T

Veronica Neri

Intelligenza artificiale generativa, *deepfakes* e identità vulnerabile. L'etica dell'incertezza come risposta a un rischio (in)controllabile

1. *Premessa*

Il 30 novembre 2022 i mass media annunciano al grande pubblico la messa a punto di sistemi di intelligenza artificiale (IA) generativa. L'IA sta vivendo la sua primavera e ha acquisito in poco tempo un posto di primo piano nell'arena pubblica, non solo tra esperti e mondo della ricerca. Pochi lustri prima, intorno agli anni '90 del secolo scorso, si assiste, invece, in molteplici ambiti disciplinari, a un altro cambiamento che ha permeato la società, la così detta *iconic turn*, ancora in corso, e che ha interessato anche i sistemi di IA generativa.

La svolta algoritmica e la svolta iconica insieme hanno dato vita a sistemi di IA altamente performanti e performativi di generazione di immagini e *deepfake*[1].

Un simile scenario si incardina nella società globale del rischio come definita da Beck[2]. Le innovazioni nel campo dell'IA visiva, come ogni invenzione non solo tecnologica che si è susseguita nel corso del tempo, hanno generato opportunità, ma anche rischi, pericoli e minacce. L'incertezza sulle implicanze del loro impiego sugli individui, sulla comunità e sulla società in generale hanno aperto la porta a nuove sfide etiche.

Si tratta di sistemi in grado di agire simulando – senza comprenderne il senso – i processi mentali degli individui, dipendendo solo in parte dalla

[1] K. Hill, *La tua faccia ci appartiene*, Orville Press, Milano 2024; A. Pinotti, A. Somaini, *Teoria dell'immagine. Il dibattito contemporaneo*, Raffaello Cortina Editore, Milano 2009.

[2] Cfr. U. Beck, *World Risk Society*, Polity Press, Cambridge 1999 (*La società globale del rischio*, trad. di W. Privitera, Carocci, Roma 2013); Id., *Conditio humana. Il rischio nell'età globale*, Laterza, Roma-Bari 2011.

 DOI: 10.4454/5amq3312

volontà umana. Aprono a nuove preoccupazioni circa la salvaguardia della propria identità, la *privacy*, la tracciabilità dei dati, la manipolazione delle decisioni umane, discriminazioni strutturali, l'(in)transparenza delle procedure fino alla creazione e propagazione di nuovi immaginari sociali e/o al rafforzamento di vecchi intrisi di *bias*, incidendo sulle nostre scelte etiche, estetiche, pubbliche e informative[3].

Se il rischio – e il pericolo – connaturato a tali sistemi visivi allenta, fino a annullare, il nostro controllo, occorre ripensare il concetto di rischio alla luce di tali cambiamenti tecnologici. Sembra un rischio da intendersi non solo come un qualche cosa di (anche solo parzialmente) calcolabile[4] – sulla scia di quanto la più recente regolamentazione europea presuppone –, quanto di un pericolo, ovvero della probabilità che un evento, in un arco temporale definito, si verifichi indipendentemente dalla decisione umana, e di incertezze, eventualità non pronosticabili né misurabili. Come nella realtà oggettuale non possiamo calcolare la probabilità che il rischio di frodi o manipolazioni avvenga con una certa regolarità e frequenza statistica, così nella dimensione aperta dall'IA simili eventi possono solo estendersi in ragione dell'aumentare delle possibilità offerte dagli strumenti tecnologici che li consentono, ma non possiamo controllarli né quantificarli statisticamente; né possiamo calcolare l'impatto effettivo dell'IA generativa di immagini sulla creazione di nuovi immaginari sociali in termini probabilistici[5].

Il rischio svela pertanto gli stati di incertezza e di vulnerabilità – recepita come la predisposizione a subire un danno – ai quali è esposta l'umanità di fronte alle tecnologie artificiali. Non possiamo adottare dunque il principio di probabilità e calcolo razionale, riconducendo tutto a una impostazione positivista *tout-court* ovvero ai fattori tecnici insiti nel *design* stesso del sistema, peraltro anch'essi permeati dalla soggettività di chi lo realizza. Con il presente contributo si vuole mostrare come occorra ricalibrare il concetto di rischio dal punto di vista dell'etica non solo individuale, ma soprattutto pubblica e sociale, poiché pertiene la società nella sua complessità, includendo tutte le possibili categorie umane senza discriminazioni al medesimo tempo.

Alcune strategie di etica pubblica possono indirizzare l'opinione pubblica e la cittadinanza alla consapevolezza e alla co-responsabilità di tutti i soggetti coinvolti, dagli ideatori, ai governi (che debbono prendersi in carico

[3] https://www.agendadigitale.eu/cultura-digitale/cose-il-rischio-cosi-la-filosofia-ci-aiuta-a-capire-il-senso-dellai-act/

[4] G. Sturloni, *La comunicazione del rischio per la salute e per l'ambiente*, Mondadori, Milano 2018, pp. 5-10.

[5] C. Taylor, *Gli immaginari sociali moderni*, trad. it. P. Costa, Booklet, Milano 2005.

regolamentazioni specifiche) fino agli utilizzatori. Ciò per indebolire lo stato di incertezza e di paura nei quali il rischio getta i soggetti in generale e quelli più vulnerabili in particolare, cercando di renderli quantomeno più informati, consapevoli e critici.

Come scrive Lagadec rispetto ai primi dispositivi tecnologici con l'IA generativa di immagini siamo di fronte alla nozione del «rischio tecnologico maggiore» poiché la fragilità dei sistemi e i pericoli che fanno correre agli esseri umani aumentano il senso di vulnerabilità dei medesimi e diminuiscono il senso di fiducia nei confronti di una società ossessionata dalla sicurezza e dal controllo[6].

