UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

Dipartimento di Comunicazione ed Economia

Corso di Laurea Magistrale in

Pubblicità, Comunicazione Digitale e Creatività d'Impresa

A. A. 2024/2025

REALE O ARTIFICIALE? IL RUOLO DELLA CAPTION NEL RICONOSCIMENTO DI IMMAGINI GENERATE DALL'INTELLIGENZA ARTIFICIALE

Relatore:
Prof. Marco Furini

Laureanda:
Giorgia Cancelli

Indice

INTRODUZIONE

Negli ultimi anni, i rapidi sviluppi dell'intelligenza artificiale generativa hanno trasformato profondamente la produzione e la diffusione dei contenuti visivi. Strumenti sempre più avanzati consentono oggi di creare immagini altamente realistiche, spesso indistinguibili da fotografie autentiche, rendendo sempre più complesso per gli utenti valutare l'affidabilità delle informazioni visive. In parallelo, l'ambiente informativo digitale è sempre più mediato da sistemi di personalizzazione algoritmica, che selezionano e organizzano i contenuti in base agli interessi degli utenti. Come osserva Pariser, "questi motori creano un universo di informazioni per ciascuno di noi (...) che modifica radicalmente il modo in cui incontriamo idee e informazioni" (Pariser, 2011, sez. *Introduction*).

Questa trasformazione si inserisce in un ecosistema digitale già profondamente segnato dalle logiche della personalizzazione, dalla rapidità di consumo e dalla competizione per l'attenzione. In questo senso, si può parlare dell'avvio di una vera e propria "era della personalizzazione", iniziata "il 4 dicembre 2009" (*ivi*). Da quel momento, la selezione dei contenuti online è diventata sempre più dipendente da filtri algoritmici capaci di modellare ciò che gli utenti vedono e leggono nei propri ambienti digitali. In questo scenario, i social media rappresentano contesti particolarmente rilevanti, poiché le immagini vengono fruite rapidamente, spesso senza un'analisi approfondita, all'interno di flussi informativi orientati all'engagement e alla permanenza sulla piattaforma.

Alla luce di questi cambiamenti, il presente lavoro si propone di analizzare la capacità degli utenti di distinguere tra immagini reali e immagini generate da intelligenza artificiale, ponendo particolare attenzione al ruolo di alcuni fattori che possono influenzare il processo di valutazione. In particolare, lo studio indaga l'effetto della presenza di una caption testuale associata all'immagine, oltre che il contributo di variabili individuali e percettive, quali la sicurezza percepita nella risposta, la propensione alla condivisione e il possibile emergere di profili differenziati di utenti nel rapporto tra accuratezza, percezione soggettiva e comportamento di diffusione.

A partire da queste premesse, la ricerca si basa su un disegno sperimentale che ha previsto la somministrazione di un questionario online, articolato in due versioni, in cui i partecipanti sono stati chiamati a classificare una serie di immagini come reali o generate da intelligenza artificiale. Inoltre, per ciascuno stimolo sono state raccolte informazioni relative al livello di sicurezza della risposta e, per una parte delle immagini, all'intenzione di condivisione. Il

disegno sperimentale ha consentito di analizzare in modo controllato l'effetto della caption, alternando la sua presenza o assenza tra le due versioni del questionario.

L'obiettivo principale dello studio è quello di comprendere in che misura gli utenti siano effettivamente in grado di riconoscere contenuti visivi generati da IA e quali fattori contribuiscano a rendere tale processo più o meno accurato. In particolare, la ricerca si propone di verificare se la presenza di una caption possa influenzare la capacità di riconoscimento, se esistano differenze tra immagini reali e sintetiche in termini di accuratezza e in che modo variabili soggettive, come la percezione di autenticità e il livello di sicurezza dichiarato, incidano sui comportamenti di valutazione e condivisione.

Il lavoro si articola in cinque capitoli. Il primo capitolo introduce il contesto teorico di riferimento, analizzando l'evoluzione dell'ecosistema dei social media e i principali meccanismi che ne regolano il funzionamento, con particolare attenzione al ruolo degli algoritmi, alle dinamiche di engagement e ai processi di formazione delle abitudini e dell'attenzione degli utenti. Il secondo capitolo esamina la letteratura esistente sul riconoscimento delle immagini sintetiche, sulla media literacy e sui processi cognitivi coinvolti nella valutazione dei contenuti visivi. Il terzo capitolo descrive il disegno della ricerca, gli strumenti utilizzati e le caratteristiche del campione. Il quarto capitolo presenta i risultati dell'analisi dei dati, mentre il quinto discute i principali risultati emersi, evidenziandone le implicazioni, i limiti e le possibili direzioni future di ricerca.

1
SCENARIO DIGITALE E DIPENDENZA

1.1 L'ecosistema dei social media

I social media possono essere definiti come "un insieme di applicazioni basate su Internet che si fondano sui principi ideologici e tecnologici del Web 2.0 e che consentono la creazione e lo scambio di contenuti generati dagli utenti" (Kaplan & Haenlein, 2010, p. 61). Questa definizione mette in evidenza due elementi chiave: da un lato, l'infrastruttura tecnologica che ha reso possibile la diffusione dei social media, rappresentata dal Web 2.0; dall'altro, gli User Generated Content (UGC), cioè la partecipazione attiva degli utenti nella produzione e condivisione dei contenuti online.

Il concetto di Web 2.0, introdotto nel 2004, rappresenta "un nuovo modo di utilizzare il World Wide Web da parte di sviluppatori di software e degli utenti finali" (*ivi*, p. 60), cioè non più come uno spazio statico in cui i contenuti vengono creati e pubblicati da pochi soggetti, ma come una piattaforma aperta e partecipativa, dove tutti possono contribuire attivamente alla costruzione collettiva dell'informazione. In questo nuovo ecosistema digitale, blog, wiki e piattaforme collaborative diventano esempi emblematici di una comunicazione orizzontale e interattiva.

Kaplan e Haenlein individuano alcune tecnologie di base che hanno reso possibile il funzionamento e la diffusione del Web 2.0. Tra queste, Adobe Flash, che ha permesso di integrare animazioni e contenuti multimediali nelle pagine web; RSS (Really Simple Syndication), che consente di ricevere automaticamente aggiornamenti da siti e blog in tempo reale; e AJAX (Asynchronous JavaScript and XML), una tecnologia che rende le pagine web più dinamiche e interattive, permettendo di aggiornare i contenuti senza dover ricaricare l'intera pagina. Questi strumenti hanno rappresentato il punto di svolta tecnico necessario per passare da un web "di lettura" a un web "di partecipazione".

Se il Web 2.0 costituisce dunque la base tecnologica dei social media, l'User Generated Content ne rappresenta il cuore pulsante. Con questa espressione si intende "la somma di tutti i modi in cui le persone utilizzano i social media" (*ivi*, p. 61). Il termine si diffonde a partire dal 2005 e descrive il modo in cui le persone, grazie alle nuove piattaforme digitali, possono contribuire attivamente alla creazione e diffusione delle informazioni.

Affinché un contenuto possa essere definito "generato dall'utente", deve rispondere a tre criteri fondamentali: prima di tutto, deve essere pubblicato su un sito o su una piattaforma accessibile al pubblico (oppure a una comunità online delimitata); in secondo luogo, deve implicare un minimo grado di originalità o sforzo creativo da parte di chi lo produce; infine, dev'essere realizzato al di fuori di contesti e pratiche professionali (*ivi*).

Attraverso questi principi, i social media trasformano gli utenti da semplici lettori passivi a partecipanti attivi, capaci di creare, commentare e diffondere contenuti in rete.

Kaplan e Haenlein propongono una classificazione dei social media basata su due dimensioni principali: da un lato, la presenza sociale e ricchezza dei media, dall'altro, il livello di auto-presentazione e auto-rivelazione.

La presenza sociale fa riferimento al grado di contatto – visivo, acustico o fisico – che un determinato mezzo consente di instaurare tra gli interlocutori: la comunicazione è più intensa quanto più è alta la sensazione di "presenza" dell'altro. La ricchezza dei media, invece, indica la quantità e la varietà di informazioni che un mezzo è in grado di trasmettere in un determinato intervallo di tempo. Le piattaforme che offrono un'elevata presenza sociale e una grande ricchezza mediale (come i mondi virtuali tridimensionali) favoriscono un'interazione più immersiva e realistica, mentre quelle basate principalmente su testi (come blog o wiki) si collocano su un livello inferiore.

La seconda dimensione riguarda l'aspetto sociale: l'auto-presentazione rappresenta il modo in cui gli utenti costruiscono un'immagine di sé agli occhi degli altri, mentre l'auto-rivelazione si riferisce alla quantità di informazioni personali che scelgono di condividere (*ivi*, p. 61-62).

Combinando questi due criteri, gli autori individuano sei tipologie principali di social media: progetti collaborativi, blog, comunità di contenuti, siti di social networking, mondi virtuali di gioco e mondi virtuali sociali. Questa tassonomia permette di comprendere la varietà delle piattaforme oggi esistenti, le loro logiche di funzionamento e le differenti opportunità che offrono agli utenti e alle aziende.

I progetti collaborativi consentono la creazione congiunta di contenuti da parte di più utenti contemporaneamente, rappresentando "la manifestazione più democratica di UGC" (*ivi*, p. 62). Si suddividono in due sottocategorie principali: i wiki, che permettono di aggiungere e modificare testi, e le applicazioni di social bookmarking, che consentono di raccogliere e valutare link o contenuti multimediali. Wikipedia è l'esempio più noto di wiki, mentre Delicious rappresenta un caso emblematico di social bookmarking. L'idea di fondo è che la collaborazione, "lo sforzo congiunto di molti attori porti a un risultato migliore di quello che qualsiasi individuo potrebbe ottenere da solo" (*ivi*). Queste prime forme di partecipazione

collettiva costituiscono la base su cui si è sviluppata l'interattività sociale che caratterizza le piattaforme contemporanee.

I blog rappresentano la prima forma di social media. Si tratta di siti web che raccolgono contenuti organizzati in ordine cronologico inverso, e che possono spaziare da diari personali a raccolte tematiche di informazioni. Oltre ai blog testuali, oggi esistono blog che integrano altri formati mediali, come immagini o video, diffusi attraverso piattaforme come Justin.tv. Molte aziende utilizzano i blog per comunicare con dipendenti, clienti e azionisti, ma questo canale non è privo di rischi: commenti negativi o proteste online possono intaccare la reputazione aziendale, generando "informazioni potenzialmente dannose nello spazio online" (*ivi*, p. 63).

