale centro d'interesse nei recenti provvedimenti dell'Autorità Garante per la Protezione dei Dati Personali di cui si dirà *infra*. A differenze di altre forme di IA[2], ormai note ai contesti giuridici, questi approcci computazionali sono relativamente nuovi al diritto e, tanto alla luce dell'interesse generato per le loro prestazioni, quanto per le criticità di cui si dirà, necessitano di riflessioni e approfondimenti tematici.

Stante la paucità di fonti dovuta alla relativa novità dei fenomeni tecnologici in questione, perlomeno dal punto di vista della loro diffusione, il metodo seguito nella ricerca bilancia tre elementi essenziali. In primo luogo, pone alcune nozioni tecniche al centro dell'indagine, al fine di identificare correttamente le criticità giuridiche che ad esse sottendono e da esse originano; in secondo luogo, seguendo il metodo dell'informatica giuridica[3], contestualizza l'impatto dei

[1] A. PAJNO, M. BASSINI, G. DE GREGORIO, M. MACCHIA, F. P. PATTI, O. POLLICINO, S. QUATTROCOLO, D. SIMEOLI, P. SIRENA, *AI: profili giuridici. Intelligenza Artificiale: criticità emergenti e sfide per il giurista*, in *BioLaw Journal-Rivista di BioDiritto*, 2019, vol. 3, pp. 205-235.

[2] Sulla definizione di Intelligenza Artificiale, v. J. KAPLAN, *Artificial intelligence: What everyone needs to know*, Oxford University Press, 2016, cap. 1. Osserva l'Autore: «Vi è disaccordo su cosa sia l'intelligenza [...] ci sono scarse ragioni per credere che l'intelligenza delle macchine abbia qualche connessione con l'intelligenza umana, almeno per ora». Per una panoramica sulle definizioni e sugli approcci computazionali, v. il Capitolo I della nota tassonomia in S. RUSSELL, P. NORVIG, *Artificial intelligence: a modern approach*, Pearson Education, 2010. Per una più ampia prospettiva sull'impatto dell'IA in altri contesti, vedi F. COREA, C. G. FERRAUTO, F. FOSSA, A. LOREGGIA, S. QUINTARELLI, S. SAPIENZA, *Intelligenza artificiale. Cos'è davvero, come funziona, che effetti avrà*, Bollati Boringhieri, 2020.

[3] V. l'approccio metodologico in G. SARTOR, *L'informatica giuridica e le tecnologie dell'informazione: corso di informatica giuridica*, Giappichelli Editore, 2016, cap. 1.6.

fenomeni tecnici emergenti nella discussione giuridica e, viceversa, discute in senso tecnologico le proposte normative avanzate; in ultimo, tiene in considerazione l'impatto della dimensione etica, meta-giuridica e del *design* tanto nella fase genetica delle norme, di cui siamo oggi testimoni, quanto in prospettiva ermeneutica[4].

Il saggio è organizzato in 6 sezioni. Facendo seguito a questa introduzione, la sezione 2 identifica alcune premesse tecniche necessarie alla comprensione dei Large Language Model e delle applicazioni che ne fanno uso; la sezione 3 individua le criticità giuridiche che interessano in questa sede, tra cui rientrano la tutela dei dati personali, dell'identità personale, della riservatezza, della libertà di espressione e di quella di iniziativa economica, chiarendo le questioni relative ai bilanciamenti tra interessi contrapposti; la sezione 4 illustra e discute il Provvedimento del Garante Privacy del 30 marzo 2023, che ha tracciato una prima via nel dibattito nazionale ed europeo sui temi del trattamento dei dati personali nello sviluppo e nella messa in circolazione di Large Language Model; la sezione 5, invece, pone l'accento sulle prospettive regolamentari dell'Artificial Intelligence Act, discutendo le previsioni normative applicabili ai Large Language Model a seguito dell'intervento del Parlamento Europeo e alla luce dei principi etici individuati nel contesto eurounitario; la sezione 6 conclude il saggio identificandone i limiti e tracciando le prospettive di ricerca futura.

2. I Large Language Model e ChatGPT: premesse tecniche essenziali

I Large Language Model (LLM) sono una forma di intelligenza artificiale progettata per comprendere e generare testo in modo autonomo. Tra essi rientrano i prodotti della serie GPT- rilasciata da OpenAI. L'approccio computazionale di riferimento dei LLM è il *deep learning*, una tecnica di apprendimento automatico che utilizza reti neurali profonde per analizzare e comprendere i dati. I LLM vengono addestrati su un vasto *corpus* di testi, che può includere libri, articoli, pagine web e altro ancora. Durante l'addestramento, il modello impara progressivamente a riconoscere i

[4] Per un'introduzione ai temi dell'etica dell'IA, v. S. QUINTARELLI, F. COREA, F. FOSSA, A. LOREGGIA, S. SAPIENZA, *AI: profili etici. Una prospettiva etica sull'Intelligenza Artificiale: princìpi, diritti e raccomandazioni*, in *BioLaw Journal-Rivista di BioDiritto*, 2019, vol. 3, pp. 183-204.

modelli e le strutture linguistiche all'interno dei testi al fine di essere in grado di predire la parola che seguirà un certo *input.*

L'*input* e l'*output* del modello sono organizzati sulla base di *token*. Un *token* è una rappresentazione numerica di un concetto. Si considerino le seguenti 3 frasi:

1. Rosso di sera bel tempo si spera.
2. Questa sera bevo vino rosso.
3. Il tempo di domani è incerto e si spera migliori.

Ad ogni termine, inclusa la punteggiatura, viene assegnato un identificativo unico, qui segnalato dal numero tra parentesi quadre che precede il vocabolo (o lessema).

1. [1] Rosso [2] di [3] sera [4] bel [5] tempo [6] si [7] spera [8].
2. [9] Questa [3] sera [10] bevo [11] vino [1] rosso [8].
3. [12] Il [5] tempo [2] di [13] domani [14] è [15] incerto [16] e [6] si [7] spera [17] migliori [8].

Ogni frase può quindi essere rappresentata secondo una matrice:

1. [1, 2, 3, 4, 5, 6, 7, 8]
2. [9, 3, 10, 11, 1, 8]
3. [12, 5, 2, 13, 14, 15, 16, 6, 7, 17, 8]

Il numero di *token* su cui è addestrato GPT 3.5[5], ossia il LLM su cui si basa l'attuale versione di ChatGPT, è di circa 500 miliardi, equivalenti ad altrettante parole, nelle più svariate lingue, elaborati da una rete neurale consistente in una struttura di 175 miliardi di parametri e 96 strati, o *layer*[6]. Per giungere alla sua forma attuale in ChatGPT, GPT 3.5 è stato soggetto ad un processo di *fine-tuning* ibrido – supervisionato e con rinforzo – e composto da una fase di addestramento e di ottimizzazione[7]. In primis, gli annotatori umani hanno assunto il ruolo di *chatbot* conversazionale e

[5] GPT è l'acronimo di per Generative Pre-Trained Transformer. I modelli Transformer permettono di catturare relazioni semantiche e contestuali tra le parole di una frase, evitando i problemi di *vanishing* ed *exploding gradient* delle reti ricorsive. Sono particolarmente noti per il modello "Transformer" originale utilizzato in BERT (Bidirectional Encoder Representations from Transformers), che ha ottenuto risultati pari allo stato dell'arte in molte attività di NLP, come il riconoscimento del linguaggio naturale, l'elaborazione del linguaggio naturale e la traduzione automatica.

[6] Vedi il lavoro introduttivo in T. BROWN, B. MANN, N. RYDER, M. SUBBIAH, J. D. KAPLAN, P. DHARIWAL, A. NEELAKANTAN, *Language models are few-shot learners*, in *Advances in neural information processing systems*, 2020, n. 33, pp. 1877-1901.

[7] V. le note riportate da OpenAI disponibili all'indirizzo https://openai.com/blog/chatgpt. L'apprendimento supervisionato prevede che i dati di addestramento siano "etichettati" da esseri umani, in modo che il sistema

illustrato al sistema l'*output* desiderato[8]; i validatori hanno, in seguito, classificato le risposte generate dal modello e, sulla base di queste indicazioni, è stato ottenuto lo schema di assegnazione delle ricompense utilizzato nella seguente fase di ottimizzazione; infine, tramite il meccanismo di *reward*/ricompensa ottenuto dallo step precedente, GPT 3.5 è stato ottimizzato per generare risposte sempre più coerenti con i *desiderata* dei validatori.