Il contributo si articolerà in tre parti. Dopo una introduzione sull'immagine artificiale e le *deepfakes* si affronterà il tema della vulnerabilità dell'individuo scontrandosi con *bias*, allucinazioni, protezione dei dati e il rischio di identità plurime, fasulle e incontrollabili; la terza e ultima parte affronterà il tema del rischio e dell'etica dell'incertezza rispetto alle immagini artificiali, cercando di proporre spunti di riflessione alle minacce etiche emergenti.

2. *Intelligenza generativa di immagini e il caso delle* deepfakes

La comunicazione visiva risulta sempre più supportata dall'IA. Questa tendenza ha, da una parte, potenziato alcuni ambiti disciplinari, dall'altra, ha imposto una verifica dei contenuti digitali a chi ne fruisce e, dunque, indebolito la fiducia nelle immagini in generale. Numerosi sistemi «che si comportano come se fossero intelligenti»[7], ad esempio ChatGPT per i testi e DALL-E per le immagini, solo per citare i sistemi di IA generativa più noti, hanno contribuito allo sviluppo della diagnostica clinica in ambito sanitario, alla semplificazione dei processi di traduzione, alla creazione di testi in diverse lingue adattati ai pubblici di riferimento[8], o, nel campo della giustizia, ai riconoscimenti facciali, fino all'ambito della creatività nel settore pubblicitario e artistico o al contesto giornalistico (per descrivere fatti e eventi)[9]. Tali strumenti hanno al-

[6] P. Lagadec, *La civilisation du risque. Catastrophes technologiques et responsabilité sociale*, Le Seuil, Paris 1981; G. Liuzzo *et al.*, *The Term Risk: Etymology, Legal Definition and Various Traits*, in «Italian Journal of Food Safety» 3 (1), 2014, p. 2269.

[7] M.R. Taddeo, *Costruire l'etica dell'intelligenza artificiale*, in *Il potere del pifferaio magico*, a cura di G. Fregonara, UPI, Pisa 2021 p. 166.

[8] D. Baidoo-Anu, L. Owusu Ansah, *Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning*, in «Journal of AI»7, 1 (2023), pp. 52-62; A. Barale (a cura di), *Arte e intelligenza artificiale. Be my GAN*, Jaca Book, Milano, 2020.

[9] L. Gaur (eds.), *Deepfake. Creation, Detection and Impact*, CRC Press, Boca Raton 2024,

tresì prodotto immagini e audiovideo tanto realistici quanto falsi indistinguibili da video realizzati con media tradizionali, manipolando l'opinione pubblica e/o ledendo la *privacy* e la reputazione di alcune persone.

Si tratta di sistemi che, attraverso interrogazioni sotto forma di stringhe testuali (prompt) da parte dell'utente, offrono risposte come se fossero individui in carne e ossa, mettendo in scena un vero e proprio dialogo persona-macchina.

Cristianini, in questo scenario, individua tre livelli di azione dell'IA generativa «*l'agente* che incontriamo nel mondo (per esempio ChatGPT), il *modello* interno che questo usa per prendere decisioni (per esempio GPT-3) e l'algoritmo che crea tale modello partendo dai dati (per esempio, il Trasformer). Per modello si intende il modello di mondo preso in considerazione che deve suggerirci quali eventi e situazioni sono probabili e quali non lo sono. Tale modello plasma solo una parte di mondo alla volta e consente di interagire con il mondo stesso divenendo una forma di comprensione del mondo stesso»[10]. Lo stesso accade quando la risultante della stringa è una immagine. Si tratta di una immagine generata, rispetto a quelle tradizionali frutto dell'immaginazione e della fantasia umane, «grazie a una specifica capacità di uni-formare» degli apparati[11]. Questa affermazione flusseriana, ancora attuale per la specifica tipologia di immagini tecniche che sono le immagini artificiali, mette in luce la capacità algoritmica di generare segni visivi efficaci, realistici o surreali come se fossero la risultante di una intensa attività immaginativa, senza però esserlo a pieno titolo, senza anzi conoscere come si sia arrivati al risultato e chi lo abbia prodotto, l'essere umano e/o la macchina. L'immaginazione algoritmica – se ad essa si può fare appello – risulta ad oggi ben lontana da quella umana[12]. L'immagine tecnica può definirsi, riprendendo Flusser, quale una immagine «generata da un apparato "artificiale", a seguito del predominio graduale del modello algoritmico basato su pixel, che ricostruisce una unità attraverso segni puntiformi»[13]. Occorre pensare alle immagini tecniche non tanto come «tentativi dell'es-

p. 91 ss.; M. Filimowicz (eds.), *Deep Fakes: Algorithms and Society*, Routledge, London 2022; C. Canali, R. Pedrazzi, *L'opera d'arte nell'epoca dell'intelligenza artificiale*, Jaka Book, Milano 2024.

[10] N. Cristianini, Machina sapiens. *L'algoritmo che ci ha rubato il segreto della conoscenza*, il Mulino, Bologna 2024, p. 35.

[11] V. Flusser, *Immagini. Come la tecnologia ha cambiato la nostra percezione del mondo*, Fazi Editore, Roma 2009, p. 15.

[12] E. Finn, *What Algorithms Want. Imagination in the Age of Computing*, MIT Press, Cambridge (MA) 2017; F. Restuccia, *Il contrattacco delle immagini, tecnica, media e idolatria da Vilém Flusser*, Meltemi, Milano 2021.

[13] V. Flusser, *Immagini*, cit., p. XIII.

sere umano estraniato dal mondo di farsi una immagine di questo mondo», quanto «di conseguenze del progresso scientifico»[14].

Attraverso *Generative Adversarial Networks* (GAN) e i Diffusion models, due dei sistemi più noti ed efficaci di generazione di immagini artificiali, è possibile realizzare immagini, anche artistiche, molto variegate per tipologia, per stile, per soggetti, per fine, ecc.[15]. Ciò che distingue una immagine digitale generata con programmi di computer grafica "tradizionali" dalle produzioni delle intelligenze artificiali è relativo, oltre al risultato estetico, al processo creativo. Il programma, dopo l'input dell'essere umano, assume un ruolo autonomo nella creazione dell'immagine[16].