Le comunità di contenuti hanno come obiettivo principale la condivisioni di contenuti multimediali tra utenti, come testi, fotografie, video o presentazioni PowerPoint. Esempi noti sono BookCrossing, Flickr, YouTube e Slideshare. Queste piattaforme non sempre richiedono un profilo personale, e rappresentano un'importante opportunità per le imprese, che le utilizzano per campagne pubblicitarie o iniziative di marketing partecipativo. Ad esempio, nel 2007, Procter & Gamble promosse un concorso per il suo farmaco da banco Pepto-Bismol, esortando gli utenti a caricare su YouTube video brevi in cui cantavano le malattie che il prodotto è in grado di combattere (*ivi*). Al tempo stesso, presentano dei limiti in ambito legale, connessi alla diffusione di materiali protetti da copyright.

I siti di social networking, come Facebook e MySpace, permettono agli utenti di creare un profilo personale, connettersi con amici o colleghi e condividere messaggi, immagini, video e altri contenuti. Essi rappresentano lo spazio più tipico della costruzione dell'identità online, in cui gli individui sviluppano pubblicamente le proprie relazioni sociali. Secondo Kaplan e Haenlein, molte aziende sfruttano questi ambienti per creare comunità di marca o condurre ricerche di mercato basate sull'osservazione etnografica digitale (*netnography*) (*ivi*, p. 64).

I mondi virtuali di gioco (o *virtual game worlds*) sono ambienti tridimensionali interattivi che riproducono dinamiche simili a quelle reali, ma in contesti ludici e regolati. In questi spazi, gli utenti si muovono attraverso avatar e seguono regole precise. "I mondi virtuali rappresentano la manifestazione ultima dei social media, poiché offrono il più alto livello di presenza sociale e ricchezza dei media tra tutte le applicazioni discusse finora" (*ivi*). Anche se non offrono una totale libertà di auto-presentazione, questi mondi rappresentano potenti strumenti di marketing e comunicazione, utilizzati, ad esempio, da aziende come Toyota per promuovere i propri prodotti (*ivi*).

Infine, i mondi virtuali sociali (o *virtual social worlds*) offrono un livello massimo di libertà espressiva e di interazione. A differenza dei mondi di gioco, qui non esistono regole predefinite,

e gli utenti possono vivere una vera e propria "vita parallela" attraverso avatar digitali. Second Life rappresenta il caso più noto: una piattaforma che consente di creare, acquistare e vendere contenuti virtuali in cambio di una valuta interna (Linden Dollars) convertibile in denaro reale. Questi ambienti si configurano anche come nuove opportunità per il marketing, la comunicazione e la formazione aziendale.

La classificazione proposta dagli autori evidenzia come i social media non costituiscano un insieme omogeneo, ma un ecosistema complesso e in continua evoluzione, dove la partecipazione, la collaborazione e la rappresentazione del sé assumono forme diverse a seconda della piattaforma e delle sue finalità. Questa varietà, oltre ad arricchire il mondo digitale, riflette le diverse modalità con cui le persone costruiscono relazioni, condividono informazioni e producono significato nella società contemporanea.

Tuttavia, per comprendere appieno la natura dell'ecosistema digitale contemporaneo, è necessario fare un'ulteriore distinzione tra social media e social network sites (SNS), due termini spesso utilizzati come sinonimi ma che indicano realtà differenti.

Secondo la definizione proposta da Boyd e Ellison, i siti di social network sono "servizi basati sul web che permettono agli individui di:

1. costruire un profilo pubblico o semi-pubblico all'interno di un sistema delimitato;
2. articolare una lista di altri utenti con i quali condividono una connessione;
3. visualizzare e attraversare la propria lista di connessioni e quelle stabilite da altri all'interno del sistema" (Boyd & Ellison, 2007, p. 211).

La differenza tra "social network site" e "social networking site" non è solo terminologica: il secondo termine enfatizza l'aspetto del "fare networking", ovvero instaurare nuove relazioni, spesso tra sconosciuti. Al contrario, la maggior parte dei social network sites non si concentra tanto sulla creazione di nuovi legami, ma più sul mantenimento e la visibilità di reti sociali preesistenti. Ciò che li rende unici, infatti, non è tanto la possibilità di conoscere persone nuove, quanto la capacità di approfondire le connessioni sociali già esistenti (*ivi*).

Ogni sito presenta poi diverse modalità di gestione della visibilità e dell'accesso ai profili. Alcune piattaforme, come i primi Friendster o Tribe.net, consentivano la visibilità pubblica dei profili anche ai non iscritti, mentre altre, come Linkedin o Facebook, controllano ciò che l'utente può vedere in base ad account premium o a reti di appartenenza. La struttura e le regole di accesso rappresentano dunque un elemento distintivo delle varie piattaforme e influenzano le dinamiche di relazione al loro interno (*ivi*, p. 213).

Inoltre, gli SNS condividono una struttura di base composta da profili personali, liste di amici ("Friends"), contatti ("Contacts") e seguaci ("Fans"), insieme a forme di interazione, come

commenti pubblici, messaggi privati o contenuti multimediali condivisi. La costruzione del profilo, spesso arricchito da foto, descrizioni personali e interessi, rappresenta un atto di auto-rappresentazione pubblica, quello che Sundén[1] descrive come il "digitare se stessi per esistere" – "type oneself into being" (*ivi*, p. 211).

Attraverso la visibilità dei contatti e delle interazioni, ogni utente costruisce e mostra una versione pubblica della propria rete sociale, partecipando a una forma di comunicazione relazionale che si basa tanto sulla connessione quanto sulla rappresentazione di sé.

Tuttavia, non tutti i social network sono nati con questa funzione. Alcune piattaforme oggi centrali, come QQ, Cyworld o Skyrock, avevano inizialmente scopi differenti – rispettivamente, la messaggistica istantanea, la discussione comunitaria e il blogging – e solo in seguito hanno integrato funzionalità sociali più complesse (*ivi*, p.213). Altre piattaforme, come MySpace, Linkedin o Facebook, hanno invece costruito fin dall'inizio la propria identità intorno all'idea di rete e di connessione.

In sintesi, mentre i social media rappresentano un ecosistema ampio e diversificato di strumenti e piattaforme per la creazione e la condivisione di contenuti, i social network sites costituiscono una sottocategoria specifica, centrata sulla costruzione, la gestione e la rappresentazione delle relazioni sociali online. Questa distinzione è essenziale per comprendere come le pratiche di partecipazione, visibilità e connessione si siano sviluppate e trasformate nel tempo, aprendo la strada alle piattaforme che oggi dominano la nostra quotidianità digitale.

Per comprendere l'evoluzione dell'attuale ecosistema digitale è utile ripercorrere la storia dei primi siti di social network, osservando come siano nati e come abbiano definito e modificato nel tempo le modalità di relazione online (fig. 1.1).

[1] In Sundén, J. (2003). *Material Virtualities*. New York: Peter Lang.

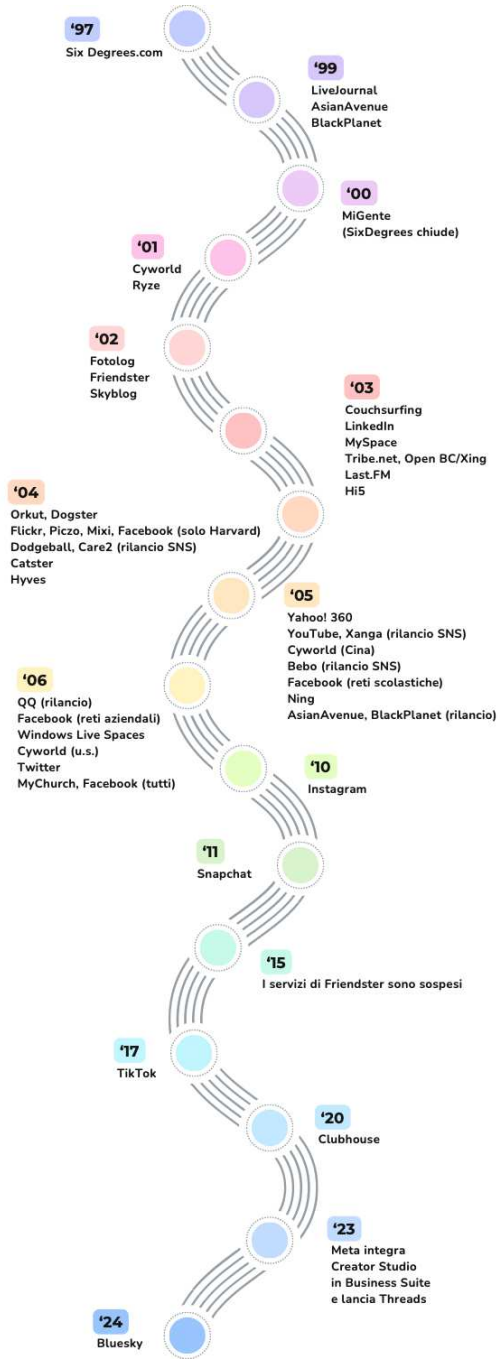


Figura 1.1 – Cronologia delle principali piattaforme di social network e relative date di lancio (elaborazione grafica dell'autrice, a partire da Boyd & Ellison, 2007, con aggiornamenti successivi)

Secondo Boyd e Ellison, il primo sito definibile come social network site fu SixDegrees.com, lanciato nel 1997. Esso permetteva agli utenti di creare un profilo, elencare i propri Amici e navigare tra le connessioni dei propri contatti. Per la prima volta, una piattaforma metteva insieme gli elementi che sono oggi centrali nei social network: profilo, lista di amici e accesso alle connessioni altrui.

SixDegrees raggiunse milioni di utenti, ma non riuscì a sostenersi economicamente e chiuse nel 2000. In quel momento, gran parte delle persone non aveva ancora una rete di amici online

abbastanza estesa da rendere l'esperienza attrattiva e interessante: il servizio offerto era in un certo senso "in anticipo sui tempi", come ha ammesso il suo stesso fondatore (*ivi*, p. 214).

Negli anni successivi, nacquero diversi siti comunitari che iniziarono a introdurre funzioni simili. Tra questi, LiveJournal, AsianAvenue, MiGente e BlackPlanet, che permettevano di creare profili personali e connettersi con altri utenti, seppure in modo meno strutturato. Parallelamente, Cyworld in Corea del Sud e LunarStorm in Svezia anticipavano le logiche che caratterizzeranno le piattaforme successive.

