I rivelatori di deepfake rischiano di fare più male che bene

Complice il desiderio di viralità fomentato dagli algoritmi dei *social*, la creazione di **disinformazione** attraverso l'applicazione di sistemi d'intelligenza artificiale è ormai massicciamente attestata. Questi prodotti multimediali fasulli – noti come ***deepfake*** – stanno approfittando del rapido sviluppo tecnologico delle IA per diventare progressivamente sempre più convincenti e persuasivi. La natura ingannatoria di questi artefatti ha spinto le aziende digitali a introdurre molteplici soluzioni che promettono di **identificare l'artificialità** di scatti, video o audio, tuttavia i risultati riscontrati sono tanto scarsi da dimostrarsi addirittura pericolosi.

La demistificazione della disinformazione è caratterizzata da un procedimento tendenzialmente molto lento e oneroso. Ancor peggio, anche quando si riesce a sbugiardare una menzogna, il successo di tale impresa è effimero se i suoi contenuti manipolatori hanno già raggiunto gli animi dei più. L'idea di poter confidare su di una macchina che è in grado di **discernere istantaneamente la verità dalla bugia** è attraente, tuttavia molti di coloro che hanno fatto uso dei *detector* hanno scoperto a loro spese quanto simili scorciatoie siano perlopiù campate per aria.

La guerra israelo-palestinese non ha fatto altro che accendere un nuovo riflettore sul problema, perlomeno da che il commentatore politico Jackson Hinkle ha messo in dubbio su X la natura di un'immagine grottesca recentemente diffusa dal Primo Ministro israeliano, Benjamin Netanyahu. La foto dovrebbe rappresentare il **cadavere carbonizzato di un bambino,** tuttavia il corpo è tanto mal ridotto che la sua natura è difficile da distinguere. Hinkle non ha però messo in dubbio la natura del soggetto, quanto la sua origine. L'uomo ha sottoposto l'istantanea al servizio analitico gratuito del portale ***AI or Not***, il quale l'ha identificata come il frutto di un'intelligenza artificiale, una posizione che però non trova riscontro nell'opinione degli specialisti interpellati da testate quali il *New York Times* e il *Wall Street Journal*.

Il professore di Berkeley **Hany Farid**, noto per la sua ricerca sui generatori di immagini, ha fatto notare su *404 Media* che la tecnologia odierna non sia ancora in grado di sviluppare linee precise e consistenti come quelle illustrati dalla controversa immagine, un'opinione condivisa da molti. Tutto indica che nell'equazione non rientri perlomeno nessuna IA. «Il fatto che AI or Not abbia un **alto tasso di errore** nell'identificare immagini AI compresse, in particolar modo quelle fotorealistiche, riduce considerevolmente la sua utilità per i ricercatori», sosteneva già a settembre il team investigativo di *Bellingcat*.

Portali e servizi di verifica automatizzata adottano d'altronde sistemi di analisi che sono diversi tra di loro e le meccaniche che li alimentano sono spesso nascoste all'interno di insondabili "**scatole nere**". Non è possibile sapere come funzionano – se funzionano –,

inoltre ogni azienda può fornire opinioni profondamente divergenti da quelle avanzate dalla concorrenza. Le corporazioni che detengono il controllo delle intelligenze artificiali di più alto profilo, sollecitate da una notevole pressione politica, stanno timidamente introducendo alcune **filigrane** che dovrebbero virtualmente essere in grado di risolvere questo genere di fraintendimenti, ma simili *escamotage* non sono altro che appendici posticce. Non sono stati pensati come parte fondamentale del servizio offerto al pubblico e hanno dimostrato di essere a loro volta tendenzialmente inutili.

Sebbene la disinformazione sia sempre esistita, la celere diffusione dei *social* e dei *deepfake* ha contribuito a creare un contesto che la nostra società non ha ancora imparato a metabolizzare. L'igiene digitale del soggetto medio è ancora molto acerba, quindi simili contenuti inquinano non poco il discorso pubblico, inoltre le prestazioni fallimentari delle soluzioni tecniche proposte dal Mercato rischiano di creare una **disinformazione di secondo livello** andando ad autenticare o smentire grossolanamente dei contenuti che possono farsi carico di profondi messaggi politici. Tenendo in considerazione la polarizzazione del dibattito pubblico, ora più che mai si dimostra necessario approcciare con sospetto e scetticismo i contenuti che vengono somministrati sul web e sui canali televisi, magari pretendendo che chi si fa carico di diffondere video e foto approfondisca adeguatamente fonte e contesto del materiale che distribuisce. O perlomeno che se ne assuma la responsabilità.

[di Walter Ferri]