Una GAN, in sintesi, è composta da due reti avversarie che hanno l'obiettivo di migliorarsi vicendevolmente. Da un lato, abbiamo una rete *generator*, il cui compito consiste nel produrre nuove immagini da un data set quanto più ampio possibile (in cui a ogni immagine è associata una etichetta testuale che ne descrive il contenuto); dall'altro, abbiamo una rete *discriminator*, che deve confrontarsi con i risultati del *generator* e segnalare se l'immagine si discosta (e in che misura) dalla richiesta dell'essere umano, se sembra eccessivamente falsa, se il risultato di una "allucinazione" del sistema, ecc. Nel mostrare un possibile errore o una incongruenza la rete avversaria apprende e permette anche alla rete *generator* di apprendere a propria volta. Come afferma Goodefellow:

> The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistiguishable from the genuine articles[17].

L'immagine finale si raggiunge quando il *generator* crea una nuova im-

[14] M. Menon, *Vilém Flusser e la «rivoluzione dell'informazione». Comunicazione, etica, politica*, Edizioni ETS, Pisa 2011, p. 42; V. Flusser, *Kommunikologie*, Fischer Taschenbuch, Frankfurt 1998, p. 102.

[15] L'invenzione delle GAN è tradizionalmente attribuita a Ian Goodfellow e ai suoi collaboratori: J. Goodfellow *et al.*, *Generative Adversarial Nets*, ArXiv:1406.2661 [Cs, Stat], June 2014: https://arxiv.org/pdf/1406.2661.pdf; M. Jovanović *et al.*, *Generative Artificial Intelligence: Trends and Prospects*, in «Computer» 55, 10 (2022), pp. 107-112. Relativamente ai Diffusion Models, cfr. A. Bordas, *What is generative in generative artificial intelligence? A design-based perspective*, in «Research in Engineering Design», 35 (2024), pp. 427-443 e H. Cao *et al.*, *A survey on generative diffusion Model*, arXiv:2209.02646

[16] A. Barale, *Arte e intelligenza artificiale: alcune domande*, in Ead. (a cura di), *Arte e intelligenza artificiale. Be my GAN*, Jaca Book, Milano, 2020, pp. 7-18.

[17] I.J. Goodfellow *et al.*, *Generative Adversarial Nets*, cit., p. 1.

magine che viene percepita dal *discriminator* "autentica"[18].

I più recenti Diffusion models sono invece fondati su un processo di diffusione che trae origine da un prompt testuale e da un dataset di coppie (image, caption). Viene applicato rumore casuale (noising process) alle immagini di addestramento e, successivamente, viene appresa la funzione inversa di denoising, la quale cerca di invertire il processo iniziale e di ricostruire un'immagine condizionata da un input testuale (plausibilmente compatibile ad esso e coerente con le sue richieste). Nel corso dell'addestramento il modello impara a prevedere il rumore aggiunto a una immagine a ogni passo del processo di diffusione per poterlo poi sottrarre correttamente nel reverse process. Il modello apprende pertanto, passo dopo passo, relazioni generali tra immagini da generare e prompt[19].

Nel 2018 presso la casa d'aste Christie's viene venduto il ritratto di Edmond de Belamy, un'opera realizzata dall'IA generativa e da tre sperimentatori del collettivo parigino *Obvious*. Nel medesimo anno l'artista Klingemann realizza con l'IA generativa le *Memories of Passersby I*, una serie di ritratti di identità inventate. In tempi ancora più recenti la fotografa italiana Zanon ha pubblicato uno pseudo-reportage sulla guerra in Ucraina attraverso la piattaforma Midjourney, mentre il fotografo tedesco Eldagsen partecipa al premio "Sony World Photography Awards" con una immagine realizzata interamente con IA. Come scrive Cohen nel caso dell'IA «[l]a creatività non risiede né nel programmatore né nel programma, ma nel dialogo tra programma e programmatore»[20]. Emerge un sistema di co-autorialità non semplice da gestire dal momento che una collaborazione e cooperazione pienamente consapevole tra essere umano e macchina richiama «the lack of a common language between AI and humans»[21].

Il sistema genera immagini combinando insieme segni visivi di un data set

[18] Esistono diverse tipologie di GAN differenti per l'architettura della rete e per come vengono addestrate. Cfr. Z. Wang, Q. She, T.E. Ward, *Generative Adversarial Networks: A Survey and Taxonomy*, in «arXiv:1906.01529» 4 june (2019). Sul concetto di autenticità, cfr. C. Taylor, *The Ethics of Authenticity*, Harvard University Press, Cambridge (MA) 1992; C. Guignon, *On Being Authentic*, Routledge, London 2004; Id., *Authenticity*, in «Philosophy Compass» 3(2), 2008, pp. 277-290.

[19] Cfr. A. Bordas, *What is generative in generative artificial intelligence?*, cit.; O. Sanseviero *et al.*, *Hands-On Generative AI with Transformers and Diffusion Models*, O'Reilly Media, Sebastopol 2024; https://www.agendadigitale.eu/cultura-digitale/creare-immagini-dallimmaginazione-il-potere-dei-modelli-di-diffusione/.

[20] Cit. tratta nell'intervista pubblicata in V. Tanni, *Arte e intelligenza artificiale. Una storia che inizia negli Anni Cinquanta*, Artribune, 30/06/2023: https://www.artribune.com/progettazione/new-media/2023/06/arte-intelligenza-artificiale-storia/.