Il vero salto di qualità arrivò infatti con Ryze.com, lanciato nel 2001 per aiutare la gente a sviluppare reti professionali. Inoltre, nacquero progetti come Linkedin, Tribe.net e, soprattutto, Friendster, che nel 2002 introdusse il concetto di connessione tra "amici di amici" (*ivi*, p. 215). Friendster ebbe una rapida crescita iniziale ma anche un'altrettanta rapida crisi: problemi tecnici, una gestione rigida dell'accesso ai profili e il blocco dei cosiddetti "Fakesters" (profili falsi o satirici) provocarono la fuga di molti utenti, delusi e tutt'altro che fidelizzati. Tanti di loro migrarono verso MySpace, nato nel 2003 con un approccio più flessibile e orientato all'espressione personale. Inizialmente, MySpace attirò le band indie-rock che erano state espulse da Friendster e i loro fan, diventando presto un punto di riferimento per i giovani e i creativi. La possibilità di personalizzare i profili e di condividere musica e contenuti visivi trasformò la piattaforma in uno spazio di socialità e auto-rappresentazione. Nel luglio 2005 venne acquistato dalla News Corporation per 580 milioni di dollari e in seguito nacquero preoccupazioni per la sicurezza: si verificarono casi di rapporti sessuali tra adulti e minori (*ivi*, p. 217).

Nel 2004 apparve Facebook, inizialmente riservato agli studenti di Harvard, poi esteso ad altre istituzioni, e infine a chiunque. La sua struttura chiara, basata su reti ad accesso limitato e profili reali, favorì un senso di fiducia e di appartenenza che la differenziava da MySpace (*ivi*, p. 218). Con l'apertura ad un pubblico più ampio e l'introduzione di nuove funzioni (come la "News feed" e le "Applications"), Facebook divenne la piattaforma dominante del decennio successivo, cambiando radicalmente il modo in cui le persone gestivano la propria identità e le relazioni online e segnando il passaggio da un web comunitario a un web centrato sulle relazioni personali e sul tracciamento costante delle interazioni sociali.

Nel frattempo, altri siti si sviluppavano a livello globale: Orkut in Brasile e India, Mixi in Giappone, Hyves nei Paesi Bassi, Hi5 in America Latina, e Bebo nel Regno Unito e in Oceania (*ivi*, p. 217). Questa diffusione mondiale mostrò come la logica del social networking potesse adattarsi a contesti culturali diversi, mantenendo però un nucleo comune basato su profili visibili, connessioni e interazioni pubbliche.

"L'ascesa dei siti di social network indica un cambiamento nell'organizzazione delle comunità online" (*ivi*, p. 219) ed ha segnato un passaggio decisivo nella storia di Internet: da spazi organizzati intorno a comunità di interesse (come i forum) si è passati a spazi "strutturati come reti personali (o 'egocentriche'), con l'individuo al centro della propria comunità" (*ivi*). Come osservano Boyd e Ellison, i social network sites non si strutturano più attorno a un tema o a un argomento, ma intorno alle persone stesse e alle loro connessioni, rispecchiando in modo più realistico le reti sociali della vita offline (*ivi*).

L'evoluzione delle piattaforme di social media ha quindi trasformato non solo le modalità di comunicazione e di relazione tra gli individui, ma anche le forme di interazione quotidiana con la tecnologia. Queste piattaforme permettono di collaborare alla condivisione di contenuti, di costruire un'identità online e di creare reti sociali sempre più interconnesse. Per questo motivo, i social media sono diventati parte integrante delle abitudini digitali contemporanee.

È proprio su questa dimensione comportamentale che si concentra il modello di Nir Eyal, il quale analizza come i prodotti digitali si insinuano nelle routine quotidiane degli utenti fino a far parte dei loro meccanismi psicologici e decisionali.

Perché le abitudini sono così importanti per le aziende? Eyal definisce le abitudini come "comportamenti che si eseguono in totale (o quasi totale) assenza di pensiero cosciente che guidano circa la metà delle nostre azioni quotidiane" (Eyal, 2015, p. 17). Sono dunque quelle azioni che compiamo senza (o quasi) sforzo cognitivo. Un esempio può essere il salutare appena si entra in un negozio, o accendere la macchina del caffè ogni mattina dopo il risveglio. Giorno dopo giorno, l'abitudine si rafforza e l'azione diventa routinaria. Inoltre, attraverso le abitudini, il nostro cervello apprende anche comportamenti più complessi.

Le abitudini guidano le nostre azioni quotidiane: per questo motivo, riuscire a controllarle in maniera efficace può risultare fondamentale per le aziende. In questo modo, esse possono influenzare i potenziali clienti ad utilizzare un prodotto di loro spontanea volontà, senza dover ricorrere a *call-to-action* esplicite. "I prodotti che formano abitudini modificano il comportamento degli utenti e creano il loro coinvolgimento senza sollecitazione" (*ivi*, p. 19), spingendoli a tornare ad utilizzare il prodotto nei momenti in cui si verificano situazioni di routine.

Se la formazione di abitudini ha successo, le aziende possono godere di diversi vantaggi, come: "un maggior valore del cliente nell'arco della vita, una maggiore flessibilità nella determinazione del prezzo, una crescita accelerata e un più forte vantaggio competitivo" (*ivi*, p.32).

Un esempio emblematico è rappresentato da Instagram, un'app che è riuscita a creare un prodotto in grado di formare abitudini negli utenti, tanto da essere entrata nelle loro routine quotidiane (*ivi*, p. 35).
Questa capacità dei prodotti digitali di inserirsi nella quotidianità degli utenti è alla base del successo dei social media, il cui design mira esattamente alla costruzione di routine e comportamenti automatici. Nel paragrafo seguente si analizzeranno più da vicino i meccanismi che rendono possibile la formazione di queste abitudini digitali.

1.2 Algoritmi, engagement e tempo di permanenza

A partire dal 4 dicembre 2009, con un post quasi inosservato sul blog di Google, è iniziata una nuova era nella storia di Internet. Da quel momento, come scrive Eli Pariser, "Google avrebbe utilizzato cinquantasette segnali (...) per formulare ipotesi su chi fossimo e su quali tipi di siti potessero piacerci" (Pariser, 2011, sez. *Introduction*).
L'azienda ha infatti introdotto la ricerca personalizzata per tutti: da quel giorno, i risultati delle ricerche non sarebbero più stati identici per ogni utente, ma adattati in base al profilo, alla cronologia, alla posizione geografica e a una serie di indizi comportamentali raccolti durante la navigazione. Come osserva Pariser, "non esiste più un Google standard" (*ivi*): ognuno di noi riceve una versione del browser calibrata sulle proprie abitudini, preferenze e interessi.
Questa decisione ha segnato l'inizio di quella che l'autore chiama "l'era della personalizzazione".
Il principio alla base è semplice: più un contenuto è rilevante per l'utente, maggiore è la probabilità che quest'ultimo vi presti attenzione, interagisca e compia azioni misurabili, come clic, like, commenti o acquisti. Così, l'obiettivo delle piattaforme diventa massimizzare l'engagement e il tempo di permanenza, trasformando l'esperienza online in un flusso continuo di stimoli su misura.
Dietro a questa apparente "premura" per l'utente si nasconde un modello economico preciso. "La corsa a sapere il più possibile su di te è diventata la battaglia centrale di quest'epoca per giganti di Internet come Google, Facebook, Apple e Microsoft" (*ivi*). Ogni nostra azione online – che si tratti di cercare un volo, guardare un video, mettere un like o lasciare un commento – genera dati che possono essere raccolti, analizzati e persino venduti. Ogni "segnale di clic" diventa una merce, un pezzo di un vasto mercato che alimenta l'intero ecosistema digitale. In pochi millisecondi, i nostri comportamenti possono essere messi all'asta al miglior offerente, e le nostre preferenze trasformate in previsioni algoritmiche (*ivi*).

In questo contesto, l'utente diventa sia prodotto che produttore. Come spiega Pariser, "gli editori stanno perdendo" perché ora "gli utenti, non i siti, sono al centro dell'attenzione" (*ivi*, sez. *The user is the content*). L'informazione non è più mediata solo dai giornali o dai media tradizionali, ma viene continuamente organizzata e selezionata da piattaforme che conoscono i nostri gusti meglio di noi stessi. In passato, erano i redattori professionisti a decidere quali notizie fossero rilevanti; oggi, questa responsabilità è passata agli algoritmi e ai sistemi di raccomandazione: "i redattori professionisti sono costosi, il codice è economico" (*ivi*).

Per capire questo cambiamento, è utile considerare la distinzione tra modello *pull* e modello *push* (*ivi*, sez. *A new middleman*). Nel modello *pull-based*, tipico della prima fase del web, l'utente aveva il controllo: decideva da chi ricevere aggiornamenti e a quali fonti accedere, cercando attivamente i contenuti di suo interesse. Oggi, nel modello *push-based*, è il contrario: sono gli algoritmi a decidere cosa mostrarci, spingendo i contenuti direttamente verso di noi. Pariser osserva che gli appassionati di Internet inizialmente vedevano nel modello pull un progresso, un modo per liberare l'utente dal bombardamento dei mass media. Tuttavia, questo approccio richiedeva attenzione e impegno, mentre il modello push rende tutto automatico: l'utente riceve un flusso costante di notizie, video e post senza dover scegliere attivamente. L'obiettivo è quello di rendere la fruizione talmente continua e scorrevole da non creare punti di arresto, e quindi massimizzare il tempo di permanenza.

Un esempio significativo è il progetto YouTube LeanBack, pensato per fondere i paradigmi pull e push. Si tratta di una sorta di "canale televisivo personale" che seleziona e riproduce contenuti in sequenza in base ai gusti espressi dall'utente, con l'obiettivo di ridurre progressivamente il suo intervento. Come osserva Pariser, LeanBack offre "i vantaggi del push e del pull" insieme, creando un'esperienza che "consente all'utente di fare sempre meno" (*ivi*).

Il rovescio della medaglia è che la piattaforma decide cosa vediamo, e se l'algoritmo "gioca sporco", la realtà che ci viene mostrata può risultare distorta (*ivi*).

La formula alla base di questo sistema è tanto semplice quanto efficace: rilevanza = attenzione = profitto. Pariser riassume tutto in modo molto chiaro: "più le offerte di informazione sono personalmente rilevanti, più pubblicità possono vendere, e più è probabile che tu acquisti i prodotti che ti propongono" (*ivi*, sez. *Introduction*).