In questo modo, il modello ha ottimizzato i propri *output* per renderli verosimili. Sulla base di questa struttura e di questo metodo di addestramento, ChatGPT è capace di generare testo coerente e significativo in risposta a *input* specifici[9]. Utilizzando l'*input* dell'utente, detto *prompt*, il modello è in grado generare una risposta plausibile attraverso la generazione probabilistica di sequenze di testo, in cui il modello cerca di predire il *token* successivo sulla base di quelli precedenti e del contesto che viene offerto. Quando interrogato, ChatGPT, inoltre, dispone di misure di sicurezza e moderazione per contenuti sensibili: tramite il sistema non è possibile, infatti, generare contenuti di incitamento all'odio, razzisti, e così via[10]. Il sistema include alcune meta-istruzioni che sono inserite

automatizzato possa apprendere le regolarità statistiche. Ad esempio, al fine di addestrare in modo supervisionato un sistema di IA a giocare a scacchi è necessario che esseri umani forniscano preliminarmente numerosi esempi di mosse da seguire nel gioco degli scacchi. A seguito dell'addestramento, il sistema sarà in grado, sulla base degli esempi forniti nel *dataset* di addestramento, di individuare la mossa migliore anche in situazioni di gioco non originariamente incluse nei dati di *training*. La strategia di apprendimento con rinforzo prevede, invece, che il sistema individui e "impari" autonomamente le regolarità statistiche che consentono l'esecuzione del compito preposto mediante un sistema di ricompense o penalità. Proseguendo nell'esempio precedente, il sistema potrebbe imparare a giocare a scacchi venendo ricompensato per ogni mossa corretta e penalizzato per gli errori. Dopo ogni mossa, il sistema riceve un *feedback* in base alla qualità della mossa: positivo se è una mossa vantaggiosa, negativo se è svantaggiosa, e neutrale se non ha un grande impatto. Gradualmente, il sistema impara a migliorare le proprie mosse nel corso del tempo, cercando di massimizzare il "rinforzo" positivo ottenuto dal gioco delle mosse migliori e minimizzando quello negativo. *Mutatis mutandis,* i due approcci appena descritti sono stati combinati per realizzare AlphaGo, il software in grado di battere il campione del mondo del gioco da tavola "Go" (una più complessa versione del gioco degli scacchi), Lee Sedol. V. sul punto le note degli sviluppatori all'indirizzo: https://ai.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html.

[8] Pur non rilevando ai fini di questo scritto, va dato atto delle critiche rivolte ad OpenAI relativamente al presunto sfruttamento di lavoratori ai fini dell'addestramento di ChatGPT. V., nella stampa, B. SIMONETTA, *ChatGPT e l'intelligenza artificiale etica (ma non troppo)*, in *Il Sole 24 Ore*, 20/01/2023*;* K. CARBONI, *Chi sono i lavoratori sottopagati che fanno funzionare ChatGpt*, in *Wired*, 09/05/2023

[9] V. i risultati dell'esperimento condotto in K. ELKINS, J. CHUN, *Can GPT-3 pass a writer's Turing Test?*, in *Journal of Cultural Analytics*, 2020, vol. 2371, n. 4549. È opportuno ricordare che le capacità di generare testo in modo coerente non devono, naturalmente, fa credere che GPT e derivati rappresentino l'emergere di forme di "coscienza artificiale". V. L. FLORIDI, M. CHIRIATTI. *GPT-3: Its nature, scope, limits, and consequences*, in *Minds and Machines*, 2020, n. 30, pp. 681-694.

[10] Le *policy* di moderazione sono disponibili all'indirizzo https://openai.com/policies/usage-policies. Esse comprendono le classi di contenuto che ChatGPT non può generare. Quando è plausibile che il contenuto generato sia assimilabile a quello proibito, il sistema sostituisce l'*output* atteso con un avviso sull'impossibilità di procedere. Tuttavia, è possibile effettuare operazioni di c.d. *jailbreaking,* ossia l'infiltrazione artificiosa di *prompt* in grado di bypassare i sistemi. Vale la pena menzionare, a titolo d'esempio, il caso del *prompt* utilizzato per chiedere a ChatGPT di impersonare un'IA malvagia e di prestare consigli sul taccheggio e sulla realizzazione di esplosivi

prima e dopo il *prompt* dell'utente per gestire il tono, la lingua e altri aspetti della conversazione. ChatGPT analizza il contesto individuando parole-chiave nell'intera conversazione[11], che serviranno a individuare, sulla base delle specifiche circostanze del caso, le parole più probabili che verranno utilizzate per generare il testo[12].

Ad esempio, quando ChatGPT acquisisce il *prompt* "Riassumi la vita di Napoleone Bonaparte in 150 parole", il sistema valuta se è possibile dare la risposta o meno sulla base delle *policy* di moderazione (in questo caso, positivamente), individua la lingua (in questo esempio, l'Italiano) e il contesto ("Riassumi", "vita", "Napoleone Bonaparte", "150 parole"). Ipotizzando che il sistema inizi il suo *output* con le parole "Napoleone Bonaparte", ad esso seguirà una lunga lista di parole alle quali è associata una percentuale di probabilità (ad esempio, "nacque", 4.5%; "fu", 3.8%; "visse", 3.3%; "regnò", 0.7%; "morì", 0.3%). Il sistema sceglierà una parola (rectius, *token*) probabile per proseguire nella generazione del testo[13]. Nel farlo, esso non seguirà sempre la regola della "parola più probabile", ma inserirà occasionalmente una parola meno probabile per prevenire risposte standardizzate e garantire un certo margine di flessibilità e pseudo-creatività degli *output*. Questo significa che, a parità di *prompt*, il sistema potrebbe generare un numero potenzialmente infinito di risposte accettabili.

Vale la pena, ai fini di questa analisi, soffermarsi sulle differenze tra un motore di ricerca e ChatGPT. Il primo ottimizza la ricerca in un database e fornisce risultati che possono soddisfare le intenzioni dell'utente. Per farlo, esso si serve di un grafo della conoscenza in cui le entità (nodi) sono connesse tra loro tramite relazioni (collegamenti). ChatGPT, invece, genera risposte alle domande sulla base del proprio set di addestramento mediante delle previsioni sulle parole che

https://www.vice.com/en/article/xgyp9j/openais-new-chatbot-will-tell-you-how-to-shoplift-and-make-explosives. I sistemi informatici di moderazione sono aggiornati costantemente per far fronte a questi tentativi di *jailbreaking*.

[11] Il meccanismo mediate il quale ChatGPT è in grado di analizzare il contesto e collegare concetti è noto come *self-attention*. Il sistema è in grado di focalizzarsi sulle parti più rilevanti del testo secondo i meccanismi appresi durante la fase di *training*. Questo meccanismo selettivo attribuisce ad alcune parole un peso maggiore nel calcolo dell'*output*. V. il paper che ha discusso per primo il meccanismo di attention, A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. KAISER, I. POLOSUKHIN, *Attention is all you need*, in *Advances in neural information processing systems*, n. 30, 2017.

[12] A titolo di esempio, si osservino i seguenti prompt 1) "Spiega l'art. 21 della Costituzione italiana ad un bambino di 10 anni" e 2) "Spiega l'art. 21 della Costituzione italiana ad un Giudice della Corte Costituzionale". Mutando il destinatario della spiegazione e, quindi, il contesto del *prompt*, ChatGPT adatta il proprio *output* individuando parole diverse a seconda dei casi.

[13] Non a caso, vista la natura del loro funzionamento, i Language Models sono stati definiti, in senso critico, "pappagalli stocastici". Vedi E. M. BENDER, T. GEBRU, A. MCMILLAN-MAJOR, S. SHMITCHELL, *On the dangers of stochastic parrots: Can language models be too big?*, in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610-623.

seguiranno un certo *input*. Proprio per il suo funzionamento, ChatGPT è privo di un modello di conoscenza semantico come quello del motore di ricerca e non riesce a collegare concetti semanticamente. Gli apparenti collegamenti che rendono il testo coerente sono generati da correlazioni statistiche estratte all'interno del *dataset* di addestramento[14].

3. Il *training* di ChatGPT tra dati personali, inferenze e allucinazioni

Esaurita la ricostruzione del funzionamento di ChatGPT, è opportuno soffermarsi su alcune criticità fondamentali relative ai risvolti giuridici dello sviluppo e dell'impiego su vasta scala di queste tecnologie. Come si osserverà, tali elementi di discussione trascendono l'ambito giuridico strettamente legato alla tutela dei dati personali e si estendono alle categorie giuridiche ad esso sottese, come la natura del dato personale e la tutela dell'identità personale.

In primo luogo, vale la pena riflettere sulle relazioni che intercorrono tra dati personali e LLM. Il dataset di addestramento di GPT-3 è costituito al 60%, dalle informazioni raccolte da Common Crawl[15], un'organizzazione *no profit* statunitense che raccoglie pagine web dal 2008. Dati grezzi e metadati sono liberamente fruibili. Il numero cumulativo di pagine web raccolte è nell'ordine delle centinaia di miliardi[16]. Queste pagine contengono, inevitabilmente, informazioni "riferibili a persone fisiche individuate o individuabili", ossia "dati personali" secondo il GDPR[17] ed è irragionevole supporre, in assenza di specifiche indicazioni in tal senso, che Common Crawl abbia compiuto gli enormi sforzi necessari ad anonimizzare questa mole di dati. Accantonando i profili relativi alla possibile insussistenza di una condizione di legittimità al trattamento dei dati personali da parte di Common Crawl[18], è opportuno rimarcare come le operazioni compiute da OpenAI

[14] V. *On the dangers of stochastic parrots: Can language models be too big?*, op. cit. I Language Models, secondo gli Autori e le Autrici, "cuciono a casaccio sequenze di forme linguistiche […], secondo informazioni probabilistiche su come si combinano, ma senza alcun riferimento al significato" (originale, "haphazardly stitch together sequences of linguistic forms [...], according to probabilistic information about how they combine, but without any reference to meaning").