[21] Cit. di Ali Nikrang in G. Stocker, M. Jandl, A.J. Hirsch (eds.), *The Practice and Art and AI*, Ars Electronica. Ars, Technology, Society, Linz 2023, p. 30.

iconografico tanto ampio da non poter essere 'contenuto' da nessuna mente umana. Possono essere rappresentate relazioni iconiche impensate, unicamente frutto dell'autonomia del sistema stesso. Il risultato può sembrare molto efficace ed esteticamente convincente, per espressività, per l'iper-realtà e/o per bellezza, ma altresì fuorviante, risultato di allucinazioni[22]. Si profilano al contempo sfide di ordine etico su possibili rischi emergenti, come vedremo più avanti, inerenti la distorsione di immaginari sociali, la generazione di *bias* e pregiudizi di genere, etnia e professione, una certa uniformazione della 'creatività', fino ad usi dichiaratamente malevoli come il furto di identità, l'utilizzo (non accordato) dei dati personali e la perdita di privacy.

Nell'alveo del visivo artificiale si inscrivono, infine, le *deepfakes*. Il termine *deepfake* richiama l'unione di *deep*, che evoca i sistemi di *deep learning* (DL), e *fake*, falso. Il sistema che le realizza impiega algoritmi di DL per produrre e modificare immagini, video e audio e generare un media sintetico/falso[23]. Non pertiene più solo immagini statiche, ma video e/o audiovideo che si insinuano in circuiti "intrasparenti". Il sistema algoritmico che ne sta alla base consente in pochi passaggi di creare *ex novo* volti di persone o di alterarne di esistenti (modificando, ad esempio, il colore dei capelli, degli occhi o della pelle, solo per citare alcune possibilità) o ancora, di incrociare, sovrapponendoli, più volti insieme per ottenerne uno unico "nuovo". Può essere impiegata altresì per alterare o generare corpi, ambienti, spazi, luoghi, oggetti, ma anche animali e quant'altro si desideri. Può dare vita anche ad audio immaginari[24]. Se ne può fare, nel complesso, un uso benevolo, creando, per esempio, influencer virtuali e video didattici; può ben essere utilizzato per l'assistenza sanitaria e farmaceutica, per realizzare sfilate virtuali a basso costo da vedere su uno schermo o altre applicazioni di *entertainment*, ricostruire fatti e eventi nell'ambito della giustizia o, come accadeva con la VR, ambienti e personaggi storici; ma possono subentrare, di contro, approcci malevoli, non etici, come lo sviluppo di algoritmi di *deep learning* in grado di trasferire volti di celebrità in video pornografici, in atti di *cyberbullismo* o di violenza più in generale, per diffamare o propagandare idee e pensieri fasulli, scorretti, ecc.

[22] G. Finocchiaro, *Intelligenza artificiale. Quali regole*, il Mulino, Bologna 2024, p. 24.

[23] Le sue origini risalgono in realtà al 1997 e all'ideazione di un programma di riscrittura video: cfr. C. Bregler, M. Covell, M. Slaney, *Video Rewrite: Driving Visual Speech with Audio*, in «Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques» 24 (1997), pp. 353-360; N. Schick, *Deepfakes. The Coming Infocalypse*, Twelve, New York 2020.

[24] A. Tversky, D. Kahneman, *Judgment under Uncertainty: Heuristics and Biases*, in D.J. Levitin (ed.), *Foundations of cognitive psychology: Core readings*, MIT Press, Cambridge (MA) 2002, pp. 585-600.

Esemplare il caso di *deepfake* sottoforma di audiovideo, datato 2017, in cui il Presidente degli Stati Uniti d'America Obama, con lo sguardo dritto nella telecamera, pronuncia frasi mai pronunciate (c.d. "effetto perturbante")[25].

Simili processi rendono indistinguibile il falso dall'originale. Il risultato dipende dalle regole delineate nel design del modello, dai big data di riferimento fino ai sistemi di controllo approntati. Appare chiara l'urgenza di generare nuovi algoritmi di rilevamento delle *deepfakes* per smascherare casi di uso malevolo dell'IA, i c.d. *deepfake detector*[26].

Simili contenuti artificiali possono essere ideati, generati e propagati da chiunque ne abbia l'intenzione. Per questo motivo può essere di ausilio una indagine sui possibili rischi e sui pericoli che l'IA generativa di immagini fa correre all'essere umano, attraverso la lente dell'etica, della comunicazione e delle linee di azione pubbliche e politiche.

3. *Manipolazione,* bias*, allucinazioni e il rischio di identità plurime*

Ogni essere umano può essere potenzialmente leso da immagini artificiali o *deepfake*. Impossibile prevedere le probabilità di tale eventualità in termini statistici. Nel nostro ecosistema massmediale, caratterizzato da infocrazia e disinformazione visive, l'IA generativa di immagini costituisce una minaccia in costante sviluppo. Ci espone a minacce difficilmente pronosticabili, che ci inducono a vivere nell'incertezza. È una esposizione che incide sulla nostra visione del mondo e sulla nostra identità e che ha implicazioni nelle diverse dimensioni di vita dell'essere umano, offline, *online* o, come teorizza Floridi, *onlife*[27].

Rischio, pericolo e incertezza appaiono parole chiave sulle quali appuntare brevemente l'attenzione prima di delineare le sfide etiche rilevanti che pertengono, più in generale, le immagini artificiali e più nello specifico, le *deepfakes* e che plasmano la *conditio humana* contemporanea.

[25] Cfr. M. Marini, *Video fake facilissimi da realizzare con un algoritmo: Obama "vittima" eccellente*, in «La Repubblica»: https://www.repubblica.it/tecnologia/2017/07/13/news/video_fake_facilissimi_da_realizzare_con_un_algoritmo_obama_vittima_eccellente-170703930/, 13 luglio 2017 (ultimo accesso 26 agosto 2024)

[26] Y. Li, S. Lyu, *Exposing DeepFake Videos By Detecting Face Warping Artifacts*, in «arXiv:1811.00656v3» (2019); L. Gaur (eds.), *Deepfake. Creation, detection and Impact*, CRC Press 2024, p. 91 ss.; M. Filimowicz, *Deep Fakes: Algorithms and Society*, Routledge, London 2022; N. Schick, *Deepfakes: The Coming Infocalypse*, Twelve, New York-Boston 2020.

