NOTE IN PRIMA LETTURA SULLE DISPOSIZIONI PENALI DEL C.D. DDL "INTELLIGENZA ARTIFICIALE"*

Fabio Coppola**

In attesa della pubblicazione dell'"AI Act" nella Gazzetta Ufficiale dell'Unione Europea, il 23 aprile 2024 il Consiglio dei Ministri ha licenziato il testo del DDL contenente "Disposizioni e delega al Governo in materia di intelligenza artificiale".

All'art. 1 del disegno di legge, rubricato "Finalità e ambito di applicazione", si definiscono gli scopi del progetto normativo, riassumibili con tre "keywords":

a) delineare i "principi in materia di ricerca, sperimentazione, sviluppo, adozione e applicazione dei sistemi e modelli di intelligenza artificiale";

b) promuovere "un utilizzo corretto, trasparente e responsabile, in una dimensione antropocentrica" dell'IA;

c) garantire "la vigilanza sui rischi economici e sociali e sull'impatto sui diritti fondamentali dell'intelligenza artificiale.

All'art. 2 il testo chiarisce altresì quale sia il sistema di intelligenza artificiale preso in considerazione dalla disciplina, intendendosi con esso "un sistema automatizzato progettato per funzionari con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare un output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali". Sembra, dunque, conformarsi alla più recente versione dell'"AI Act" Europeo[1] e rivolgersi in primo luogo alla c.d. IA forte, ossia quella capace di auto-apprendere e di produrre autonomamente *output* decisionali sulla base dell'addestramento ricevuto[2].

Passando alle "Disposizioni penali", contenute nell'art. 25 del DDL, si propone di introdurre tra le circostanze aggravanti comuni "l'avere commesso il fatto mediante l'impiego di sistemi di

* Il contributo riproduce, con alcuni adeguamenti bibliografici, l'intervento al Seminario "*L'evoluzione del Bullismo dalla Scuola al Digitale. Esplorando attraverso il Cinema l'Intenzionalità dietro le Azioni*", tenutosi presso l'Aula 1 del Dipartimento di Scienze Giuridiche, Università degli Studi di Salerno, il 14/05/2024.

** Ricercatore di Diritto Penale presso il Dipartimento di Scienze Giuridiche dell'Università degli Studi di Salerno.

[1] In tal senso, cfr. D. De Pasquale - M. Pappalardo, *Intelligenza artificiale: prime osservazioni sul ddl sull'utilizzo dei sistemi di A.I.*, in *Il Quotidiano Giuridico*, 30 aprile 2024.

[2] Per una dettagliata analisi e definizione, si rimanda alla tradizionale opera di S.J. Russell – P. Norvig, *Artificial Intelligence. A modern approach*, IV ed., London 2022, 1032 ss.

intelligenza artificiale, quando gli stessi, per la loro natura o per le modalità di utilizzo, abbiano costituito mezzo insidioso, ovvero quando il loro impiego abbia comunque ostacolato la pubblica o la privata difesa, ovvero aggravato le conseguenze del reato".

Vengono poi previste una serie di circostanze aggravanti per alcune specifiche fattispecie, tutte fondate sull'"impiego dei sistemi di intelligenza artificiale". Così, il delitto di truffa è procedibile d'ufficio[3] ed è punito con una pena da 1 a 5 anni e della multa da euro 309 a euro 1.549 di reclusione in ragione dell'utilizzo dei sistemi di IA per commettere il fatto. La frode informatica è punita con una pena da 2 a 6 anni di reclusione e della multa da euro 600 a euro 3.000 se il fatto è commesso mediante sistemi di IA. L'utilizzo dell'IA diventa un fattore aggravante anche per le fattispecie di riciclaggio, 'reimpiego', autoriciclaggio, aggiotaggio e la manipolazione del mercato. Viene infine incriminata con la multa da euro 51 a euro 2.065 la riproduzione ed estrazione di "opere o altri materiali in violazione (…) [del diritto di autore], anche attraverso i sistemi di intelligenza artificiale".

La disposizione che presenta maggiore interesse è la proposta di introduzione dell'art. 612-*quater* c.p. – "Illecita diffusione di contenuti generati o manipolati artificialmente".

La collocazione all'interno dei Delitti contro la libertà morale e subito dopo l'art. 612 *ter* c.p. – "Diffusione illecita di immagini o video sessualmente espliciti" – sembra evidenziare l'interesse tutelato dalla fattispecie in quello di evitare la circolazione abusiva di contenuti dannosi, generati o manipolati tramite gli strumenti di intelligenza artificiale e capaci di indurre in errore circa la loro genuinità[4].

Verifichiamo se il "design" della fattispecie è riuscito nell'intento sopra enunciato.

Si tratta di un reato comune e di evento, in quanto incrimina "chiunque cagiona ad altri un danno ingiusto".