Prendiamo Amazon, ad esempio, che riesce a generare miliardi di dollari grazie alla sua abilità di anticipare i gusti di ogni cliente e di presentarli al momento giusto nella sua vetrina virtuale. Netflix, d'altra parte, ottiene fino al 60% dei suoi noleggi grazie ai suggerimenti personalizzati, dimostrando quanto l'analisi algoritmica possa influenzare i comportamenti di consumo (*ivi*). Anche YouTube continua a perfezionare i suoi sistemi di raccomandazione, puntando a

trattenere gli utenti il più a lungo possibile, offrendo contenuti che massimizzano il tempo di permanenza e l'interazione (*ivi*, sez. *A new middleman*).

Tutto ciò crea un effetto di chiusura informativa, che Pariser definisce *filter bubble*, ovvero la "bolla dei filtri": un universo personalizzato di informazioni che modifica il nostro modo di entrare in contatto con idee e notizie. Ogni utente vive in una versione unica di Internet, costruita su misura dalle piattaforme e invisibile agli altri. La filter bubble introduce tre dinamiche fondamentali: prima di tutto, "sei solo al suo interno" – ogni individuo si trova in un ambiente informativo unico e non condiviso; in secondo luogo, "la filter bubble è invisibile" – non abbiamo idea di quali criteri guidino la selezione dei contenuti; infine, "non scegli di entrarvi", perché i filtri personalizzati si attivano automaticamente, senza un consenso esplicito (*ivi*, sez. *Introduction*).

Questa invisibilità rende il fenomeno particolarmente insidioso. L'utente tende a credere che i risultati mostrati da Google, i video suggeriti da YouTube o i post visibili nel feed di Facebook siano oggettivi e neutrali, quando in realtà riflettono – e rafforzano – le sue stesse convinzioni. Pariser mette in evidenza come la personalizzazione possa amplificare il bias di conferma, ovvero la nostra inclinazione a cercare, selezionare e ricordare solo le informazioni che confermano le nostre credenze. Di conseguenza, "un ambiente informativo basato sui clic favorisce i contenuti che confermano le nostre convinzioni, non quelli che le mettono in discussione" (*ivi*, sez. *The Adderall society*).

L'algoritmo, progettato per massimizzare la pertinenza e il coinvolgimento, finisce per creare un circolo vizioso: mostrare più frequentemente ciò che provoca reazioni prevedibili – indignazione, curiosità, desiderio – perché è ciò che riesce a mantenere l'attenzione più a lungo. Questo porta a una forma di "determinismo informativo", in cui "ciò su cui hai cliccato in passato determina ciò che vedrai dopo", fino a rimanere intrappolati "in una versione statica e sempre più ristretta di te stesso – un loop infinito di te" (*ivi*, sez. *Introduction*).

Dal punto di vista cognitivo, Pariser spiega come la filter bubble possa alterare l'equilibrio mentale che regola l'apprendimento e la creatività. I nostri schemi, cioè le strutture cognitive che utilizziamo per interpretare il mondo, tendono naturalmente a stabilizzarsi, ma per crescere devono anche essere messi alla prova da nuove informazioni. La bolla dei filtri rompe questo equilibrio: circondandoci di idee che già condividiamo, riduce la nostra esposizione a prospettive diverse e limita la possibilità di rielaborare e modificare le nostre conoscenze. In altre parole, la personalizzazione "aumenta la proporzione di contenuti che confermano ciò che crediamo" (*ivi*, sez. *The Adderall society*), rendendo più difficile apprendere qualcosa di nuovo o inaspettato.

Le conseguenze si estendono anche alla creatività e all'innovazione. Secondo Pariser, la filter bubble "limita artificialmente l'ampiezza del nostro orizzonte delle soluzioni" (*ivi*): le idee che incontriamo diventano sempre più omogenee e prevedibili, e mancano di quella diversità che genera nuove connessioni. Infatti, le scoperte più significative nascono "dall'introduzione di idee del tutto casuali – proprio quelle che i filtri sono progettati per eliminare" (*ivi*).

In questo senso, Pariser parla di una possibile "società dell'Adderall", dove l'iperfocalizzazione, una concentrazione estrema su ciò che già ci interessa, prende il posto di una conoscenza più ampia e della capacità di sintesi. Un ambiente troppo filtrato, privo di sorprese e discontinuità, può ridurre la curiosità e, di conseguenza, l'apprendimento (*ivi*).

La personalizzazione, quindi, non agisce solo a livello tecnico, ma anche psicologico e sociale. Da un lato, ci offre un'esperienza apparentemente confortevole – "un mondo su misura, (...) popolato dalle nostre persone, cose e idee preferite" – ma dall'altro ci isola in spazi informativi chiusi, dove "non siamo mai annoiati, non siamo mai infastiditi, e i nostri media diventano un perfetto riflesso dei nostri interessi e desideri" (*ivi*, sez. *Introduction*). Tuttavia, questo comfort ha un prezzo: "rendendo tutto più personale, potremmo perdere alcune delle caratteristiche che rendevano Internet così attraente all'inizio" (*ivi*).

Le implicazioni sono profonde. "Ciò che è buono per i consumatori non è necessariamente buono per i cittadini": la logica algoritmica, orientata al coinvolgimento e al profitto, non coincide con quella del bene comune o dell'informazione pubblica. La struttura stessa dei media digitali, "controllando ciò che vediamo e ciò che non vediamo", finisce per modellare anche "il carattere della nostra società" (*ivi*). In questo modo, la filter bubble non solo orienta i consumi e le preferenze individuali, ma modifica la percezione collettiva della realtà.

Il legame tra algoritmi, engagement e tempo di permanenza è chiaro: la personalizzazione non è solo un effetto collaterale, ma il vero motore dell'economia dell'attenzione. Ogni piattaforma compete per catturare e mantenere il tempo degli utenti, utilizzando strategie algoritmiche sempre più avanzate che imparano dai comportamenti passati per prevedere quelli futuri. L'utente che scorre il feed di Facebook o guarda video su YouTube non è semplicemente un consumatore di contenuti, ma una fonte continua di dati che alimentano e migliorano il sistema. In questa dinamica, la nostra attenzione diventa la risorsa fondamentale su cui si basa la redditività delle piattaforme: più restiamo connessi, più il sistema ha successo.

Come scrive Pariser, "gli algoritmi che orchestrano le nostre pubblicità stanno cominciando a orchestrare anche le nostre vite" (*ivi*). Comprendere come funzionano questi meccanismi significa quindi riflettere non solo su come vengono selezionate le informazioni, ma anche su

come queste influenzano i nostri comportamenti, la nostra percezione e la nostra capacità di scelta.

Su questi temi si concentra anche l'analisi di Tristan Harris, che nel suo articolo *How Technology Hijacks People's Minds* (2016) descrive una serie di "dirottamenti" psicologici che le applicazioni usano per manipolare il comportamento degli utenti.

Il primo di questi meccanismi, che Harris chiama "controllo del menù", riguarda la percezione della scelta. Oltre a rispondere ai bisogni degli utenti, le piattaforme definiscono i menù da cui possiamo scegliere. Come in un trucco di magia, l'illusione della libertà nasconde una selezione preimpostata: non ci chiediamo "cosa non è incluso nel menù?", ma semplicemente scegliamo tra le opzioni che ci vengono proposte. La tecnologia restringe lo spazio della scelta consapevole.

Un secondo meccanismo è quello della ricompensa variabile, preso in prestito dal funzionamento delle slot machine. Ogni volta che controlliamo la posta o scorriamo un feed, speriamo inconsciamente di "vincere", cioè di trovare una nuova notifica, un like o un messaggio. Ciò che genera dipendenza è proprio l'imprevedibilità della ricompensa: secondo Harris, ora abbiamo tutti una slot machine in tasca (Harris, 2016, sez. *Hijack #2: Put a Slot Machine In a Billion Pockets*). Questo principio, che si approfondità nei paragrafi successivi, è stato ampiamente studiato dalla psicologia comportamentale, e sta alla base di molte piattaforme digitali. Inoltre, spiega il motivo per cui controlliamo il nostro smartphone così spesso: in media 150 volte al giorno (*ivi*). La promessa di una potenziale gratificazione, che non è mai garantita ma sempre presente, alimenta il ciclo dell'engagement.

Un altro aspetto di manipolazione è la paura di perdere qualcosa di importante (FOMSI, *Fear of Missing Something Important)*. I social media alimentano questa ansia presentandosi come fonti indispensabili di aggiornamenti e connessioni. La preoccupazione di "rimanere indietro" o di non vedere una notizia, un invito o un messaggio significativo ci spinge a rimanere connessi. Come osserva Harris, la paura è più intensa prima di disconnettersi che dopo: quando finalmente ci stacchiamo, non succede nulla di catastrofico. Tuttavia, questa consapevolezza si manifesta solo quando siamo già fuori dal sistema, perché "non ci manca quello che non vediamo" (*ivi*, sez. *Hiajck #3: Fear of Missing Something Important (FOMSI)*), mentre all'interno la pressione a rimanere iperconnessi è costantemente riattivata.

A queste dinamiche si aggiungono altri meccanismi sociali profondamente umani, come il bisogno di approvazione e la reciprocità. Le piattaforme sfruttano il nostro desiderio di riconoscimento e la pressione sociale a ricambiare gesti o messaggi. Come sottolinea l'autore, un "mi piace" o una menzione non sono atti spontanei, ma eventi spesso suggeriti

dall'algoritmo, che sa quando siamo più vulnerabili al giudizio degli altri (per esempio, dopo aver cambiato la foto del profilo). Vedremo come anche Linkedin e altri sistemi di messaggistica giocano sulla reciprocità per mantenere vivo il ciclo dell'interazione.

Un'altra strategia davvero efficace è quella delle "scodelle senza fondo", che consiste nell'eliminare qualsiasi punto di arresto. Harris fa un paragone con un esperimento di Brian Wansink: quando una ciotola di zuppa si riempie da sola, le persone continuano a mangiare senza neanche accorgersi di quanto hanno consumato (*ivi*, sez. *Hijack #6: Bottomless bowls, Infinite Feeds, and Autoplay*). Allo stesso modo, i feed infiniti di Facebook, Instagram o TikTok, insieme alla riproduzione automatica (*autoplay*) dei video su YouTube o Netflix, ci spingono a "consumare" informazioni senza limiti, allungando indefinitamente il tempo che passiamo online.