[27] L. Floridi, *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*, Oxford University Press, Oxford 2014, p. 4; D.J. Chalmers, *Più realtà. I mondi virtuali e i problemi della filosofia*, Raffaello Cortina Editore, Milano 2023.

Sulla scia di Le Breton, con rischio si intende «una conseguenza aleatoria di una situazione, ma non in termini di una minaccia, di un danno possibile». Si rimanda al termine italiano "risco", forma antica di "rischio", al latino *resecare*, con il significato di rimuovere tagliando e al latino classico *rixare*, "litigare" oltre che a *resecum*, colui che taglia. Anche la parola spagnola *riesgo*, roccia tagliata, scoglio, rimanda al medesimo concetto. Sembra cioè il momento in cui le strade si incrociano superando un prevedibile momento di pericolo, una 'roccia'. Il rischio sembra dunque appartenere alla categoria di incertezza quantificata, un pericolo possibile che può derivare da determinate circostanze ed eventualità, una sorta di misura dell'incertezza. Il rischio rispetto al pericolo lascia all'uomo ancora una responsabilità[28]. Come sottolinea Beck con la metafora della navigazione il rischio si corre ogni qual volta si naviga e ci si espone alla possibilità che uno scoglio possa danneggiare la nave[29]. Con l'AI generativa gli scogli sono frequenti e possono far addirittura affondare la nave.

Luhmann delinea, sulla scia di Knight[30], una distinzione di base tra rischio e pericolo: il primo termine implica una responsabilità soggettiva (dovuta alle scelte dell'essere umano), mentre il secondo è intrecciato a minacce che sfuggono al controllo dell'individuo[31], sebbene possa essere percepito in modo diverso da ciascuna e ciascuno.

L'incertezza si riferisce, invece, a qualche cosa di non prevedibile né quantificabile. Solo nello sviluppo degli eventi e delle circostanze si potrà rivelare o meno un pericolo. Sia Douglas che Luhmann affrontano il concetto sebbene con posizioni differenti. Se Douglas analizza l'incertezza come un fenomeno culturale, controllato attraverso la costruzione sociale delle categorie di purezza e rischio, Luhmann vede l'incertezza come una conseguenza della complessità sociale, che i sistemi devono ineludibilmente affrontare[32].

[28] Sul rischio, cfr.: M. Douglas, *Risk and Blame. Essays in Cultural Theory*, Routledge, London 1992 (*Rischio e colpa*, a cura di G. Bettini, il Mulino, Bologna 1996, pp. 33-34); N. Luhmann, *Risk. A sociological Theory*, Aldine de Gruyter, New York 1993 (*Sociologia del rischio*, trad. it. di G. Corsi, Mondadori, Milano 1996); M. Douglas, A. Wildavsky, *Risk and culture*, University of California Press, Berkeley 1982, D. Le Breton, *Sociologia del rischio*, a cura di A. Romeo, Mimesis 2017, pp. 14-15; P.L. Bernstein, *Against the Gods. The Remarkable Story of Risk*, Wiley, New York 1996.

[29] U. Beck, *La società del rischio. Verso una seconda modernità*, cit.

[30] Cfr. F.H. Knight, *Risk, Uncertainty, and Profit* (1921), Beardbooks, Washington 2002 (*Rischio, incertezza e profitto*, trad. it. di M. Giorda, La Nuova Italia, Firenze 1960).

[31] N. Luhmann, *Sociologia del rischio*, cit., p. 17.

[32] *Ivi*; M. Douglas, *Purity and Danger: An Analysis of Concepts of Pollution and Taboo* (1966), Routledge, London 2022.

Relativamente alle sfide etiche aperte dalle immagini in generale generate dall'IA esse possono, in primo luogo, riflettere lo stato di incertezza e di paura dell'uomo contemporaneo. In specie a causa di *bias* (o *contro-bias*) insiti nei dati di addestramento del sistema. Se l'IA viene alimentata da dati non bilanciati o parziali, potrebbe generare immagini distorte che rinforzano stereotipi o pregiudizi fondati su discriminazioni di genere, religione, etnia, cultura, professione, classe sociale sulla base dell'immaginario sociale di riferimento. Per *bias* [dal gr. "epikáros", dal fr. e provenz. ant. *biais* «obliquo»] non si intende tanto la deviazione sistematica da una norma, quanto le inclinazioni e la predisposizione al pregiudizio[33]. Di fronte a alcune immagini artificiali possono emergere *bias* cognitivi, automatismi o scorciatoie mentali dalle quali si generano credenze (non sempre eticamente orientate) e dalle quali si traggono decisioni veloci. Sono giudizi che, talvolta, riflettono le disuguaglianze sociali della realtà oggettuale e che impattano su decisioni, comportamenti e sviluppo del pensiero in contesti incerti, talaltra, ne generano di nuove. Sulla base del meccanismo dei *confirmation bias*, secondo cui le persone prediligono ricevere immagini che confermano le proprie preferenze e credenze, portando a negare qualsiasi evidenza contraria, tali meccanismi saranno ulteriormente corroborati e riproposti fino a che il sistema viene allenato con tale procedura. Nel mondo pubblicitario, ad esempio, sono spesso rappresentati individui che impersonano certe professioni, fondandosi su stereotipi e pregiudizi di genere, di etnia e sociali propri della realtà oggettuale, sedimentatisi nel tempo. E più questi personaggi sono permeati da stereotipi sociali che suggeriscono caratteristiche valoriali condivise più la pubblicità funziona e diventa efficace. Si contrae il nostro spazio di autonomia e aumenta lo spazio di azione dei 'difetti' algoritmici o pregiudizi (*bias*) iniqui, così perpetuando o esacerbando nuove, esistenti e/o passate logiche discriminatorie e forme di diseguaglianza. Un secondo aspetto, che discende dal precedente, riguarda la disinformazione e la manipolazione e persuasione sociali. Immagini false influenzano le decisioni individuali e l'opinione pubblica, giocando sulla credulonerìa, sulle emozioni e sulla vulnerabilità di chi guarda. Le identità personali degli individui *onlife*, finite nei data set di riferimento, vengono ridotte ad aggregati di dati in vendita.