Prevede poi una particolare nota modale con cui deve prodursi tale danno, ossia "mediante invio, consegna, cessione, pubblicazione o comunque diffusione di immagini o video di persone o di cose ovvero di voci o suoni in tutto o in parte falsi, generati o manipolati mediante l'impiego di

[3] L'aggravante non rientra, infatti, tra quelle procedibili a querela in ragione della c.d. riforma Cartabia. In argomento, cfr. G. L. Gatta, *L'estensione del regime di procedibilità a querela nella riforma Cartabia e la disciplina transitoria dopo la l. n. 199/2022*, in *Sistema penale* 1 (2023).

[4] Per una completa ricostruzione della fattispecie, si rimanda, per tutti, ad A.R. Castaldo, *Il delitto di diffusione illecita di immagini o video sessualmente espliciti ex art- 612-ter c.p. e i profili più critici, in Iura and Legal Systems* 1 (2021) 49-53 e al lavoro monografico di E. Lo Monte, *L'art. 612-ter c.p. Diffusione illecita di immagini o video sessualmente espliciti. Tra buoni propositi, denegato diritto all'oblio e morti "social"*, Torino 2021, 1-185. Sul carattere plurioffensivo della fattispecie, rimandiamo alle considerazioni di M. Bianchi, *L'incriminazione del "revenge porn": il nuovo delitto di "diffusione illecita di immagini o video sessualmente espliciti"*, in A. Cadoppi, S. Canestrari, A. Manna, M. Papa (cur.), *Diritto penale*, III, Milano 2022, 6568-6569.

sistemi di intelligenza artificiale, atti a indurre in inganno sulla loro genuinità o provenienza".

Il delitto è punibile a querela, salvo che il reato sia connesso con uno procedibile d'ufficio o sia commesso ai danni di persona incapace, per età o per infermità, o di una pubblica autorità a causa delle funzioni esercitate. Il *range* edittale da 1 a 5 anni di reclusione consente l'accesso all'art. 131 *bis* c.p. non essendo il reato, a differenza del suo 'simile' (art. 612 *ter* c.p.), incluso tra i reati ostativi alla concessione del beneficio della particolare tenuità del fatto.

Terminato questo breve *excursus* sulla struttura della fattispecie, ritorniamo, approfondendole, sulle modalità della condotta.

Pare a chi scrive che il redattore abbia inteso ancorare la punibilità del danno ingiusto a quattro distinti elementi, che il giudice sarà chiamato ad accertare:

i) la circolazione (mediante le modalità descritte) del contenuto dannoso;

ii) la falsità totale o parziale del contenuto;

iii) la produzione o manipolazione del contenuto tramite strumenti di IA;

iv) la capacità del contenuto ad indurre in errore circa la sua genuinità.

Il legislatore ha infatti richiamato i precedenti requisiti in una *consecutio* temporale separata esclusivamente dalla virgola, costruendo pertanto una fattispecie a 'più cifre', dove cioè ciascun elemento richiamato va accertato singolarmente e autonomamente affinché possa ritenersi perfezionata. Pertanto, la falsità totale o parziale del contenuto non esaurirà il terreno di indagine dell'interprete. Si dovrà altresì accertare che il contenuto, oltre ad essere falso, sia stato generato o manipolato da un sistema di IA e che abbia una idonea capacità decettiva.

Proviamo ora a sottoporre la struttura della fattispecie ad alcuni *test* di operatività concreta per meglio delineare l'ambito di potenziale applicazione del reato.

Immaginiamo, ad esempio, la condotta di colui che invii ripetutamente alla vittima dei video generati con l'IA in cui, tramite il c.d. "deep fake", la persona offesa appaia in atteggiamenti ridicoli e poco decorosi che in realtà non ha mai assunto. Se, per l'abitualità della condotta molesta, si provocasse nella persona offesa il danno ingiusto consistente nel "perdurante e grave stato di ansia o di paura", ovvero "un fondato timore per l'incolumità propria o di un prossimo congiunto o di persona al medesimo legata da relazione affettiva", oppure lo si costringesse "ad alterare le proprie abitudini di vita", l'autore risponderebbe del più grave reato di "stalking". In questa ipotesi, dunque, l'offesa sarebbe già ampiamente presidiata, ancorché realizzata tramite strumenti di intelligenza artificiale.

Facciamo un altro esempio.

Immaginiamo, come avviene durante la nota trasmissione satirica "Striscia la notizia", che il personaggio pubblico del momento venga riprodotto, sempre grazie agli strumenti di IA, in situazioni goffe e/o imbarazzanti, ma immediatamente percepibili come fotomontaggi o video-collage e trasmesse, grazie alla televisione, a un elevato numero di spettatori.

Opportunamente, in questo caso, la fattispecie non troverà applicazione in quanto si tratterebbe di un prodotto digitale incapace di indurre in errore circa la sua genuinità e comunque 'coperto' (entro i suoi limiti) dalla scriminante del diritto di satira.

Affacciamoci infine su uno scenario *pop* di stringente attualità.