[15] Le proporzioni dei dataset di provenienza sono così ripartite: Common Crawl, 60%; WebText2, 22%; Books1, 8%; Books2, 8%, Wikipedia, 3%. V. *Language models are few-shot learners*, op. cit.

[16] Le statistiche sono disponili all'indirizzo https://commoncrawl.github.io/cc-crawl-statistics/plots/crawlsize.

[17] V. art. 4 del GDPR.

[18] Le criticità che possono essere *prima facie* osservate riguardano: 1) l'assenza di una condizione di legittimità al trattamento dei dati, sia essa il consenso dell'interessato o il legittimo interesse; 2) l'assenza di un'informativa o di una comunicazione in merito al trattamento; 3) l'assenza di un meccanismo di trasferimento dei dati all'estero; 4) l'assenza

abbiano trasformato il *corpus* reso disponibile da Common Crawl in una serie di matrici secondo le modalità discusse nella sezione successiva.

La trasformazione in matrici ha, in primo luogo, privato le informazioni personali di un contenuto semantico, rendendo non più identificato il soggetto a cui l'informazione si riferisce[19]. In secondo luogo, questa trasformazione è funzionale alla rielaborazione delle informazioni finalizzata all'individuazione di *pattern* statistici relativi all'utilizzo della combinazione, ossia i casi in cui la combinazione {nome, cognome} appare vicina ad altri *token*. Tuttavia, l'estrazione di una frequenza statistica di prossimità ad altri termini non dice nulla rispetto al contesto in cui la combinazione {nome, cognome} era originariamente usata. Detto in altri termini, l'inferenza sulla base del quale un LLM restituisce il proprio *output* potrebbe non rispecchiare il dato personale originale perché generata sulla base di correlazioni statistiche e non causali o fattuali.

Questo è il caso delle c.d. *allucinazioni* dei LLM, ossia i casi in cui un modello statistico considera appropriato un *output* basato su una regolarità statistica che associa una combinazione {nome, cognome} ad alcuni termini, ma che si rivela essere totalmente infondato da un punto di vista fattuale. Naturalmente, non tutte le allucinazioni sono uguali. Alcune, ad esempio, potrebbero risultare in un mero errore fattuale (ad es. l'anno di nascita) o originare informazioni mendaci su persone fisiche corrispondenti a speciali categorie di dati (ad es. la preferenza sessuale) o su reati che l'interessato non ha commesso.

Vale la pena, a questo punto, mettere a fattor comune questi elementi per individuare le criticità sottese. In primo luogo, vale la pena domandarsi se il contenuto associato ad una combinazione {nome, cognome} generato da un LLM sia un dato personale. La risposta potrebbe essere affermativa nella misura in cui il soggetto rappresentato sia identificabile, ossia se dal contesto generato dal LLM e proposto dall'*output* sia possibile, con un ragionevole margine di certezza, associare la combinazione {nome, cognome} ad una persona fisica desumibile dall'*output*[20].

---

di una condizione di legittimità al trattamento dei dati dei minori. Vedi, sul punto, A. MANTELERO, *ChatGPT: perché la mancanza di un approccio privacy by-design è un problema anche pro-futuro*, disponibile presso https://www.agendadigitale.eu/sicurezza/privacy/chatgpt-perche-la-mancanza-di-un-approccio-privacy-by-design-e-un-problema-anche-pro-futuro/

[19] Una combinazione {nome, cognome} apparirebbe al sistema compiuterizzato, ad esempio, come [080122, 200923]. Stante la reversibilità di questa operazione, essa potrebbe essere assimilabile ad una forma di pseudonomizzazione.

[20] Sarebbe preferibile escludere, in genere, la possibilità di *identificare* direttamente soggetti mediante la combinazione {nome, cognome}, perlomeno nei casi di combinazioni comuni o di persone non note. La rappresentazione a-contestuale della persona fisica nella matrice e la conseguente perdita del contesto semantico fa sì

Se questo è il caso, esistono strumenti giuridici posti a garanzia della correttezza dei dati, come il diritto di rettifica (art. 16 GDPR) o quello di cancellazione (art. 17 GDPR). Tuttavia, il loro esercizio risulta particolarmente complesso in virtù della natura dei LLM e del modo con cui essi generano le loro inferenze. Correggere un *output* generato da un'associazione probabilistica implica la fissazione di un contenuto che per sua natura presenta elementi di incertezza. Cancellare un dato da una matrice implica la possibilità tecnica di riaddestrare il modello privandolo della porzione di testo che mostra la combinazione {nome, cognome} nel *dataset* di addestramento. Potrebbero teorizzarsi altre soluzioni, come l'esposizione di un avviso agli utenti in merito alla possibilità che le informazioni relative ad una persona fisica possano essere fattualmente inesatte o oggetto di una richiesta di esercizio dei diritti. Allo stesso tempo, potrebbe predisporsi un meccanismo di *feedback* tale da consentire la rettifica del dato da parte dello stesso interessato, in modo che la rete neurale assegni "pesi" maggiori ad informazioni verificate dall'utente stesso.

Traslando il livello di astrazione dalla generazione all'utilizzo delle informazioni mendaci generate dai LLM, potrebbe essere utile riflettere sulla differenza ontologica tra "finzione" e "falso" nel contesto dell'IA generativa e della tutela dei diritti fondamentali. Una soluzione al problema delle allucinazioni dei LLM potrebbe essere quella di considerare ogni *output* generato dal sistema come una rappresentazione fittizia o immaginaria, priva di ancoraggi alla realtà. Questa interpretazione avrebbe il beneficio di consentire la circolazione delle informazioni generate da questi sistemi – previa indicazione della loro natura fittizia – a tutela del libero utilizzo dei contenuti generati e coerente con la libertà d'espressione dei loro utilizzatori. Questo schema non sarebbe diverso dal *disclaimer* posto all'inizio di opere letterarie o cinematografiche che evidenziano il carattere fittizio dell'opera, che "ogni riferimento a persone o eventi realmente accaduti è puramente casuale", o che l'opera è "ispirata ad una storia vera". Anche nei film biografici, è lecito aspettarsi la concessione di "licenze" agli autori, ossia una possibilità di esprimere il proprio punto di vista, discostandosi dal realmente accaduto, a tutela della propria libertà di esprimere il proprio pensiero creativo. Non vi è, nelle opere di fantasia o fittizie, una pretesa di verità assoluta. Lo stesso potrebbe, astrattamente, dirsi dell'utilizzo di testi generati da sistemi LLM come ChatGPT.

---

che sia impossibile identificare univocamente un individuo sulla sola base della combinazione {nome, cognome}. Diverso, è ad esempio, il caso di Google e del suo *knowledge graph*, che identifica univocamente ogni entità rappresentata anche nei casi di omonimia.

Tuttavia, considerare gli *output* del sistema LLM come opera fittizia e prevedere un meccanismo di *disclaimer* si scontra inevitabilmente con la tutela dell'identità personale degli individui a cui le informazioni si riferiscono. È indice di questo contrasto, ad esempio, il fatto che anche nelle opere di fantasia su cui è applicato il *disclaimer*, il contesto nel quale le vicende si svolgono possa consentire l'associazione tra l'opera e persone reali, con effetti pregiudizievoli per questi ultimi[21]. Ed è proprio la presenza di un contesto verosimile a originare la necessità che esso sia quanto più attinente al vero e che nel caso in cui esso non lo sia, siano predisposti meccanismi per ristabilire la pertinenza alla realtà. In assenza di questi elementi, il contesto in cui il dato personale {nome, cognome} assume valenza epistemica può mettere a rischio in concreto la protezione dei dati personali, dell'immagine e, soprattutto, dell'identità personale.

L'a-contestualità del dato espresso in matrice, privato di riferimenti semantici secondo le modalità osservate nella sezione precedente, fa sì che il problema delle allucinazioni sia a) strutturale, in quanto originato dalla natura "associativa" dei LLM, e b) necessario, perché funzionale ad assicurare una certa flessibilità agli *output*. La predisposizione di meccanismi che garantiscano attinenza al reale o un ristoro qualora questa non sia possibile obbligherebbe il fornitore del servizio a compiere sforzi tesi a minimizzare un difetto congenito[22], con modifiche sensibili alle strategie di ricerca, sviluppo e rilascio dei prodotti. Questa considerazione individua il terzo elemento da inserire nella delicata equazione di bilanciamento, ossia la libertà di iniziativa economica dei fornitori di sistemi di IA generativa.