Un ulteriore pericolo concerne il concetto di co-autorialità che contraddistingue le immagini artificiali. Se da una parte incide sui diritti d'autore e

[33] N. Cristianini, *La scorciatoia. Come le macchine sono diventate intelligenti senza pensare in modo umano*, il Mulino, Bologna 2023, p. 104.

sul copyright[34], intrecciandosi con aspetti di pertinenza giuridica, dall'altra artisti e creativi potrebbero sentirsi minacciati da sistemi che concorrono alla realizzazione di opere d'arte. Il rischio nel quale si può incorrere è di ottenere un prodotto visivo frutto di allucinazioni oppure immagini che si autocensurano sulla base dei valori guida del sistema (impostati durante il design dell'applicazione) oppure, ancora, cadere nel nichilismo, nel non credere più alle immagini che circolano nei massmedia e di ingenerare una crisi di fiducia nella capacità degli individui di distinguere tra realtà e finzione. Questo può avere conseguenze profonde sulla comunicazione visiva in generale, sul giornalismo e sul modo in cui interagiamo con il mondo digitale, ma anche in ambito di giustizia poiché difficile è stabilire l'autenticità delle prove visive.

In sintesi, le immagini create dall'intelligenza artificiale offrono grandi potenzialità comunicative e creative, ma sollevano altresì importanti questioni etiche (e legali) che richiedono una regolamentazione e una riflessione sui principi adeguati alla base dell'IA e di chi la utilizza per mitigarne i rischi e i pericoli.

Sulla generazione di immagini iperrealistiche e *deepfake* che falsificano la realtà rischi e pericoli si moltiplicano. Possono essere utilizzate per ingannare o manipolare l'opinione pubblica in modo più diretto e impattante di mere immagini statiche. Emergono al riguardo questioni etiche riferite alla violazione della *privacy* e dell'identità personale. L'IA può generare immagini di persone che non esistono, utilizzando dati provenienti da fotografie o informazioni personali tratte liberamente online senza consenso del soggetto ritratto. Ciò solleva questioni di *privacy*, poiché i dati visivi possono essere sfruttati per generare nuove identità o per comprometterne altre, senza che le persone coinvolte ne siano consapevoli. Inoltre, possono essere realizzate immagini o offensive e lesive della dignità dell'essere umano senza responsabilità diretta.

Come scrivono Thaler e Sunstein gli algoritmi, anche dunque quelli relativi al visivo, sono i nuovi «architetti della scelta», in grado di rimodellare e "architettare" i contesti e gli ambienti in cui formiamo i nostri gusti e compiamo le nostre scelte e, inoltre, formiamo (o inventiamo nel caso delle *deepfakes*) la nostra identità personale[35].

[34] Noto il caso del dicembre 2023 che vede il New York Time fare causa a OpenAI e a Microsoft per violazione del diritto di autore, ovvero per l'uso non autorizzato di milioni di suoi articoli per l'addestramento di chatbot. Cfr. https://www.ilsole24ore.com/art/new-york-times-fa-causa-openai-e-microsoft-uso-copyright-AFtsWfBC?refresh_ce=1.

[35] S. Tiribelli, *Identità personale e algoritmi. Una questione di filosofia morale*, Carocci, Roma 2023, p. 67; R.H. Thaler, C.R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Penguin Books, London 2009.

L'identità dell'essere umano con l'IA si scontra con alcuni rischi che anche il Regolamento europeo sull'intelligenza artificiale (AI Act) ha cercato di affrontare adottando il *risk based approach*, ovvero la suddivisione di gruppi di sistemi di IA in livelli di rischio. Come scrive Dadà al riguardo[36] nell'*AI Act* europeo il termine 'rischio' «appare più di 350 volte»[37]. Si cerca sia di ottemperare alle esigenze di tutela contro le minacce provocate dai sistemi sia di far evolvere il progresso tecnologico. Il documento propone quattro livelli di rischi, dal più elevato al meno impattante: 1) sistemi a rischio inaccettabile, 2) sistemi ad alto rischio, 3) sistemi con rischio per la trasparenza e 4) sistemi a basso o a minimo rischio[38]. Relativamente ai rischi provocati dalle immagini artificiali e dalle *deepfakes* è il terzo livello quello deputato a tentare una forma di tutela. In ragione della difficoltà a comprendere se sono prodotti realizzati o meno da mano umana la regolamentazione impone l'obbligo di trasparenza sulla loro origine in vista di una IA più affidabile e, dunque, più etica[39].

Relativamente al rischio del terzo tipo le questioni etiche non riguardano tanto aspetti tecnici, quanto culturali, legati all'immaginario sociale e alla nostra identità individuale e sociale.

Rimodulare i data set di riferimenti dai quali l'IA attinge per creare immagini o *deepfake* è un problema pertanto sia culturale e sociale che tecnico. Sono gli esseri umani con i loro principi, valori, cultura e *bias* a fare una selezione dell'archivio iconografico dal quale il sistema attinge o a non farla, lasciando alla rete il ruolo di *hard disk* esterno di riferimento.

Da questa breve disamina affiora un'idea di rischio probabilistica ovvero fondata sulla probabilità del presentarsi di un evento dannoso e della gravità delle conseguenze che esso ha ingenerato ($R = P \times D$)[40].

Nel caso delle *deepfakes* appare chiaro che questa definizione sia difficile da applicare. Il controllo su certi dati e su determinate scelte appare utopistico pur limitando un sistema che ci sfugge sempre più di mano. Con

[36] S. Dadà, *Rischio e Intelligenza Artificiale. Un'analisi concettuale tra razionalità e incertezza*, in «Il pensiero critico» I (2014), pp. 47-66.

[37] https://artificialintelligenceact.eu/ (AI Act, approvato in ultima istanza il 12 luglio 2024).