In tutti questi casi, la logica di fondo resta la stessa: massimizzare il tempo di permanenza. Come sottolinea Harris, "l'interruzione fa bene al business" (*ivi*, sez. *Hijack #7: Instant Interruption vs. 'Respectful' Delivery*). Le notifiche, la visibilità del "messaggio letto" e l'autoplay sono strumenti che incoraggiano risposte rapide e abbassano la soglia dell'attenzione. Il risultato è una "tragedia dei beni comuni": miliardi di interruzioni ogni giorno, attenzione frammentata, produttività ridotta (*ivi*).

In risposta a questo, Harris propone un cambio di paradigma: passare da una tecnologia che compete per il nostro tempo a una che lo protegge, perché "il tempo delle persone è prezioso", e che tuteli i nostri valori, senza amplificare le nostre dipendenze (*ivi*, sez. *Summary And How We Can Fix This*). L'autore ci invita a progettare dispositivi e interfacce che promuovano la consapevolezza e la scelta, trattando il tempo e l'attenzione come beni da preservare, non da sfruttare.

Per concludere, Pariser analizza il ruolo degli algoritmi nella creazione di universi informativi personalizzati, mentre Harris mette in luce come questi algoritmi si traducano in esperienze di design pensate per massimizzare l'engagement. Entrambi, sebbene da prospettive diverse, descrivono la stessa logica economica, ovvero quella di un'industria che monetizza l'attenzione e costruisce sistemi che orientano, frammentano e invadono il nostro tempo.

Come già accennato, la formazione di abitudini digitali è una componente essenziale per le aziende che vogliono emergere in un mercato sempre più competitivo e in continua evoluzione. Esistono tuttavia dei requisiti senza i quali le abitudini non possono formarsi? Per scoprirlo, basterà riportare in un piano cartesiano due fattori: la frequenza – sull'asse delle ordinate – che indica quanto spesso si verifica un comportamento, e l'utilità percepita – sull'asse delle ascisse – che rappresenta "quanto è utile e gratificante quel comportamento, nella mente dell'utente,

rispetto ad altre soluzioni alternative" (Eyal, 2015, p. 27). Nel grafico, l'area celeste sopra la curva segna la zona dell'abitudine, dove la frequenza e l'utilità percepita sono sufficienti a trasformare un comportamento desiderato in un'abitudine (fig. 1.2). Dunque, un comportamento può diventare abitudine solo se si manifesta con una certa frequenza e genera un sufficiente vantaggio percepito.

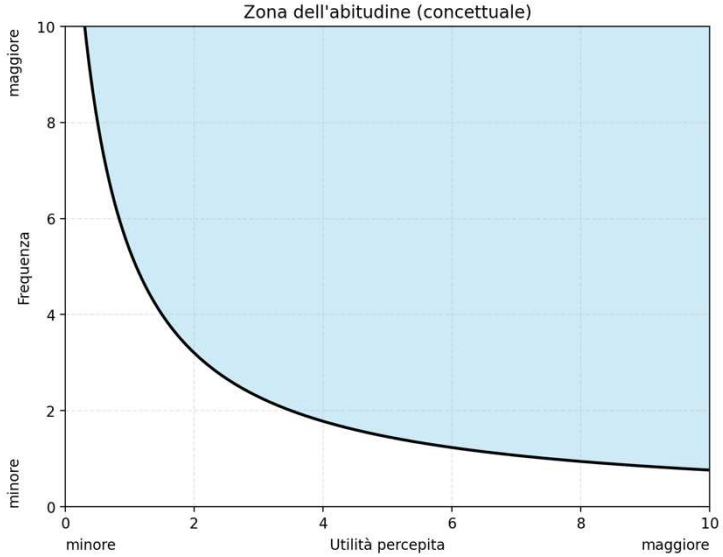


Figura 1.2 – Rappresentazione concettuale della zona dell'abitudine (rielaborazione dell'autrice da Eyal, 2015). Grafico realizzato in Python.

Eyal evidenzia che fino ad oggi non è stata definita una "scala temporale universale che valga per tutti i comportamenti che si trasformano in abitudini" (*ivi*, p. 28). Alcune abitudini possono formarsi in poche settimane, mentre altre potrebbero richiedere mesi.
Nelle prossime pagine, si analizzerà come le aziende riescano a creare abitudini attraverso un processo specifico: il modello del gancio.

1.3 Meccanismi di dipendenza e attenzione

"Il modello del gancio spiega la logica che sta alla base della progettazione di molti prodotti e servizi di successo che formano abitudini e che usiamo quotidianamente" (Eyal, 2015, p. 15). È uno strumento pratico, un processo che si articola in quattro fasi: trigger, azione, ricompensa variabile, investimento (fig. 1.3). Di seguito, un'attenta descrizione dei quattro elementi, con relativi esempi.

Figura 1.3 – Modello del gancio di Nir Eyal (Dell'Olio, 2020)

La prima fase è rappresentata dal trigger: un elemento che aziona il comportamento, un innesco. Per spiegarne il concetto, Eyal paragona le abitudini alle perle. Nelle ostriche, ciò che innesca la formazione di una perla è l'arrivo di un fattore esterno, "di irritazione", inaspettato e indesiderato. Questo fa sì che l'ostrica, come meccanismo di difesa, inizi a rivestire il parassita di strati luccicanti. Allo stesso modo, "le nuove abitudini hanno bisogno di un fondamento su cui svilupparsi": i trigger, che ci spingono a compiere azioni (*ivi*, p. 36).

I trigger possono essere di due tipi: esterni e interni. I trigger esterni comunicano all'utente l'azione successiva, il prossimo passo da fare, usando stimoli sensoriali. Un esempio online può essere l'invito ad effettuare il log-in tramite un messaggio di posta elettronica, che presenta un pulsante colorato in evidenza. Questo elemento richiama l'attenzione e la curiosità dell'utente, spingendolo a compiere il passo successivo: cliccare sul pulsante. Al contrario, i trigger interni non si possono vedere o percepire, ma si legano a emozioni e comportamenti già esistenti nella mente dell'utente. "Collegare trigger interni con un prodotto è il traguardo ambito dalla tecnologia di consumo" (*ivi*, p. 41). Come sottolinea Eyal, le nostre emozioni, specialmente quelle negative, sono trigger interni potentissimi, capaci di influenzare profondamente i nostri comportamenti quotidiani. Sentimenti come la solitudine, la demotivazione, l'incertezza o la frustrazione scatenano il desiderio di compiere un'azione che allievi il disagio provocato da tali emozioni negative. "Gli utenti che trovano un prodotto in grado di alleviare il loro disagio formeranno, con il tempo, associazioni forti e positive con il prodotto stesso" (*ivi*), capaci di radicarsi nella loro memoria. In questo modo, gli utenti si "agganciano" al prodotto, instaurando un legame quasi imprescindibile che, nel tempo, si trasforma in abitudine.

Instagram ha fatto uso di diversi trigger esterni per attirare l'attenzione dell'utente, come le foto condivise su Facebook e Twitter che invitano nuovi utenti a installare l'app, o le stesse notifiche push relative alle pubblicazioni degli amici, che suggeriscono un'azione successiva. Inoltre, Instagram contribuisce ad alleviare la cosiddetta FOMO (*fear of missing out*), ovvero la paura di essere tagliati fuori o di perdersi qualcosa.

Il trigger suggerisce all'utente che cosa deve fare. Così, si passa alla fase successiva, quella dell'azione: "il comportamento esibito in previsione di una ricompensa" (*ivi*, p. 12).

Secondo il modello del comportamento di Fogg, "sono necessari tre ingredienti per l'avvio di qualsiasi comportamento: (1) l'utente deve avere una motivazione sufficiente; (2) l'utente deve avere l'abilità o la capacità necessaria per completare l'azione desiderata; e (3) deve essere presente un trigger per attivare il comportamento" (*ivi*, p. 51). Tutto ciò si può riassumere nella formula C = MAT, secondo cui perché un comportamento (C) si manifesti devono essere presenti, in misura sufficiente, motivazione (M), abilità (A) e un trigger (T) (*ivi*).

La motivazione rappresenta il grado del desiderio di compiere quella data azione. Sempre secondo Fogg, sono tre i motivatori fondamentali che alimentano il nostro desiderio di agire: cercare il piacere ed evitare il dolore, cercare la speranza ed evitare la paura, cercare l'accettazione sociale ed evitare il rifiuto (*ivi*, p. 52). Ad esempio, le pubblicità che presentano corpi femminili in perfetta forma promettono il piacere per catturare l'attenzione e spingere l'utente all'azione. Allo stesso modo, le pubblicità che fanno leva sul motivatore della coesione sociale mostrano spesso gruppi di amici sorridenti e comunità affiatate in momenti di felicità condivisa.

Tuttavia, la motivazione non basta. Ciò che ci spinge a compiere un'azione è anche l'abilità: "la capacità di realizzare un determinato comportamento" (*ivi*, p. 58). Si tratta di semplificare la vita dell'utente, rendendogli più facile l'accesso e l'utilizzo di un dato prodotto o servizio: più si riducono i passaggi necessari a completare un'attività, più cresce la probabilità che l'utente la ripeta. Così, il tasso di adozione aumenta. Un esempio che porta Eyal è quello di Blogger, azienda che alla fine degli anni Novanta rivoluzionò il mondo dei blog, facilitando l'accesso alla piattaforma da parte degli utenti e permettendo loro di pubblicare contenuti online in modo estremamente semplice. "Quanto più facile è l'azione, tanto più è probabile che l'utente la compia e continui a percorrere il ciclo, verso la fase successiva del modello del gancio" (*ivi*, p. 59). Infatti, questa fase del modello include i sei elementi della semplicità individuati da Fogg:

- Tempo: quanto tempo serve per concludere un'azione.
- Denaro: il costo di iniziare un'azione.

- Sforzo fisico: lo sforzo fisico necessario per compiere un'azione.
- Cicli cerebrali: il grado di impegno e concentrazione mentali necessario per intraprendere un'azione.
- Devianza sociale: il livello di accettazione sociale di quel comportamento.
- Non-routine: il grado di coerenza di quell'azione con le routine dell'utente, o al contrario la sua capacità di interromperle.

Per comprendere meglio la forza della semplicità, è sufficiente osservare come alcuni social network abbiano potenziato le proprie funzionalità per facilitarne l'utilizzo. Uno di questi è Facebook. Molte aziende permettono agli utenti di registrarsi utilizzando le proprie credenziali di Facebook, riducendo così i passaggi necessari per accedere alla piattaforma.