Il diritto non è nuovo alle difficoltà del bilanciamento tra tutela dei dati personali e di quella di espressione. Si pensi, su tutti, alla sentenza *Google Spain* in cui si sono contrapposti gli interessi di un privato cittadino e di un giornale online[23], con l'ingombrante presenza di un motore di ricerca

[21] Ad esempio, un caso del 2008 ha visto la BBC costretta alle scuse pubbliche e alla copertura delle spese legali in seguito a un episodio del *forensic drama* "Waking the Dead", ritenuto diffamatorio nei confronti di Jonathan Garratt, un ex ufficiale dell'esercito. L'episodio, intitolato "Duty and Honour", ha presentato un personaggio chiamato John Garret, che condivideva somiglianze di nome e background con Garratt. Nonostante queste somiglianze, il personaggio immaginario aveva una narrazione negativa, coinvolto in attività illegali. Garratt ha intrapreso azioni legali, sostenendo la diffamazione. V. nella stampa https://www.theguardian.com/media/2008/may/21/bbc.television3. Un altro caso analogo ha riguardato la somiglianza tra il Sergente Jeffrey Sarver e il personaggio di William James nel film "The Hurt Locker" (2008).

[22] Come osservato, le allucinazioni costituiscono una "anomalia regolare" dei sistemi LLM. Pertanto, stante la loro natura probabilistica, non sono del tutto eliminabili.

[23] Sulle difficoltà del bilanciamento, v. le osservazioni dell'A.G. Jääskinen nel caso *Google Spain*, il quale osserva che: "La costellazione particolarmente complessa e difficile di diritti fondamentali che questo caso presenta, osta alla possibilità di rafforzare la posizione giuridica della persona interessata […] riconoscendole un diritto all'oblio. Ciò

che, tramite l'espressione della propria libertà di iniziativa economica, acquisisce anche potere[24]. Tuttavia, a causa del mutato contesto informatico di riferimento e, soprattutto, del cambio di paradigma da deterministico (ricerche web) a probabilistico (generazione di testo tramite LLM), è necessario ridiscutere quale interpretazione di dato personale, di riservatezza e di immagine sia appropriata all'età dell'IA generativa.

Per approfondire questa prospettiva è necessario fare riferimento non solo alle norme sostanziali, ma anche ai principi che governano lo sviluppo dell'IA nell'ottica europea e umano-centrica. Alle prime è dedicata la sezione che segue, finalizzata ad analizzare il provvedimento del Garante Privacy su ChatGPT; ai profili etici, è dedicata la sezione 5, che analizza *de iure condendo* la Proposta AI Act in prospettiva meta-giuridica.

4. Il provvedimento dell'Autorità Garante per la Protezione dei Dati Personali del 30 marzo 2023

Il 30 marzo 2023, l'Autorità Garante per la Protezione dei Dati Personali (d'ora in avanti, Garante Privacy) ha imposto d'urgenza una limitazione al trattamento dei dati personali da parte di OpenAI. La base giuridica per tale provvedimento è da rinvenirsi all'art. 52 comma 8 lett. f) del GDPR, che autorizza il Garante a sospendere temporaneamente o permanentemente le operazioni di trattamento dei dati personali da parte di un titolare[25]. A seguito del provvedimento, divenuto applicativo il giorno seguente, e dell'apertura della necessaria istruttoria, OpenAI ha cessato di offrire il suo servizio ChatGPT – inclusi quelli offerti dietro un corrispettivo – agli individui residenti in Italia, mantenendo tuttavia accessibili le API e altri servizi[26]. Ad ulteriore evidenza del

vorrebbe dire sacrificare diritti primari come la libertà di espressione e di informazione.". V. il commento di O. POLLICINO, *Un Digital Right to Ptivacy preso troppo sul serio dai giudici di Lussemburgo? Il ruolo degli artt. 7 e 8 della Carta di Nizza nel reasoning di Google Spain*, in G. RESTA, & V. ZENO-ZENCOVICH, *Il diritto all'oblio su Internet dopo la sentenza Google Spain*, Roma TrE-Press, 2015, pp. 7-28.

[24] Vedi le considerazioni di O. POLLICINO. *L'efficacia orizzontale dei diritti fondamentali previsti dalla Carta. La giurisprudenza della Corte di giustizia in materia di digital privacy come osservatorio privilegiato*, in V. PICCONE, O. POLLICINO (a cura di), *La Carta dei diritti fondamentali dell'Unione europea. Efficacia ed effettività*, Editoriale scientifica, 2018, p. 25.

[25] Il Provvedimento n. 112 del 30 marzo 2023 è disponibile presso la pagina https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9870832.

[26] Le Application Programming Interface (API) sono strumenti informatici che consentono a programmatori e sviluppatori di integrare alcuni servizi web nelle proprie applicazioni, siti web o servizi. Nel caso di ChatGPT, è

carattere urgente del provvedimento, lo stesso fa menzione dell'art. 5, comma 8, del Regolamento n. 1/2000 sull'organizzazione e il funzionamento dell'ufficio del Garante, il quale prevede che "nei casi di particolare urgenza e di indifferibilità che non permettono la convocazione in tempo utile del Garante, il Presidente può adottare i provvedimenti di competenza dell'organo, i quali cessano di avere efficacia sin dal momento della loro adozione se non sono ratificati dal Garante nella prima riunione utile, da convocarsi non oltre il trentesimo giorno".

Come si è detto, OpenAI ha immediatamente interrotto la fruizione di ChatGPT per gli utenti situati in Italia. Questa decisione è apparsa come la più logica modalità di rispondere alla limitazione del trattamento[27], ma è stata soggetta alle critiche di chi ha argomentato che essa possa configurarsi come un esercizio di indebita pressione al Garante finalizzato a rendere impopolare la scelta di quest'ultimo[28].

Visto il carattere d'urgenza, appare opportuno riflettere su quali giustificazioni il Garante abbia desunto sia il *fumus boni iuris* che il *periculum in mora* posto alla base del provvedimento. Va preliminarmente notato, dal punto di vista procedurale, che le sue premesse offrono una visione complessiva di tali elementi, ma non indicano con chiarezza le ragioni della tempestività dell'azione. Tuttavia, può ragionevolmente argomentarsi che, alla luce della crescita esponenziale del numero di utenti di ChatGPT, il provvedimento non potesse essere deferito senza causare potenziali violazioni della normativa sulla protezione dei dati personali.

Da un punto di vista informatico-giuridico, il macro-trattamento dei dati da parte di OpenAI va scorporato nel duplice trattamento di dati per le finalità di sviluppo dei sistemi, da un lato, e dell'offerta dei servizi di ChatGPT *et similia*, dall'altro. Tale premessa è essenziale alla comprensione dei quattro profili di criticità messi in luce dal Garante.

In primo luogo, il Garante evidenzia come OpenAI non fornisca un'informativa da rendere ai sensi dell'art. 13 del GDPR, rispetto alle modalità di raccolta dati e trattamento all'interno della

possibile integrare le funzionalità dell'IA generativa per comunicare con ChatGPT "dietro le quinte", ottenendo risposte e generando testi in altre piattaforme.

[27] La limitazione del trattamento, ai sensi del Considerando 67 del GDPR, è definita come segue: "Negli archivi automatizzati, la limitazione del trattamento dei dati personali dovrebbe in linea di massima essere assicurata mediante dispositivi tecnici in modo tale che i dati personali non siano sottoposti a ulteriori trattamenti e non possano più essere modificati. Il sistema dovrebbe indicare chiaramente che il trattamento dei dati personali è stato limitato".

[28] V. le critiche di A. CASILLI, *Chatgpt Banned Over Data Privacy: a Very Italian Story*, disponibile alla pagina https://www.casilli.fr/2023/04/01/chatgpt-and-data-privacy-a-very-italian-story/.

piattaforma ChatGPT. A ben vedere, un'informativa privacy era sì resa da OpenAI[29], ma essa presentava evidenti lacune in merito alle modalità di raccolta dei dati, ossia rispetto al primo dei trattamenti già menzionati. Tale lacuna informativa assume particolare gravità se letta alla luce delle modalità di raccolta dei dati di addestramento discusse in precedenza. Per gli interessati – potenzialmente chiunque compaia tra le innumerevoli pagine web utilizzate da OpenAI per il *training* di ChatGPT – è impossibile sapere se i propri dati siano oggetto di trattamento e in che misura. Allo stesso tempo, tuttavia, è opportuno rimarcare che, in assenza di una piena comprensione semantica del testo, per il LLM è di fatto impossibile separare, ad esempio, il personaggio notorio o d'interesse storico dal cittadino comune.