[38] Approccio presente anche oltreoceano con l'US Algorithmic Accountability Act (2023), già proposto nel 2019 con l'obiettivo di fronteggiare i rischi relativi a possibili discriminazioni e violazioni della privacy in relazione all'utilizzo di sistemi di intelligenza artificiale ed il Canadian Directive of Automated Decision-Making (2020), fondato su principi etici cardine come la trasparenza, la responsabilità, la legalità e l'equità procedurale. Cfr. https://www.congress.gov/bill/118th-congress/senate-bill/2892/text; https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592.

[39] Sul concetto di rischio e IA, cfr. S. Dadà, *Rischio e Intelligenza Artificiale*, cit.

[40] Cfr. D.Lgs 81/2008; UNI EN ISO 12100-11; G. Sturloni, *La comunicazione del rischio*, cit.

le *deepfakes* si incide sui diritti fondamentali dell'essere umano legati alla tutela della propria identità e dei dati personali: la *privacy*, recepita non tanto come «il diritto a uno spazio in cui essere lasciati soli»[41], quanto come *privacy informazionale*, intesa sia come il diritto di impedire ad altri l'accesso alle nostre informazioni personali, sia come il diritto alla tutela della propria identità personale e della propria immagine *onlife*[42]. Il Regolamento (UE) 2016/679, ad esempio[43], ha affrontato in passato gli aspetti legati alla tutela del flusso dei nostri dati personali (*data protection*) e al controllo delle nostre informazioni (*privacy* informazionale).

Con le *deepfakes* siamo dunque di fronte a rischi di furto o distorsione malevola di identità incalcolabili, quindi all'incertezza estrema. Le minacce in corso possono ledere la dignità e i diritti fondamentali dell'essere umano[44].

Conclusioni. Dal rischio ad un'etica dell'incertezza

Per eludere la paura radicale, oramai ontologica, dell'individuo nella modernità[45] potremmo convenire con l'affermazione di Luhmann secondo il quale «se ci si astiene da una certa azione non si corre alcun rischio»[46]. Sulla base di questa opzione estrema l'inazione parrebbe la soluzione che mette a riparo da ogni possibile rischio e pericolo. Nel caso dell'IA generativa approccio luhmanniano potrebbe risultare anacronistico e poco efficace. La posizione di Giddens, secondo il quale «l'inazione è sovente rischiosa e vi sono alcuni rischi che, volenti o nolenti, noi tutti dobbiamo correre» rappresenta piuttosto la risposta consapevole ad una tecnologia che oramai pervade la vita degli esseri umani e della quale dobbiamo farci carico[47]. È nel dominio dell'incertezza e del pericolo ai quali siamo costantemente

[41] S. Warren, L.D. Brandeis, *The Right to Privacy*, in «Harvard Law Review» 4, 1980, pp. 193-194.

[42] S. Tiribelli, *Identità personale e algoritmi*, cit., p. 12; L. Floridi, *The Ontological Interpretation of Informational Privacy*, in «Ethics and Information Technology» 7, 2005, pp. 185-200; Id., *The Informational Nature of Personal Identity*, in «Minds and Machines» 21, 4, 2011, pp. 549-566; C. Koopman, *How We Became Our Data: A Genealogy of the Informational Person*, University of Chicago Press, Chicago 2019.

[43] https://www.garanteprivacy.it/il-testo-del-regolamento.

[44] S. Tiribelli, *Identità personale e algoritmi*, cit.

[45] A. Giddens, *Le conseguenze della modernità*, il Mulino, Bologna 1994.

[46] N. Luhmann, *Familiarità, confidare e fiducia: problemi e alternative*, in D. Gambetta (a cura di), *Le strategie della fiducia. Indagini sulla razionalità della cooperazione*, trad. it. di D. Panzieri, Einaudi, Torino 1989, p. 130.

[47] A. Giddens, *Le conseguenze della modernità*, cit. p. 41.

esposti che dobbiamo ricercare il senso del rischio al tempo dell'IA generativa, dal momento che l'autonomia morale e decisionale dell'individuo risulta indebolita e le procedure delle scelte algoritmiche appaiono opache. Certamente «calcolare l'incalcolabile» non offre risposte utili relativamente a come agire[48]. Non è possibile misurare la probabilità oggettiva che un dato evento accada o una certa immagine venga prodotta, poiché la tecnologia e i data set sono in costante evoluzione e procedono repentinamente.

Piuttosto nel caso delle *deepfakes* e delle immagini artificiali sembra più corretto riflettere sul ruolo del pericolo e, semmai, su un'etica dell'incertezza, nonostante i concetti di rischio e di incertezza storicamente non siano mai stati nettamente separati come teorizza Knight[49]. Rischio misurabile e pericolo non misurabile creano incertezza. Ciò che appare chiaro, invece, è che la società dell'incertezza baumaniana[50] che caratterizza la postmodernità ha reso l'umanità più vulnerabile, aumentando il senso di insicurezza esistenziale e personale[51]: «come in epoca premoderna, la base simbolica delle nostre incertezze è l'ansia creata dal disordine, la perdita di controllo sui nostri corpi, sui rapporti con gli altri, il necessario per vivere, e il grado di autonomia di cui possiamo godere nella vita quotidiana»[52].

Potremmo oggi aggiungere, altresì, la perdita di controllo sui propri dati, sui propri immaginari e della propria identità. Solo nello sviluppo degli eventi e delle circostanze si rivelerà o meno un pericolo arginabile attraverso, secondo Luhmann, sistemi sociali che mirano a ridurre la complessità ontologica della società[53].