Molti dei lettori avranno visto la serie "Netflix" campione di incassi "Baby Reindeer", in cui il protagonista racconta il proprio reale trascorso di vita in cui sarebbe stato vittima di "stalking" da parte di una donna.

Ebbene, quale corto-circuito dell'interesse (rectius: morbosa ossessione) scatenato dal successo dello "show", molti utenti sono andati alla ricerca della protagonista "Martha", che nella "fiction" interpreta la presunta "stalker", trovandola nella vita reale. Essendo la sua identità ormai di dominio pubblico, di recente quest'ultima si è anche fatta intervistare per rilasciare la propria versione dei fatti.

Ma non è questo il peggio.

Alcuni hanno creato sui "social network" "Facebook" e "TikTok" dei finti profili apparentemente riconducibili alla donna in questione, attribuendole dei "post" da loro creati e contenenti frasi altamente infamanti, che confermerebbero l'indole "borderline" della donna. Ora, il tema è: questa grave aggressione alla vita privata e alla reputazione della donna rientrerebbe nel cono di operatività della fattispecie di cui all'art. 612-quater c.p.?

La risposta positiva dovrebbe essere quasi scontata se il fuoco d'offensività, come pare, ruota intorno al danno che generano le "fake news" generate o trasmesse grazie agli strumenti di intelligenza artificiale.

Eppure, la soluzione potrebbe non essere così scontata. Ad una più attenta analisi, infatti, nel caso *de quo*, sicuramente sarebbe realizzato il danno ingiusto alla reputazione della donna. Potrebbe, invece, difettare, il requisito che pretende che il contenuto "fake" sia "generato o manipolato mediante l'impiego di sistemi di intelligenza artificiale". A ben vedere, infatti, la manipolazione sarebbe opera dell'uomo e lo strumento digitale costituirebbe esclusivamente il megafono tramite il quale dare "voce" al contenuto offensivo. Difetterebbe, dunque, la "matrice IA" che rappresenta uno degli elementi di tipicità della fattispecie.

Ciò non vuol dire che questa condotta andrebbe esente da responsabilità. Si tratterebbe pur sempre di sostituzione di persona (pena della reclusione fino ad 1 anno) e diffamazione aggravata col "mezzo della stampa o qualsiasi altro mezzo di pubblicità" (pena della reclusione da 6 mesi a 3 anni o della multa non inferiore a 516 euro), al quale oggi la giurisprudenza serenamente riconduce i "social media"[5]. A seconda della tipologia e modalità di utilizzo dei dati personali, potrebbe contestarsi altresì l'art. 167 del Codice della privacy.

Resta, tuttavia, il dato dell'apparente incapacità della fattispecie di 'colpire' talune delle condotte più riprovevoli.

La evidente congestione in cui si innesterebbe l'art. 612-*quater* c.p. e il residuale spazio applicativo che ne deriverebbe, ci consentono, in conclusione, di porre una domanda: era davvero necessario?

Probabilmente, il centro di interessi preso di mira dal legislatore merita una tutela rafforzata. L'esposizione alla viralità di contenuti c.d. "deep-fake", grazie alle potenzialità dell'IA, potrà in futuro rappresentare un problema tremendamente serio e da non sottovalutare. Sono note le immagini circolate sul *web* di uno dei recenti Presidenti degli Stati Uniti d'America tratto in arresto dalla Polizia. Immagini, tanto verosimili nella rappresentazione, quanto false nel contenuto che, ad un occhio meno esperto o aduso al "fact checking" potrebbero generare la convinzione che quei fatti (id est: l'arresto del Presidente degli Stati Uniti!) sia davvero avvenuto.

Allora la questione si sposta dall'*an* dell'incriminazione al *quomodo*. Da tale angolazione, la tecnica redazionale non convince del tutto. Se il senso fosse quello di evitare la circolazione di contenuti non veritieri e dannosi, si potrebbe costruire l'incriminazione sulla condotta di colui che metta in circolazione un contenuto "fake" credibile, provocando un danno ingiusto, prescindendo dalle modalità di realizzazione uomo/macchina. Sarebbe il requisito della capacità decettiva – che suggeriremmo di mantenere - a selezionare la condotta penalmente rilevante, anziché lo strumento utilizzato.

Al più, l'utilizzo dei sistemi di IA, che effettivamente permette una più agevole moltiplicazione dei contenuti dannosi e una circolazione più spedita, potrebbe costituire una circostanza aggravante.

Infine, si potrebbe consentire una tutela *in extremis* del bene giuridico tramite una condotta di segno contrario al reato, prevedendo una ipotesi di non punibilità sopravvenuta o, più prudentemente,

[5] Una delle prime pronunce in tal senso della giurisprudenza di legittimità si rinviene in Cass. Pen., Sez. I, 02/01/2017, n. 50.

di attenuazione della pena per colui che prontamente elimini il contenuto, evitando che raggiunga livelli di viralità estremamente gravosi.