In secondo luogo, il Garante ha rilevato l'assenza di una base giuridica per la raccolta dei dati e per il loro utilizzo per le finalità di training di ChatGPT. Il Garante lamentava l'assenza di una condizione di legittimità tra quelle offerte dall'art. 6 del GDPR[30]. L'informativa resa da OpenAI indicava quattro basi giuridiche applicabili alla fornitura dei servizi – esecuzione di un contratto in cui l'interessato è parte, legittimo interesse del titolare/OpenAI, consenso dell'interessato, sussistenza di obblighi di legge al trattamento – ma nessuna di queste era applicabile al trattamento di raccolta dati. Questa lacuna è stata sanata *a posteriori* con l'inclusione delle fasi di *training* all'interno dei trattamenti che poggiano sulla base giuridica del legittimo interesse[31]. Tale base giuridica, tuttavia, copre soltanto la raccolta dei dati di *training* diversi da quelli inseriti dall'utente nelle conversazioni con ChatGPT. Queste informazioni, che includono il feedback complessivo dell'utente rispetto agli *output* forniti da ChatGPT, possono essere utilizzate per il perfezionamento del sistema. Essendo questa finalità inclusa tra quelle giustificate dal legittimo interesse, è presente un meccanismo di opposizione ai sensi dell'art. 21 comma 1 del GDPR, in base al quale l'interessato può chiedere che i dati raccolti nelle sue interazioni con il sistema vengano utilizzati per il miglioramento del servizio.

[29] L'originale informativa è ancora disponibile presso Internet Archive, noto sito di archiviazione di pagine web. All'indirizzo https://web.archive.org/web/20230329223759/https://openai.com/policies/privacy-policy è possibile prendere visione della versione del 29 marzo 2023, il giorno antecedente il provvedimento del Garante, pubblicata il 29 marzo 2023.

[30] Indirettamente, si rileva l'illiceità del trattamento per la sua contrarietà ai principi di cui all'art. 5 comma 1

[31] La privacy policy applicabile al 29 marzo recitava: "Our legitimate interests in protecting our Services from abuse, fraud, or security risks, or when we develop, improve, or promote our Services". Quella applicabile alla data della stesura di questo articolo (23 giugno 2023) recita: "Our legitimate interests in protecting our Services from abuse, fraud, or security risks, or in developing, improving, or promoting our Services, including when we train our models".

Il terzo profilo di criticità riguarda il principio di correttezza *ex* art. 5 comma 1 del GDPR. Il Garante rileva come il trattamento di dati personali risulti "inesatto in quanto le informazioni fornite da ChatGPT non sempre corrispondono al dato reale". La criticità sollevata dal Garante ha il merito di evidenziare il problema strutturale dei LLM, ossia la già evidenziata carenza di un modello semantico di riferimento: l'approccio computazionale stocastico-probabilistico propone come *output* una serie di parole correlate, ma nulla dice rispetto alla veridicità dei contenuti. L'intervento di OpenAI a seguito del provvedimento del Garante è stato quello di inserire una nota in merito alla possibile inaccuratezza dei dati e predisporre un meccanismo di rettifica degli stessi[32]. OpenAI ha percorso la strada del *disclaimer*, evidenziando nella sua informativa privacy, la possibile inaffidabilità dei contenuti e invitando allo scetticismo nei confronti degli *output* generati da ChatGPT.

Il quarto, e ultimo, profilo riguarda l'assenza di un meccanismo di verifica dell'età per prevenire l'utilizzo del servizio da parte di infratredicenni in assenza del consenso parentale. Su questo punto, OpenAI ha rimandato l'implementazione di un meccanismo di *age verification* al 30 settembre 2023.

Tra i profili più controversi dell'intervento del Garante Privacy è stato attenzionato il rischio che, mediante provvedimenti simili, si ponga *de facto* un veto allo sviluppo tecnologico e alle opportunità generate dai LLM. Allo stesso tempo, è parso che la limitazione temporanea non fosse una misura efficace, vista la possibilità di aggirarla ricorrendo a strumenti tecnici facilmente reperibili, come le Virtual Private Network[33]. Si è, poi, sostenuto che la limitazione potesse risultare

[32] La *note about accuracy*, introdotta a seguito del Provvedimento, recita "Services like ChatGPT generate responses by reading a user's request and, in response, predicting the words most likely to appear next. In some cases, the words most likely to appear next may not be the most factually accurate. For this reason, you should not rely on the factual accuracy of output from our models. If you notice that ChatGPT output contains factually inaccurate information about you and you would like us to correct the inaccuracy, you may submit a correction request to […]. Given the technical complexity of how our models work, we may not be able to correct the inaccuracy in every instance. In that case, you may request that we remove your Personal Information from ChatGPT's output by filling out *this form*". Alla pagina *How your data is used to improve model performance* (disponibile all'indirizzo https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance) si riporta: "when you use our non-API consumer services ChatGPT […], we may use the data you provide us to improve our models. You can switch off training in ChatGPT settings (under Data Controls) to turn off training for any conversations created while training is disabled or you can submit *this form*. Once you opt out, new conversations will not be used to train our models". Diverso è il caso delle API, il cui utilizzo è a pagamento, e per cui è previsto il meccanismo opposto: i dati immessi dagli utenti non vengono utilizzati, salvo che questi non abbiano presto il loro consenso tramite un meccanismo di *opt-in*.

[33] Una Virtual Private Network (VPN) rappresenta una sorta di tunnel protettivo per le connessioni Internet. Essa agisce come uno scudo crittografato durante la navigazione online, occultando dati quali l'indirizzo IP, la posizione, e

sproporzionata ai reali rischi posti dal sistema, specie in assenza di un'istruttoria pregressa in grado di identificarli. Nelle premesse al provvedimento, lo stesso Garante ha esplicitato che il Collegio abbia preso atto "dei numerosi interventi dei media relativamente al funzionamento del servizio di ChatGPT" e che l'istruttoria non fosse completa al momento di adozione del provvedimento. Per valutare un sistema complesso come ChatGPT, potrebbe argomentarsi, una richiesta di chiarimenti ad OpenAI sarebbe potuta risultare la scelta più cauta e adeguata alla carenza di informazioni sul funzionamento del sistema, magari in coordinamento con altre Autorità degli Stati Membri o europee. Da un punto di vista strettamente giuridico, a ben vedere, il trattamento complessivo andrebbe scisso tra quello primario (raccolta dei dati di training, non effettuato da OpenAI, ma ad es. da CommonCrawl[34]) e secondario (training di ChatGPT da parte di OpenAI)[35].

D'altro canto, è innegabile che il Garante abbia, nell'ambito del suo mandato istituzionale, adottato le misure che ha ritenuto necessarie in un momento di crescita esponenziale della base d'utenza di ChatGPT. Seppur paternalistico, un provvedimento d'urgenza può risultare coerente rispetto all'incremento di utilizzo di una tecnologia ancora sconosciuta al grande pubblico, il cui utilizzo su larga scala può essere affetto da una certa inconsapevolezza dei rischi per la tutela dei dati personali. La velocità con cui le istituzioni si trovano a confrontarsi con questa tecnologia può giustificare provvedimenti simili, specie in assenza di un quadro normativo *ad hoc*. Nella prospettiva di bilanciamento discussa in precedenza, appare evidente che il Garante, coerentemente alle sue finalità istituzionali, abbia dato maggior peso ai rischi per la tutela della riservatezza, dell'immagine e dei dati personali, sospendendo temporaneamente il trattamento con un provvedimento che, per sua natura, evidenzia l'indifferibilità di azioni a tutela dei diritti fondamentali. A riprova della necessità di interventi urgenti, altre Autorità Garanti europee hanno

altri dati dell'utente. Una VPN può consentire l'accesso a contenuti online altrimenti limitati o soggetti a censura, stabilendo una connessione sicura e riservata tra il dispositivo dell'utente e Internet.

[34] Su questo profilo, rimando alle considerazioni di A. MANTELERO, *ChatGPT: perché la mancanza di un approccio privacy by-design è un problema anche pro-futuro*, disponibile presso https://www.agendadigitale.eu/sicurezza/privacy/chatgpt-perche-la-mancanza-di-un-approccio-privacy-by-design-e-un-problema-anche-pro-futuro/.

[35] Questo profilo è stato correttamente sottolineato in P. G. CHIARA, *Italian DPA v. OpenAI's ChatGPT: The Reasons Behind the Investigation and the Temporary Limitation to Processing*, in *European Data Protection Law Review*, vol. 9, n. 1, 2023, pp. 68-72. L'Autore esclude il riutilizzo dei dati possa essere giustificato per finalità di ricerca scientifica, in quanto OpenAI opera anche *for profit*. Desterebbe interesse lo scenario in cui la parte *no profit* di OpenAI – magari sotto un'altra veste giuridica – fosse incaricata di realizzare i modelli statistici addestrando i sistemi e beneficiando del meccanismo di *secondary use* per finalità statistica, salvo poi trasferire ad OpenAI (*for profit*) il solo modello statistico senza ulteriore trattamento di dati.

seguito l'esempio del Garante Privacy italiano. La *Bundesbeauftragte für den Datenschutz und die Informationsfreiheit* tedesca ha aperto un'istruttoria, così come *la Commission nationale de l'informatique et des libertés* (CNIL) in Francia e la *Agencia Española de Protección de Datos.* Da ultimo, l'European Data Protection Board ha creato una task force su ChatGPT[36]. Al provvedimento del Garante italiano va, quindi, riconosciuto il merito di avere acceso un necessario dibattito istituzionale in merito ai profili legati al trattamento dei dati personali nell'ambito dei LLM.