Al riguardo Luhmann introduce, accostandolo alla nozione di rischio e di incertezza, anche il concetto di contingenza, con il quale indica la possibilità che una circostanza – diversa dalle proprie aspettative – accada nel corso del tempo e in un determinato ambiente, generando per l'appunto incertezza[54]. La comunicazione diventa, in questo frangente, un mezzo per far sì che la probabilità dell'attuarsi di determinati accadimenti diminuisca. Ma nel caso dell'IA generativa di immagini è la comunicazione (visiva) stessa a generare incertezza, vincolata dalla qualità e dalla tipologia delle immagini che il si-

[48] M. Dean, *Risk, Calcuable and Incalculable*, in «Soziale Welt» 49, 1998, pp. 25-42.
[49] F.H. Knight, *Risk, Uncertainty, and Profit*, cit., p. 205.
[50] Z. Bauman, *La società dell'incertezza*, il Mulino, Bologna 1999.
[51] Z. Bauman, *Modernità liquida*, Laterza, Roma-Bari 2002.
[52] D. Lupton, *Il rischio*, il Mulino, Bologna 2003, p. 9.
[53] N. Luhmann, *Sociologia del rischio*, cit.
[54] N. Luhmann, *Generalized Media and the Problem of Contingency*, in J.J. Loubser *et al.* (eds.), *Exploration in General Theory of Social Science*, Free Press, New York 1976, p. 509.

stema ha prodotto. Simile incertezza informativa è intrecciata alla (in)capacità (legittima) degli esseri umani di comprendere la veridicità dei messaggi visivi.

Come dunque arginare i pericoli emergenti e rispondere all'incertezza o, ancor meglio, al rischio dell'incertezza di fronte all'IA generativa di immagini?

Rischio, pericolo e incertezza rappresentano categorie sociali intrinseche alla società moderna.

Le stime degli esperti sono importanti e necessarie, ma non bastano per ricondurre il rischio e il pericolo sotto il nostro dominio[55]. Il controllo su basi matematiche può generare ulteriori conseguenze "irrazionali". Ecco perché, in questo quadro il rischio si trasforma in incertezza degli eventi, delle conseguenze degli eventi, «degli effetti collaterali e degli effetti collaterali degli effetti collaterali» degli stessi[56], venendo meno la possibilità di compensazione e la capacità di limitazione e controllo del danno poiché le immagini possono viaggiare ovunque senza sapere dove, quando e di fronte a quale sguardo.

Alla luce di quanto pretermesso l'etica sembra delinearsi quale strumento di decodifica e prospettiva di accettazione dell'incertezza, seppur non consenta di immaginarne le conseguenze[57]. Valori come l'equità, la libertà (consapevole) di scelta, la giustizia, la fiducia, il diritto alle informazione, la trasparenza, il rispetto dei diritti fondamentali degli esseri umani oltre alla valutazione (pur sempre relativa) della pericolosità (e dell'entità) del danno di alcune immagini costituiscono le basi per arginare se non almeno limitare minacce e pericoli. Sono standard morali che debbono essere impliciti al design e al data set di riferimento di ciascun sistema di IA generativa di immagini. Noi siamo esposti al rischio solo nel momento in cui decidiamo di utilizzare specifici sistemi algoritmici dei quali conosciamo i meccanismi e l'archivio iconografico di base di cui si avvalgono e, in questo senso, ci assumiamo un rischio in qualche modo calcolabile. Questa prospettiva però non può che risultare inattuabile se non illusoria, dal momento che l'opacità dei sistemi di IA e l'estrema ampiezza dei data set di riferimento rappresentano aspetti caratterizzanti dei sistemi algoritmici generativi.

Se dunque in questo quadro «i rischi sono sempre legati a decisioni, cioè presuppongono una possibilità di scelta»[58] da parte dell'essere umano, i sistemi, invece, compiono scelte autonomamente. Non vale l'assunto secondo cui «[l] e minace incalcolabili vengono trasformate dalla società industriale

[55] A. Giddens, *Il mondo che cambia. Come la globalizzazione ridisegna la nostra vita*, il Mulino, Bologna 2000, p. 40.

[56] U. Beck, *Conditio humana*, cit., p. 34.

[57] *Ivi*, p. 38.

[58] *Ivi*, p. 178.

in rischi calcolabili»[59]. Il rischio si colora di incertezza, contrariamente a quanto suggerito dal Regolamento europeo.

Possiamo dunque rispondere all'incertezza con un'etica sorretta dalla consapevolezza che miri a rafforzare le competenze degli esseri umani per avere le risorse (anche morali) per affrontare le conseguenze degli stati di incertezza e la paura di futuri distopici. Si tratta, da una parte, di una sorta di principio di precauzione che impone a ciascun individuo di avere una cassetta degli attrezzi efficace per fronteggiare e contenere i pericoli (e i danni) dell'IA visiva, dall'altra, di strategie di etica pubblica per sensibilizzare l'opinione pubblica e far agire i governi con regolamentazioni sempre più specifiche e in linea con lo sviluppo tecnologico, fino alla co-responsabilità di tutti i soggetti coinvolti, dagli ideatori, ai governi fino agli utilizzatori dei sistemi. Il vero rischio altrimenti consisterà nel trasformare il mondo in un mondo a dimensione dell'intelligenza artificiale generativa.

English title: Generative Artificial Intelligence, Deepfakes, and Vulnerable Identity: The Ethics of Uncertainty as a Response to an (Un)Controllable Risk

Abstract

The algorithmic turn has given rise to high-performance image generation systems (artificial images, deepfakes, fake images). This scenario is intertwined with the risks, dangers and threats posed by these systems, as well as the uncertainty surrounding their consequences for individuals and society as a whole. Bias, loss of control over one's own data, identity theft, opacity about data use, are just a few emerging ethical issues. Public ethics strategies can play a crucial role in guiding public opinion and citizenship towards awareness and co-responsibility, among all stakeholders from creators to governments (which must implement specific regulations) to users. Such strategies aim to mitigate the uncertainty and fear that these risks cause in individuals, equipping them to become more informed and critical in their engagement with these technologies.

Keywords: deepfake; ethics; generative artificial intelligence; risk; uncertainty.

Veronica Neri
Università di Pisa
veronica.neri@unipi.it

[59] *Ibidem*.