Un altro esempio di prodotto semplificato è Apple, che ha permesso ai possessori di un iPhone di lanciare l'app della fotocamera direttamente dal blocco schermo, per creare più facilmente e più velocemente foto e video (*ivi*, p. 62). Oggi, Apple continua a evolvere e a migliorare le proprie prestazioni: con i nuovi modelli di iPhone è possibile aprire l'app fotocamera da un apposito tasto laterale (fig. 1.3.1). Ancora una volta, l'azienda è riuscita a semplificare l'azione dell'utente. Tutto ciò dimostra come la semplicità rafforzi i comportamenti desiderati negli utenti (*ivi*, p. 64).

Figura 1.3.1 – Tasto "Controllo fotocamera" di iPhone 16 (Apple, s.d.)

Un'altra modalità utile ad aumentare la probabilità che una data azione si compia è l'utilizzo delle euristiche: scorciatoie mentali che ci aiutano a prendere decisioni rapide e a formarci opinioni. Ad esempio, secondo l'euristica della scarsità, "il valore percepito di un prodotto può diminuire, se quel prodotto inizialmente è scarso e poi diventa abbondante" (*ivi*, p. 68).

Il prossimo passo nel modello del gancio è quello della ricompensa variabile, in cui gli utenti vengono gratificati attraverso la risoluzione di un problema e, al tempo stesso, motivati a ripetere l'azione iniziata nella fase precedente.

Vari esperimenti e ricerche hanno dimostrato che "ciò che ci spinge ad agire non è la sensazione che riceviamo della ricompensa stessa, ma il bisogno di alleviare la brama di quella ricompensa": ciò che Eyal definisce "stress del desiderio" (*ivi*, p. 76).

È fondamentale comprendere l'importanza della variabilità nella progettazione e nello sviluppo di un prodotto o di un servizio. Come abbiamo visto, le abitudini ci permettono di svolgere un'attività in modo del tutto, o quasi del tutto, inconscio. Tuttavia, quando viene meno la relazione di causa ed effetto che il nostro cervello si aspetta, la nostra attenzione si riaccende improvvisamente, insieme alla curiosità. La novità ha infatti la capacità di risvegliare la nostra mente dallo stato routinario e di alimentare nuovamente interesse e coinvolgimento.

Le ricompense variabili sono rintracciabili in ogni tipologia di prodotto o servizio progettato per catturare la nostra attenzione, e si classificano in tre categorie principali: ricompense della tribù, della caccia e del sé.

"Il nostro cervello è adattato alla ricerca di ricompense che ci fanno sentire accettati, attraenti, importanti e inclusi" (ivi, p. 79). Le ricompense della tribù, o sociali, nascono dal bisogno di accettazione e rinforzo da parte degli altri. Ogni volta che pubblichiamo una foto o un tweet, ci aspettiamo dalla nostra comunità una forma di convalida sociale, che ci spinge a tornare e ritornare per ottenerne ancora.

Su Facebook, ad esempio, "il clic sul pulsante 'mi piace' offre una ricompensa variabile ai creatori dei contenuti", "una convalida tribale a quanti hanno condiviso i contenuti" e "una ricompensa variabile che li motiva a continuare a pubblicare" (*ivi*, p. 80).

Il secondo tipo di ricompensa variabile è la ricompensa della caccia, definita dalla ricerca di risorse. "Il bisogno di acquisire oggetti fisici, come cibo e altre cose che contribuiscono alla nostra sopravvivenza, fa parte del 'sistema operativo' del nostro cervello" (*ivi*, p. 84).

Eyal sottolinea che questo meccanismo non è cambiato nel tempo: i nostri antenati cacciavano per procurarsi il cibo, mentre oggi "cacciamo" informazioni, denaro o gratificazioni simboliche. Un classico esempio di ricompensa variabile della caccia è rappresentato dalle slot machine: i giocatori sono agganciati alla remota possibilità di vincere il jackpot, una vincita concessa a intervalli del tutto casuali. È proprio questa variabilità a mantenere il giocatore legato alla macchina, in un inseguimento tanto eccitante quanto imprevedibile.

Lo stesso principio vale per i social media. Twitter, ad esempio, espone l'utente a un flusso potenzialmente infinito di contenuti non ordinati per rilevanza. Questo cocktail di contenuti

eterogenei, alcuni irrilevanti e altri estremamente interessanti, spinge l'utente a scorrere ancora e ancora, finché non trova ciò che soddisfa la propria curiosità.

Infine, "le ricompense del sé sono alimentate dalla 'motivazione intrinseca'" (*ivi*, p. 87), cioè da una forma di soddisfazione e realizzazione personale.

Secondo la teoria dell'autodeterminazione, le persone ricercano la soddisfazione del desiderio di competenza. È ciò che accade, ad esempio, nei videogiochi: i giocatori provano piacere e appagamento al raggiungimento di obiettivi e premi.

Un esempio più quotidiano è rappresentato dal servizio di posta elettronica: l'utente mira a leggere tutti i messaggi non aperti, per provare una sensazione di ordine e completezza. L'app Mailbox, di proprietà di Dropbox, offre proprio "un senso di completamento e di padronanza" (*ivi*, p. 89), organizzando le e-mail in cartelle ordinate per aiutare l'utente a raggiungere quella sensazione di soddisfazione data dalla casella di posta vuota.

Eyal riporta alcune considerazioni cruciali per una progettazione efficace delle ricompense variabili. In primo luogo, è fondamentale comprendere a fondo le esigenze e le motivazioni degli utenti, così da offrire ricompense coerenti con i loro trigger interni.

In secondo luogo, è essenziale lasciare all'utente un certo grado di autonomia: "le tecnologie di maggior successo rivolte ai consumatori (...) sono quelle che nessuno ci costringe a usare" (*ivi*, p. 95). Gli utenti devono percepire di avere il controllo sul prodotto o sul servizio, non di subirlo.

Infine, Eyal distingue fra variabilità finita e variabilità infinita. La prima riguarda prodotti o servizi che, dopo un certo periodo di utilizzo, diventano prevedibili o scontati. Questi prodotti sono costretti a reinventarsi costantemente per mantenere viva la curiosità dell'utente.

La seconda, invece, appartiene a esperienze che conservano nel tempo il loro potenziale di attrazione, mantenendo l'interesse dell'utente sempre attivo "grazie a una variabilità che si conserva anche con l'uso ripetuto" (*ivi*, p. 98).

Le piattaforme social che conosciamo, come Instagram o TikTok, sfruttano proprio i contenuti generati dagli utenti (*user generated content*) per offrire la percezione di un flusso infinito di stimoli e novità.

L'ultima fase del modello del gancio prevede che gli utenti effettuino un investimento nel prodotto. "Gli utenti sono sollecitati a immettere nel sistema qualcosa che abbia valore, che aumenterà la probabilità che usino il prodotto e percorrano nuovamente il ciclo del gancio" (*ivi*, p. 108). In questo stadio, dopo aver ricevuto delle ricompense variabili, gli utenti sono spinti a compiere uno sforzo ulteriore, spesso guidati dal principio della reciprocità: tendiamo a restituire qualcosa quando percepiamo di aver ricevuto un beneficio (*do ut des*).

L'idea alla base di questa fase "è quella di fare leva sul fatto che gli utenti comprendano che il servizio migliorerà con l'uso (e con l'investimento personale)" (*ivi*, p. 109).

Eyal sottolinea come, nel tempo, questo processo produca un cambiamento dell'atteggiamento negli utenti, che avviene attraverso tre tendenze principali in grado di influenzare le nostre azioni future.

In primo luogo, "quanto più impegno profondiamo in qualcosa, tanto più è probabile che vi attribuiamo valore" (*ivi*, p. 107). Un esempio emblematico è quello dell'azienda svedese IKEA, che consente ai clienti di costruire autonomamente i propri mobili. In questo modo, gli utenti investono tempo ed energia fisica nel prodotto, e tale sforzo contribuisce ad aumentare il valore percepito del bene acquistato. Facendo leva sull'impegno personale dei clienti, IKEA riesce così a trasferire valore ai propri prodotti (*ivi*, p. 105).

In secondo luogo, "è più probabile che siamo coerenti con i nostri comportamenti passati" (*ivi*, p. 107). Quando all'utente viene richiesto inizialmente un piccolo investimento, cresce la probabilità che in futuro egli compia azioni di maggiore portata.

"Quanto più gli utenti investono su un prodotto con piccoli frammenti di lavoro, tanto più prezioso diventa il prodotto per la loro vita, e tanto meno ne mettono in forse l'utilizzo" (*ivi*, p. 120). Questo principio spiega perché le aziende che riescono a introdurre nei propri servizi una logica di impegno progressivo fidelizzano maggiormente i propri utenti.

Infine, "modifichiamo le nostre preferenze per evitare la dissonanza cognitiva" (*ivi*, p. 107).

Quando ci troviamo di fronte a nuovi stimoli, come cibi o bevande dal gusto inusuale, la prima reazione può essere il rifiuto. Tuttavia, con la ripetizione dell'esperienza, tendiamo a adattare le nostre preferenze per ridurre l'incoerenza interna tra ciò che pensiamo e ciò che osserviamo nel comportamento degli altri. In questo modo, la nostra percezione si modifica e l'oggetto prima "estraneo" diventa gradualmente familiare e accettato.

Oltre a modificare l'atteggiamento, affinché l'utente riutilizzi il prodotto nel tempo è necessario immagazzinare valore in forme diverse.

Questo valore può assumere differenti configurazioni:

- Attraverso i contenuti: creati o meno direttamente dall'utente, i contenuti acquisiscono valore nel tempo, rendendo sempre più difficile abbandonare il servizio. È il caso, ad esempio, delle fotografie e dei "mi piace" pubblicati su Facebook, che rappresentano un vero e proprio archivio personale.
- Attraverso i dati: Linkedin ha dimostrato che quanto più gli utenti condividono informazioni personali e professionali, tanto più diventano legati alla piattaforma.

- Attraverso i follower: più l'utente cura la selezione degli account da seguire, più il suo feed risulta coerente con i propri interessi, aumentando il valore percepito del servizio.
- Attraverso la reputazione: utenti, acquirenti o venditori, rimangono fedeli a un servizio nel quale hanno investito tempo e impegno per costruire la propria reputazione. Su eBay, ad esempio, chi gode di punteggi di soddisfazione più elevati ha maggiori probabilità di vendere, e dunque un incentivo a restare attivo.
- Attraverso le competenze: quando un prodotto richiede tempo per essere appreso, come Adobe Photoshop, gli utenti tendono a sviluppare un legame più profondo. Più aumenta la loro competenza, più risulta difficile abbandonare la piattaforma a favore di un concorrente.