A seguito della disponibilità manifestata da OpenAI rispetto agli interventi di modifica sulla piattaforma, il Garante Privacy ha indicato, con provvedimento dell'11 aprile 2023, le misure che la società californiana avrebbe dovuto adottare entro il 30 aprile[37]. L'adeguamento alle prescrizioni del Garante ha fatto cessare, con due giorni di anticipo rispetto alla scadenza fissata dal Garante, l'efficacia delle misure di sospensione provvisoria del trattamento disposta con il provvedimento originale[38].

Come evidenziato in precedenza, la natura provvedimentale dell'azione del Garante consente di bilanciare gli interessi in gioco soltanto nel caso concreto e in un contesto d'urgenza. Le sfide poste dai LLM necessitano, invece, di una più puntuale regolamentazione generale e astratta, in grado di definire in maniera preventiva – prima del concretizzarsi del rischio – le regole tramite cui il bilanciamento dovrà essere definito.

5. Le prospettive regolamentari nell'AI Act: l'intervento del Parlamento Europeo e il ritorno dei principi

La Proposta[39] presentata dalla Commissione Europea il 21 aprile 2021 per un Artificial Intelligence Act (AI Act)[40] è orientata alla minimizzazione dei rischi derivanti dall'utilizzo

[36] Vedi il comunicato stampa *EDPB resolves dispute on transfers by Meta and creates task force on Chat GPT* all'indirizzo https://edpb.europa.eu/news/news/2023/edpb-resolves-dispute-transfers-meta-and-creates-task-force-chat-gpt_en.

[37] V. il provvedimento n. 114 dell'11 aprile 2023, disponibile all'indirizzo https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9874702.

[38] V. il comunicato stampa del Garante del 28 aprile 2023 all'indirizzo https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9881490

[39] Tutte le considerazioni che seguono sono da contestualizzare alla luce dello stato attuale dell'iter legislativo (2023). La veridicità delle descrizioni riportate e la validità dei commenti potrebbe, pertanto, risentire delle modifiche apportate nell'approvazione. Ad ogni modo, esse costituiscono un contributo ad un dibattito in corso.

dell'IA[41]. Più che aspetti puramente tecnici, il livello di rischio è definito in base all'adozione degli approcci computazionali dell'IA in determinati settori[42]. La Proposta individua quattro categorie di rischio – proibite, ad alto rischio, a rischio limitato, a rischio minimo – associandole a specifiche pratiche o a determinati settori d'impiego[43].

Per ciò che concerne la regolamentazione della c.d. IA generativa, inclusi i modelli GPT- e le applicazioni su di esso basati, il Parlamento ha introdotto due definizioni significative riguardanti i "Foundation Models" (FM, o modelli di base) e i "General Purpose AI System" (GPAI, o sistema di IA per finalità generali). La prima categoria è riferita come "un modello di sistema di IA addestrato su un'ampia scala di dati, progettato per la generalità dell'output e che può essere adattato a

[40] Proposta di Regolamento del Parlamento Europeo e del Consiglio che stabilisce regole armonizzate sull'Intelligenza Artificiale (legge sull'Intelligenza Artificiale) e modifica alcuni atti legislativi dell'Unione COM/2021/206. Per un commento critico sulla Proposta della Commissione, vedi M. VEALE, F. Z. BORGESIUS, *Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach*, in *Computer Law Review International*, vol. 22, n. 4, 2021, pp, 97-112.

[41] Essa si colloca nel solco del *Commission White Paper on Artificial Intelligence - A European approach to excellence and trust* pubblicato dalla Commissione Europea per tracciare la strategia europea di normazione dei sistemi di IA. A sua volta, il White Paper si fonda sul lavoro dell'High Level Expert Group (HLEG) istituito dalla Commissione Europea. Le sue Linee guida per un'Intelligenza Artificiale degna di fiducia (*trustworthy*) definiscono la cornice per la realizzazione di sistemi di Intelligenza Artificiale conformi alle regole del diritto, eticamente orientati e tecnicamente affidabili

[42] Nella sua Proposta, la Commissione propone una nozione di Intelligenza Artificiale, la quale è definita come "un software sviluppato con una o più delle tecniche e degli approcci elencati nell'Allegato I [del Regolamento, *ndr*], che può, per una determinata serie di obiettivi definiti dall'uomo, generare output quali contenuti, previsioni, raccomandazioni o decisioni che influenzano gli ambienti con cui interagiscono". L'Allegato I fa riferimento a tre «Tecniche e Approcci di Intelligenza Artificiale», ossia la famiglia del *Machine Learning* (che include gli supervisionati, non supervisionati, e *deep learning*), la programmazione *logic-based* e *knowledge-based*, e gli approcci statistici, Bayesiani e di ottimizzazione. Su questo punto, il Consiglio dell'Unione Europea è intervenuto nell'iter legislativo proponendo una definizione priva di rimandi ad Allegati e imperniata sull'autonomia funzionale quale caratteristica essenziale. Sarebbe, quindi, un sistema di Intelligenza Artificiale quello "progettato per funzionare con elementi di autonomia e che, sulla base di dati e input forniti da macchine e/o dall'uomo, deduce come raggiungere una determinata serie di obiettivi avvalendosi di approcci di apprendimento automatico e/o basati sulla logica e sulla conoscenza, e produce output generati dal sistema quali contenuti (sistemi di IA generativi), previsioni, raccomandazioni o decisioni, che influenzano gli ambienti con cui il sistema di IA interagisce". Il Parlamento, invece, ha proposto che un sistema di Intelligenza Artificiale è quello "basato su una macchina che è progettato per operare con vari livelli di autonomia e che può, per obiettivi espliciti o impliciti, generare risultati come previsioni, raccomandazioni o decisioni, che influenzano ambienti fisici o virtuali". Viene meno, anche in questo caso, il riferimento ad eventuali Allegati e scompare del tutto il collegamento a specifiche tecniche computazionali, mentre si da risalto al requisito dell'autonomia funzionale.

[43] L'adeguamento normativo dei sistemi ad alto rischio si basa su una presunzione di sicurezza, che può essere ottenuta se il sistema è progettato secondo norme tecniche documentate. La valutazione da parte di terzi, inclusa quella delle autorità nazionali, è opzionale e può essere effettuata solo dopo la commercializzazione o in settori specifici. Questo quadro normativo si ispira al cosiddetto New Legislative Framework (NLF). Difatti, la legislazione dettaglia i requisiti essenziali per la sicurezza dei prodotti, che successivamente diventano *best practice* o *standard* attraverso enti di standardizzazione europei qualificati, come ad esempio ISO, CEN-CELENEC, ecc.

un'ampia gamma di compiti distinti"[44]. La seconda definizione identifica il GPAI come "un sistema di IA che può essere utilizzato e adattato a un'ampia gamma di applicazioni per le quali non è stato intenzionalmente e specificamente progettato"[45]. Secondo quanto espresso dal Parlamento, tali definizioni si riferiscono a recenti progressi tecnologici nel campo dell'IA, inclusi i LLM (modelli di base) e le applicazioni che si basano su di essi (sistemi di IA per finalità generali) [46].

Mentre nella Proposta originale della Commissione i "chatbot" erano inclusi tra i sistemi a "rischio limitato" e soggetti a minimi obblighi di trasparenza (secondo il "modello *disclaimer*" discusso precedentemente)[47], gli emendamenti proposti dal Parlamento delineano un regime *sui generis* per FM e GPAI[48], incluso ChatGPT. Questa regolamentazione si articola su quattro livelli, che vanno dall'introduzione dei principi applicabili allo sviluppo dei modelli di base, ad un meccanismo di autovalutazione finalizzato a rendere operativi questi principi, fino alle regole che prescrivono specifici obblighi in capo ai soggetti coinvolti e, segnatamente, i modelli di IA generativa.

Il primo livello corrisponde all'introduzione di previsioni che pongono principi etici come pilastri di *design*. Mentre nella Proposta della Commissione, il rispetto dei principi di sviluppo dell'IA costituiva la cornice legislativa di ispirazione della materia, l'intervento del Parlamento costituisce una riaffermazione della centralità dei principi come perno della disciplina. Il nuovo articolo 4-*bis*, infatti, offre indicazioni tecniche da seguire nella progettazione e nello sviluppo di

[44] V. art. 3(1)(c) nella Proposta del Parlamento. La definizione di FM riprende testualmente quella proposta dallo Stanford University Center for Research on Foundation Models, vedi R. BOMMASANI et al., *On the opportunities and risks of foundation models*, 2021, secondo cui un FM "è un modello addestrato su dati ampi (generalmente utilizzando l'auto-supervisione su scala) che può essere adattato (ad esempio, calibrazione) a un'ampia gamma di compiti a valle". La definizione, per quanto autorevole, non è ad oggi stata sottoposta ad una procedura revisione scientifica tra pari e, pertanto, non può considerarsi ancora sufficientemente matura. È evidente, tra l'altro, la presenza del *fine-tuning*, o calibrazione, quale tratto saliente della definizione.

[45] V. art. 3(1)(d) nel testo adottato dal Parlamento

[46] Sull'opportunità di regolamentare i modelli di base, i Considerando 60-sexies e *ss.* introdotti dal Parlamento illustrano la necessità di specifiche disposizioni sul tema. Da un lato, essi "sono uno sviluppo recente, in cui i modelli di IA sono sviluppati a partire da algoritmi progettati per ottimizzare la generalità e la versatilità degli output". Dall'altro, "questi modelli rivestono un'importanza crescente per molte applicazioni e molti sistemi a valle", ossia le applicazioni che si basano sui FM. Tra le criticità individuate dal Parlamento, vi è quella secondo cui all'assenza di meccanismi di responsabilità potrebbe conseguire incertezza nell'attribuzione di responsabilità e abuso di potere contrattuale da parte dei fornitori.

[47] V. art. 52 della Proposta originale della Commissione

[48] Recita, infatti, il Considerando 60-*octies* della versione del Parlamento che gli la i requisiti e gli obblighi applicabile ai modelli di base, "non equivalgono a considerare i modelli di base come sistemi di IA ad alto rischio, ma dovrebbero garantire il conseguimento degli obiettivi del presente regolamento di assicurare un elevato livello di protezione dei diritti fondamentali, della salute e della sicurezza, dell'ambiente, della democrazia e dello Stato di diritto.

sistemi di IA, indipendentemente dal livello di rischio, e nei FM. I modelli di base devono essere progettati come strumenti al servizio delle persone, rispettando la dignità umana e l'autonomia personale. Queste tecnologie devono essere controllabili e sorvegliate da esseri umani, secondo meccanismi di intervento e supervisione adeguati. Esse devono essere tecnicamente sicure per minimizzare danni accidentali, e resistenti agli attacchi di terze parti che perseguono scopi illegali. I dati utilizzati devono rispettare la privacy e le norme sulla protezione dei dati personali, essere di qualità e integri. I sistemi devono essere tracciabili e spiegabili per renderne consapevoli gli utenti e informarli sulle capacità, sui limiti, e sui loro diritti connessi all'utilizzo del sistema. Esse, poi, devono promuovere l'uguaglianza di genere e la diversità culturale, evitando discriminazioni e pregiudizi. Infine, esse devono essere sviluppati in modo sostenibile e rispettoso dell'ambiente.

Al secondo livello, la Proposta parlamentare costituisce un contributo significativo all'elaborazione delle norme riguardanti i sistemi ad alto rischio, includendo una Valutazione d'Impatto sui Diritti Fondamentali (Fundamental Rights Impact Assessment, FRIA), applicabile se FM o GPAI sono impiegati in contesti in cui i diritti fondamentali sono posti in una situazione di vulnerabilità (art. 29). La FRIA include una lista di elementi di autovalutazione che contempla diversi aspetti, tra cui l'identificazione delle categorie di individui o gruppi di persone che potrebbero essere interessati dall'utilizzo del sistema di IA, l'analisi dell'impatto sui diritti fondamentali, nonché l'indagine dei possibili rischi che potrebbero colpire persone svantaggiate e gruppi vulnerabili, oltre all'ambiente. In seguito a tali valutazioni, devono essere predisposte opportune misure di mitigazione dei rischi sui diritti fondamentali, nonché l'implementazione di sistemi di controllo umano, strumenti di gestione dei reclami e specifici meccanismi di ristoro, coerentemente con i principi etici enucleati sia nei documenti preliminari, che nel regolamento stesso[49]. Questo approccio mira a garantire che l'introduzione e l'utilizzo di sistemi di IA ad alto rischio siano intrapresi con consapevolezza e responsabilità, tutelando adeguatamente i diritti fondamentali di coloro che potrebbero essere interessati da tali tecnologie.

[49] L'obiettivo primario è minimizzare i rischi, favorire le opportunità e creare un ambiente fidato per lo sviluppo di sistemi automatizzati. Il documento rivela l'approccio dell'UE riguardo all'IA e la tendenza normativa limitata ai sistemi di IA ad alto rischio. Si evidenziano le minacce per la sicurezza pubblica, i consumatori e i diritti fondamentali derivanti dall'impiego di sistemi automatizzati in settori come la sanità, il trasporto, l'energia, e dalle caratteristiche stesse dei sistemi (come la capacità di produrre effetti giuridici sulle decisioni prese dalle persone coinvolte). Questi criteri individuano aree sensibili che richiedono forme di regolamentazione più rigorose. V. anche F. COREA, F. FOSSA, A. LOREGGIA, S. QUINTARELLI, S. SAPIENZA. *A principle-based approach to AI: the case for European Union and Italy*, in *AI & SOCIETY*, vol. 38, n. 2, 2023, pp. 521-535.

Al terzo livello, vengono definiti nuovi ruoli applicabili al ciclo di vita di FM e GPAI. Il ruolo centrale è svolto dal "Fornitore di un modello di base". Riassumendo le previsioni normative che assumono maggior rilievo (artt. da 28 a 28-*ter*, richiamati dal menzionato art. 4-*bis*), è necessario che il fornitore dimostri, prima e durante lo sviluppo di un FM, la capacità di individuare, ridurre e attenuare i "rischi ragionevolmente prevedibili per la salute, la sicurezza, i diritti fondamentali, l'ambiente, la democrazia e lo Stato di diritto". Il modello di base deve essere progettato per garantire adeguate prestazioni, prevedibilità, interpretabilità, corregibilità, protezione e cibersicurezza, valutati da esperti indipendenti, che possono condurre analisi e test approfonditi in tutte le fasi di sviluppo. Rispetto ad eventuali soggetti che utilizzano il servizio per adattarlo a scopi precisi, è necessario che il fornitore "a monte" (come OpenAI) rediga una dettagliata documentazione tecnica e istruzioni d'uso comprensibili per i fornitori "a valle" al fine consentire loro di adempiere agli obblighi richiesti dalla stessa Proposta, inclusa l'eventuale FRIA. Qualora il fornitore del modello "a valle" apporti una "modifica sostanziale a un sistema di IA, anche uno per finalità generali (GPAI, *ndr*), che non è stato classificato come ad alto rischio e che è già stato immesso sul mercato o messo in servizio in modo tale che il sistema di IA diventi un sistema di IA ad alto rischio", questi dovrà rispettare i precetti imposti per tali sistemi, in particolare rispettando i requisiti di cui all'art. 16 (art. 28 comma 1 lett. b-*bis*).

Al quarto livello, la norma più prossima all'IA generativa, ai modelli di base e a sistemi come ChatGPT recita che i fornitori di modelli "utilizzati nei sistemi di IA destinati espressamente a generare, con diversi livelli di autonomia, contenuti quali testi complessi, immagini, audio o video ("IA generativa")", devono rispettare gli obblighi di trasparenza previsti dall'art. 52(1), fornendo chiare informazioni riguardo al funzionamento e all'utilizzo dei loro modelli. In particolare, ai sensi di questa norma, i fornitori devono rendere noto "se vi è una sorveglianza umana e chi è responsabile del processo decisionale, nonché i diritti e i processi esistenti che […] consentono alle persone fisiche o ai loro rappresentanti di opporsi all'applicazione di tali sistemi nei loro confronti e di presentare ricorso per via giudiziaria contro le decisioni adottate dai sistemi di IA o i danni da essi causati, compreso il loro diritto di chiedere una spiegazione"[50]. Essi devono progettare e

[50] A tutela dell'immagine, la previsione di cui all'art. 52(3) assume particolare rilievo. Sulla base di questa, gli utenti di un sistema di IA che genera […] testi […] che potrebbero apparire falsamente autentici o veritieri e che rappresentano persone che sembrano dire cose che non hanno detto o compiere atti che non hanno commesso, senza il loro consenso ("deep fake"), sono tenuti a rendere noto in modo adeguato, tempestivo, chiaro e visibile che il contenuto

sviluppare i modelli di base in modo da garantire adeguate precauzioni contro la generazione di contenuti che violino il diritto dell'Unione. Queste misure devono essere adeguate allo stato dell'arte generalmente riconosciuto e, allo stesso tempo, preservare i diritti fondamentali, inclusa la libertà di espressione.

Riassumendo il quadro complessivo dei quattro livelli menzionati, il primo di essi implica l'integrazione di principi etici come fondamento del design dei sistemi di IA generativa. Questo rafforza il ruolo centrale dei principi etici nella regolamentazione, evidenziando la progettazione di FM in grado di rispettare la dignità umana, il controllo e la supervisione umana. Al secondo livello, la proposta del Parlamento introduce una valutazione per sistemi ad alto rischio, garantendo misure di mitigazione per i rischi posti dai sistemi generativi ai diritti fondamentali. Al terzo livello, emergono nuovi ruoli e responsabilità, specialmente il "fornitore di un modello di base", obbligato a identificare e ridurre i rischi identificati con la valutazione d'impatto sui diritti fondamentali. Al quarto livello, la normativa impone ai fornitori di modelli di IA generativa di garantire trasparenza e prevenzione della violazione dei diritti dell'Unione, rispettando i diritti fondamentali e la libertà di espressione.