Perché l'abitudine si consolidi, gli utenti devono attraversare molti cicli del modello del gancio, continuando a utilizzare il prodotto. Affinché questo accada, è necessaria l'attivazione di trigger esterni: "gli utenti impostano trigger futuri durante la fase di investimento, offrendo alle aziende la possibilità di coinvolgerli nuovamente" (*ivi*, p. 116).

Un esempio efficace è Pinterest, che usa trigger successivi per motivare la riattivazione dell'utente. La piattaforma invia, ad esempio, notifiche nel momento in cui altri utenti interagiscono con un contenuto salvato o contribuiscono a una bacheca condivisa, fornendo così un motivo concreto per tornare sul sito e proseguire l'esperienza.

Inoltre, Pinterest rappresenta un caso paradigmatico perché incarna tutte le fasi del modello del gancio. Innanzitutto, il trigger è costituito dalle notifiche e dalle email personalizzate che richiamano l'utente a interagire; in seguito, l'azione è il salvataggio o la condivisione di contenuti visivi che generano gratificazione immediata; poi, la ricompensa variabile risiede nella scoperta continua di nuovi pin, resa possibile dal feed algoritmico; infine, l'investimento è rappresentato dal tempo dedicato alla creazione di bacheche, alla raccolta di immagini e alla personalizzazione del proprio profilo, che aumenta progressivamente il valore percepito del servizio.

"Attraverso cicli successivi del modello del gancio, si rafforza l'attrazione degli utenti per l'esperienza: finiscono per fare sempre più affidamento sul prodotto come soluzione per i loro problemi, fino a che si forma la nuova abitudine, e la nuova routine" (*ivi*, p. 120).

Il modello del gancio di Nir Eyal mostra come le piattaforme digitali riescano a creare abitudini profonde negli utenti.

Attraverso le quattro fasi – trigger, azione, ricompensa variabile e investimento – il comportamento viene guidato in modo graduale, fino a diventare automatico. Ogni fase alimenta la successiva: un piccolo stimolo riattiva l'attenzione, genera un'azione, porta a una

gratificazione e infine a un investimento personale, che rafforza ancora di più il legame con il prodotto.

Questo ciclo continuo spiega perché i social media riescano a catturare il nostro tempo e la nostra attenzione, spingendoci a tornare più volte al giorno, spesso senza rendercene conto. Comprendere queste dinamiche aiuta a riconoscere quanto le piattaforme facciano leva su meccanismi psicologici semplici ma potenti, come il bisogno di appartenenza, la curiosità o la ricerca di ricompensa, per mantenerci coinvolti.

Il capitolo successivo esplora il lato critico di questi processi, analizzando come i social media si inseriscono nel quadro del capitalismo della sorveglianza e delle nuove forme di controllo dell'attenzione.

2
STATO DELL'ARTE E QUADRO TEORICO

2.1 Perché è difficile distinguere il reale dal sintetico

Negli ultimi anni la qualità delle immagini generate dall'intelligenza artificiale è cresciuta molto rapidamente. I modelli attuali sono in grado di produrre fotografie estremamente pulite, coerenti e ricche di dettagli, rendendo sempre più difficile distinguerle da quelle reali.
È importante specificare che queste difficoltà non riguarda soltanto gli utenti che hanno poca familiarità con il digitale: commettono errori frequenti anche persone giovani, motivate e abituate all'uso delle tecnologie. La letteratura più recente mostra che le motivazioni sono molteplici e riguardano aspetti sia percettivi che cognitivi.
Uno studio di Nightingale e Farid (2022) sul riconoscimento dei volti sintetici dimostra che i volti IA moderni sono indistinguibili da quelli reali. Anche quando viene fornito un breve training, le persone ottengono risultati vicini al caso. Inoltre, lo studio mostra che molte immagini sintetiche vengono percepite come più "affidabili" di quelli reali (fig. 2.1.1).

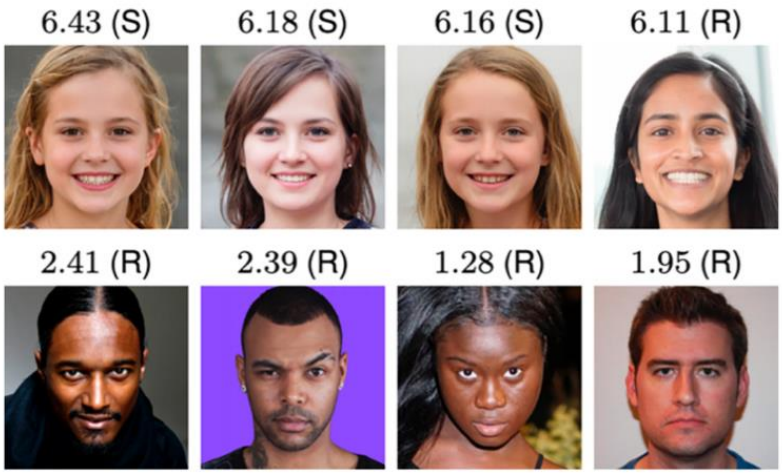


Figura 2.1.1 – I quattro volti percepiti come più affidabili (in alto) e i quattro percepiti come meno affidabili (in basso), con punteggio di fiducia su scala 1-7. In media, i volti sintetici risultano più affidabili di quelli reali (Nightingale & Farid, 2022).

"Ciò potrebbe essere dovuto al fatto che i volti sintetizzati tendono ad assomigliare maggiormente ai volti medi, che sono considerati più affidabili" (*ivi*, p. 2). Il risultato può essere considerato un traguardo per chi lavora nell'ambito, ma dovrebbe far sorgere dubbi e preoccupazioni su possibili problemi futuri legati alla democratizzazione di immagini fake, che

possono portare, ad esempio, alla creazione di profili falsi molto convincenti, o alla distribuzione non consensuale di foto o video intimi.

Questa difficoltà non riguarda solo i volti. Lu et al. (2023) hanno analizzato il riconoscimento di immagini appartenenti a otto diverse categorie, come paesaggi, animali, uomini, donne o scene con più persone. Anche in questo caso l'accuratezza è bassa e quasi quattro immagini su dieci vengono classificate in modo errato (*missclassification rate*: 38,7%). I partecipanti riescono a distinguere le immagini reali da quelle false della categoria "Multiperson" con un tasso di accuratezza del 67,5%, mentre le immagini della categoria "Object" vengono distinte correttamente con un tasso di accuratezza più basso, uguale al 50,8% (fig. 2.1.2). Gli autori sottolineano che il risultato porta a ipotizzare che gli esseri umani potrebbero percepire ogni categoria di foto in modo differente, suggerendo che "gli attuali modelli generativi basati sull'intelligenza artificiale potrebbero essere efficaci nel generare alcune categorie, ma non altrettanto efficaci nel generarne altre" (*ivi*, p. 7).

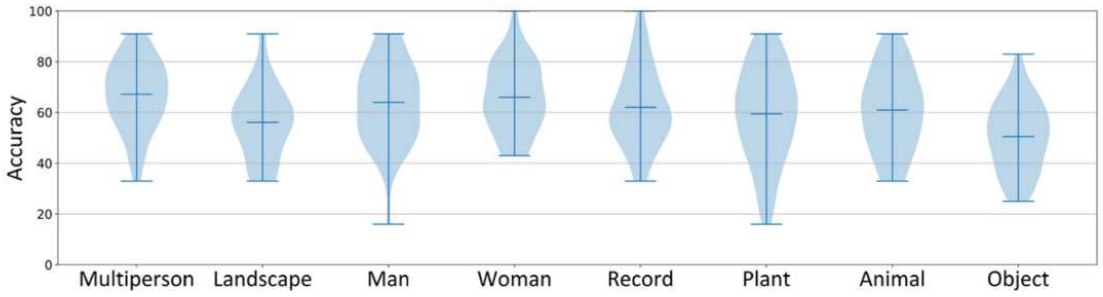


Figura 2.1.2 – Distribuzione dell'accuratezza umana nella distinzione tra immagini reali e sintetiche in otto categorie di contenuto (Lu et al., 2023).

Lo studio mostra un altro elemento interessante: le persone riconoscono meglio le immagini reali (il 66,9% delle immagini reali viene identificato correttamente), mentre tendono a scambiare per autentiche quelle generate dall'IA (il 44,2% delle immagini IA viene identificato correttamente). Ciò suggerisce che molti artefatti sono ormai talmente sottili da sfuggire all'occhio umano.

Questi risultati vengono confermati anche su larga scala: in uno studio molto recente, Roca et al. (2025) hanno raccolto oltre dodicimila partecipanti attraverso un gioco online ("Real or Not Quiz") in cui agli utenti veniva chiesto di indicare se immagini fotografiche realistiche fossero "Real" o "AI-generated". Le immagini, sia reali che sintetiche, rappresentavano situazioni quotidiane e contesti non professionali, simili a quelli che circolano comunemente sui social media, e sono state suddivise dagli autori in diverse categorie di contenuto visivo, tra cui *people*,

vehicles, *objects*, *urban* e *nature* (fig. 2.1.3). Le foto sintetiche sono state create dai modelli generativi più diffusi, come MidJourney, DALL-E 3 e Stable Diffusion. L'obiettivo degli autori è stato duplice: sensibilizzare sulla difficoltà di distinguere immagini sintetiche e raccogliere dati su larga scala. Anche qui l'accuratezza generale resta poco superiore al caso del "lancio della moneta", con un tasso di successo del 62% (immagini reali e IA), confermando la bassa capacità umana nel riconoscere contenuti sintetici (Roca et al., p. 3).

	Images seen per category	Success	Success rate
People	123,138	79,538	0.65
Vehicles	53,728	33,665	0.63
Objects	229,130	142,144	0.62
Urban	138,535	84,684	0.61
Nature	110,485	65,083	0.59

Figura 2.1.3 – Accuratezza di riconoscimento delle immagini reali e sintetiche per categoria di contenuto visivo: people, vehicles, objects, urban, nature (Roca et al., 2025).

Come mostrato in fig. 2.1.3, l'accuratezza di riconoscimento varia in base alla categoria di contenuto, risultando più elevata per le immagini raffiguranti persone e progressivamente più bassa per immagini di ambienti urbani e naturali. Coerentemente con il lavoro di Lu et al., emerge dunque che le persone hanno un'elevata capacità di riconoscere i volti, le cui "anomalie sono più facili da individuare rispetto ai paesaggi" (*ivi*), perché i modelli generativi sono molto efficaci su questa tipologia di contenuti, ingannando gli utenti specialmente con le immagini che sembrano realistiche ma non professionali (fig. 2.1.4).