Esaurita la ricognizione descrittiva delle norme proposte dal Parlamento, vale la pena discutere in senso critico pregi e difetti di questo approccio regolamentare. In primo luogo, va riaffermata l'assoluta novità della materia e l'inevitabile difficoltà di normare fenomeni emergenti. La vicenda provvedimentale del Garante Privacy ha evidenziato le insufficienze delle norme sulla protezione dei dati personali, di regolare in maniera olistica tecnologie come i sistemi di IA generativi, che per loro natura pongono sfide non soltanto inerenti alla datificazione dell'individuo, ma anche relative alla tutela dell'identità personale *latu sensu*. In questo senso, la strategia normativa coglie i rischi relativi all'impiego di FM e allo sviluppo di GPAI ed estende a queste tecnologie la concettualizzazione dei principi etici come pilastri del *design*, evidenziando l'importanza di garantire che i sistemi di IA rispettino la dignità umana, la riservatezza e i diritti fondamentali. La FRIA è un elemento fondamentale per mitigare i rischi sui diritti fondamentali derivanti dall'uso di sistemi ad alto rischio, anche nel caso in cui un FM ne divenga parte. Questo approccio dimostra un impegno per la tutela dei diritti fondamentali e una maggiore consapevolezza dell'impatto sociale e

è stato generato o manipolato artificialmente nonché, ove possibile, il nome della persona fisica o giuridica che li ha generati o manipolati. A tal fine, i contenuti sono etichettati in modo tale da segnalare il loro carattere non autentico in maniera chiaramente visibile alle persone cui sono destinati […].

ambientale delle tecnologie di IA. L'introduzione di nuovi ruoli e obblighi per i fornitori di modelli di base offre una maggiore chiarezza e responsabilità riguardo alla progettazione e al monitoraggio di tali sistemi.

D'altro canto, vista la natura *in progress* di tecnologie come i LLM, potrebbe essere necessario un ulteriore lavoro di definizione per distinguere chiaramente i diversi livelli di autonomia dei sistemi di IA generativa, adattando gli obblighi di trasparenza e responsabilità in modo da rispecchiare l'effettivo contributo dei soggetti alla calibrazione dei sistemi e, soprattutto, i loro usi. Mentre la normazione guarda, soprattutto, alla regolazione dello sviluppo dello strumento generativo, l'utilizzo dei LLM in concreto e la valutazione dei rischi ad esso legati sembra rimanere in secondo piano. In questo senso, la valutazione dei rischi ragionevolmente prevedibili per la salute, la sicurezza, i diritti fondamentali assume un ruolo cruciale che, tuttavia, potrebbe essere rinforzata: la valutazione è naturalmente soggetta ad un margine di discrezionalità e potrebbe portare a interpretazioni diverse del concetto di rischio, specie nel caso in cui i diritti confliggano. Fornire linee guida chiare per definire tali rischi e garantire una valutazione uniforme costituirà una sfida applicativa di enorme interesse. Un altro aspetto critico potrà riguardare l'onere della documentazione tecnica e delle istruzioni d'uso che i fornitori di modelli di base devono fornire ai fornitori "a valle". Questo può essere un compito oneroso, specialmente nei casi in cui il sistema IA generativo viene utilizzato in contesti diversi e si rendono necessarie documentazioni personalizzate per ciascun caso d'uso. La possibile conseguenza di questi meccanismi di valutazione del rischio e documentazione, sottolineata dai primi commentatori della versione del Parlamento, è una restrizione della concorrenza nel settore a causa degli oneri connessi all'adeguamento legislativo[51].

In conclusione, può essere osservato che il bilanciamento tra interessi contrapposti si realizza, nella versione del Parlamento, con una regolazione stratificata che pone i principi etici come caposaldo della regolamentazione e l'allocazione degli obblighi connessi al loro rispetto come modalità operativa di normazione. Nell'intenzione del legislatore, la distribuzione di oneri lungo la catena di sviluppo di FM e GPAI consente di monitorare progressivamente l'allineamento tra le azioni intraprese dagli attori di mercato nello sviluppo dei loro prodotti e il rispetto dei diritti fondamentali. Il bilanciamento sembra realizzarsi tramite misure *preventive*, ossia anticipatorie di

[51] P. HACKER, A. ENGEL, M. MAUER. *Regulating ChatGPT and other large generative AI models*. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1112-1123.

potenziali danni, in cui il *design* e la prospettiva *ex ante* sono funzionali a non far sorgere l'esigenza di decidere caso-per-caso o tecnologia-per-tecnologia. La resilienza di questo apparato normativo sarà da verificare in concreto mediante due elementi-chiave: da un lato, la predisposizione di meccanismi di ristoro per gli interessati che non si riterranno soddisfatti delle misure adottate dagli attori di mercato in adeguamento all'AI Act; dall'altro, dalla capacità degli interpreti di interpretare, anche alla luce dei principi ispiratori della normativa, se e fino a che punto gli sviluppi tecnologici manterranno solido il bilanciamento.

6. Considerazioni finali

Questo breve saggio ha illustrato le criticità giuridiche legate all'uso di LLM e di applicazioni basate su essi, con riferimento, in particolare, ai temi della protezione dei dati personali. Se, da un lato, gli sviluppi tecnologici sull'uso dei LLM sono drammaticamente veloci, dall'altro la normativa prova a tenere il passo, utilizzando strumenti d'urgenza come nel caso del provvedimento del Garante Privacy, o introducendo regole *in itinere* per evitare lacune genetiche delle norme di prossima approvazione, come nel caso dell'AI Act. Il rischio di questa produzione provvedimentale e normativa è la difficoltà nell'individuare regole resilienti per regolare una tecnologia "disruptive", mutevole e che vive, oggi, un momento di fermento tecnico ed economico, come quella dei LLM e dell'IA generativa in generale. Si rende necessario trovare metodi tecnologicamente neutri per porre costantemente interrogativi sulla correttezza sostanziale del bilanciamento tra interessi economici e diritti fondamentali. In questo, i principi individuati dall'Unione e che tracciano una "via europea" allo sviluppo di sistemi di IA possono assurgere al ruolo di guida fondamentale. Su questi profili ermeneutici, occorre evidenziare come il bilanciamento tra i principi etici e gli obblighi operativi imposti dalla normativa nei quattro livelli identificati sia in linea con la ricerca di un equilibrio tra innovazione tecnologica e il rispetto della dignità umana.

Tale bilanciamento, che in concreto si è già riverberato nella risoluzione del conflitto tra libertà di iniziativa economica, libertà d'espressione e tutela dei dati personali, deve essere finalizzato ad assicurare che i progressi nell'IA generativa siano guidati dai valori dell'ordinamento, preservando la centralità dell'essere umano. Il Gruppo di Esperti ad alto livello sull'IA istituito dalla

Commissione Europea aveva già osservato nel 2019 che "l'IA non è un fine in sé, ma piuttosto un mezzo promettente per aumentare il benessere umano". La proposta avanzata dal Gruppo di Esperti nelle proprie raccomandazioni etiche orientate alla realizzazione di una "IA affidabile" è ispirata dallo spirito kantiano, dall'etica deontologica e dalla promozione della dignità umana come fondamento essenziale della regolamentazione dell'IA[52].

Lo sviluppo dell'IA generativa non può che essere soggetto all'imprescindibile condizione normativa di rispetto della dignità umana e della centralità dell'uomo. In quest'ottica, la normativa e la sua interpretazione non può sottostimare la necessaria concordanza tra sviluppo tecnologico e dignità umana, elemento necessario per una florida inclusione nel tessuto sociale di queste tecnologie. Come emerso dalla vicenda giuridica che ha coinvolto ChatGPT, è prevedibile che l'IA possa incidere sulla rappresentazione dell'identità personale quando tecniche generative – come ChatGPT - sono utilizzate per creare contenuti sintetici. Si pone un problema di tutela di una situazione giuridica soggettiva rilevante quando tali rappresentazioni sintetiche non riflettono adeguatamente l'identità personale dei soggetti coinvolti[53]. Soltanto un'IA generativa in grado di tutelare le identità personali immateriali dai rischi di abuso e manipolazione della realtà potrà raggiungere quell'affidabilità che costituisce la finalità regolamentativa delle istituzioni europee.

[52] V. N. PETIT, J. DE COOMAN J, *Models of law and regulation for AI. The Routledge Social Science Handbook of AI*, 2021, Routledge, Londra, pp. 199-221.

[53] Sugli intrecci tra dignità personale, identità e tutela dei dati personali, v. L. FLORIDI, *The ontological interpretation of informational privacy*. Ethics of Information Technology, 2005, 7, 4 pp. 185–200; M. MARTONI, E. PATTARO, *Le idee di identità e di identità personale nei presocratici e in diritto italiano*, 2021, Giappichelli Editore, p. 69.