Figura 2.1.4 – Esempi di immagini generate dall'IA con stile fotografico amatoriale, utilizzate dagli autori per illustrare come contenuti realistici ma non professionali risultino particolarmente ingannevoli per gli utenti. Le immagini, generate con Flux Pro, non sono state utilizzate nel gioco Real or Not Quiz (Roca et al., 2025).

Inoltre, effettuando una breve valutazione della qualità delle immagini (*Image Quality Assessment*), è emerso che molti partecipanti identificano erroneamente immagini reali lasciandosi guidare da segnali poco affidabili, come una saturazione del colore troppo elevata, luci troppo innaturali o un'estetica particolarmente "pulita", tipica delle immagini IA.

Al contrario, le immagini IA più difficili da riconoscere risultano essere quelle più "naturali", meno rifinite (*ivi*, p. 7).

Accanto ai fattori percettivi ci sono anche elementi cognitivi e demografici. Il lavoro di Totti et al. (2024), che analizza come gli utenti valutano la veridicità di notizie accompagnate da immagini reali o sintetiche, mostra che l'età, la familiarità con l'IA e l'autostima hanno un ruolo rilevante. Gli individui più anziani sbagliano di più e dichiarano minore fiducia nelle proprie capacità, mentre i più giovani tendono a sopravvalutare le proprie competenze, sentendosi più capaci di quanto non siano realmente. Lo studio mette in evidenza che la difficoltà non dipende solo dall'immagine in sé, ma anche dalle aspettative e dalle strategie che ogni persona utilizza nel momento in cui deve giudicare un contenuto.

Infine, il lavoro di Li et al. (2025) offre un quadro più preciso sui segnali visivi usati dalle persone per giudicare l'autenticità di un'immagine. Dal grafico (fig. 2.1.5) emerge che i partecipanti si basano maggiormente su indizi superficiali ("*low-level features*") come la texture (14,27%), il colore complessivo (12,65%), la chiarezza (11,60%) o i bordi (8,17%).

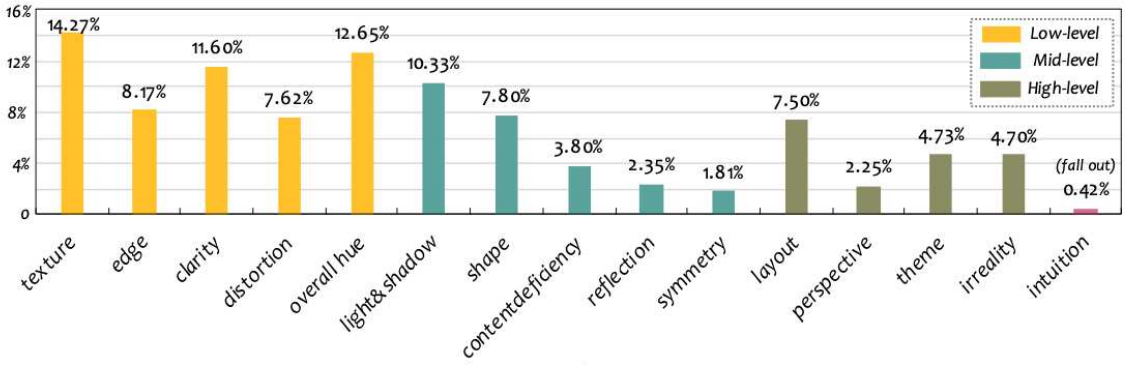


Figura 2.1.5 – Tassonomia dei criteri utilizzati dagli utenti per valutare l'autenticità delle immagini, suddivisi in segnali low-level, mid-level e high-level, con statistiche di utilizzo (Li et al., 2025).

Al contrario, criteri come l'analisi delle ombre o delle forme ("*mid-level features*") compaiono molto meno, mentre elementi come il layout o la coerenza del tema ("*high-level features*") richiedono uno sforzo cognitivo maggiore e vengono quindi utilizzati molto raramente. Il dato interessante è che quasi nessuno usa la pura intuizione: gli utenti non si affidano ad una sensazione generale, ma a dettagli che spesso possono risultare fuorvianti. La tassonomia sviluppata dagli autori conferma che gli esseri umani tendono a basarsi su segnali che ritengono

importanti ma che non sono necessariamente affidabili, evidenziando l'utilità di un intervento di media literacy che possa riorientare l'attenzione verso indizi più validi. Lo studio mostra anche un fenomeno importante per questa tesi: dopo molte esposizioni a contenuti sintetici, alcune persone iniziano a fidarsi meno anche delle immagini reali. Questo rischio di "*over-skepticism*" è cruciale per progettare interventi educativi equilibrati, che aiutino a riconoscere i contenuti sintetici senza però generare una sfiducia generale.

Nel complesso, la letteratura analizzata evidenzia come la difficoltà nel distinguere immagini reali e sintetiche non sia riconducibile unicamente al livello di sofisticazione tecnica dei modelli generativi, ma derivi dalla combinazione tra caratteristiche percettive delle immagini, tipo di contenuto rappresentato e strategie di valutazione adottate dagli individui. In particolare, le immagini che raffigurano scene quotidiane, ambienti urbani, eventi naturali o situazioni di cronaca risultano particolarmente ingannevoli e più difficili da classificare, poiché gli artefatti tipici della generazione artificiale sono meno evidenti e più difficili da isolare. Questo tipo di contenuti, inoltre, è lo stesso che circola più frequentemente sulle piattaforme social, contribuendo a rendere il problema non solo teorico, ma profondamente legato alle pratiche quotidiane di fruizione delle immagini online.

Un ulteriore elemento critico riguarda il modo in cui le persone costruiscono il proprio giudizio di autenticità. Come mostrato dagli studi, gli utenti non si affidano a un'intuizione globale, ma cercano indizi specifici che ritengono informativi, spesso concentrandosi su aspetti superficiali o estetici che non sempre si rivelano affidabili. Questo approccio, se da un lato riflette un tentativo razionale di valutazione, dall'altro espone al rischio di errori sistematici, soprattutto in un contesto in cui i modelli generativi sono sempre più efficaci nel produrre immagini visivamente plausibili ma non professionali. La conseguenza è una crescente incertezza percettiva, che può tradursi sia in una sovrastima dell'autenticità delle immagini sintetiche sia, al contrario, in una progressiva sfiducia nei confronti delle immagini reali.

Questi risultati suggeriscono che il riconoscimento delle immagini generate dall'intelligenza artificiale non può essere affrontato esclusivamente come un problema di percezione individuale, ma deve essere considerato all'interno di un ecosistema più ampio, in cui interagiscono limiti umani, tecnologie di rilevamento automatico e competenze critiche. Da un lato, emerge l'esigenza di strumenti tecnici in grado di supportare il riconoscimento dei contenuti generati dall'IA; dall'altro, diventa evidente la necessità di comprendere in che modo le persone interpretano e valutano le immagini, soprattutto quando queste vengono inserite in contesti comunicativi complessi come quelli dei social media. Su queste basi si innestano, nel paragrafo successivo, le riflessioni sui progressi e sui limiti dei detector automatici e,

successivamente, sul ruolo della media literacy come possibile strategia complementare per affrontare le sfide poste dalla diffusione delle immagini sintetiche.

2.2 Detector automatici: progressi e limiti

Se da un lato gli esseri umani faticano a riconoscere le immagini generate dall'intelligenza artificiale, dall'altro ci si aspetterebbe che i sistemi automatici possano compensare queste difficoltà. Negli ultimi anni sono stati sviluppati molti modelli di detection capaci di identificare contenuti sintetici in modo rapido e spesso più accurato rispetto all'occhio umano. Tuttavia, come mostra la letteratura tecnica recente (Dehghani & Saberi, 2025), la detection è diventata sempre più complessa a causa del rapido miglioramento dei modelli generativi. Così, la relazione tra generazione e rilevazione diventa una "corsa agli armamenti" ("*arms race*"), poiché la prima cresce più velocemente della seconda: "il rapido ritmo dell'innovazione nella generazione di deepfake richiede una costante evoluzione delle tecniche di rilevamento" (ivi, sez. *Background and related work*). In questo senso, ogni progresso nell'IA generativa riduce l'efficacia degli strumenti di riconoscimento.

Un primo elemento critico riguarda l'evoluzione stessa dei contenuti sintetici. Le prime tecniche basate su GAN presentavano difetti riconoscibili, come incoerenze nella texture o nei bordi, che potevano essere identificati tramite metodi tradizionali (*ivi*, sez. *Introduction*). Con l'arrivo dei modelli di diffusione, questo approccio è diventato meno efficace: la qualità delle immagini sintetiche è cresciuta enormemente e i nuovi modelli sono in grado, ad esempio, di generare scambi di identità, sincronizzazione labiale e manipolazioni facciali altamente realistiche (*ivi*, sez. *Diffusion models*). Questo cambiamento rende la maggior parte dei detector classici meno affidabili, soprattutto quando devono riconoscere immagini provenienti da modelli non presenti nel dataset di addestramento.

Uno dei problemi più discussi è proprio la scarsa generalizzazione. Molti detector funzionano bene sui dataset su cui vengono addestrati, ma male di fronte a immagini nuove, magari prodotte da modelli o tecniche leggermente differenti o più recenti. Gli autori sottolineano che "le attuali strategie di rilevamento, sebbene in fase di miglioramento, continuano a riscontrare difficoltà nell'identificare deepfake sempre più realistici, evidenziando la necessità di ulteriori ricerche su algoritmi di rilevamento più robusti e generalizzabili" (*ivi*, sez. *Conclusions*). Perciò, è necessario sviluppare algoritmi di detection più efficaci e, allo stesso tempo, incrementare programmi di educazione alla media literacy (*ivi*).

Gli studi più recenti propongono soluzioni alternative ai classici detector. ZeroFake (Sha et al., 2024), ad esempio, non si basa su un classificatore tradizionale, ma sfrutta il comportamento dei modelli di diffusione durante il processo di inversione, introducendo una strategia *zero-shot* che non richiede un addestramento specifico sul dataset. L'idea alla base è la seguente: nel momento in cui le immagini vengono manipolate, quelle generate dall'IA mantengono una maggiore stabilità interna rispetto a quelle reali, poiché prodotte dalla stessa tipologia di modello (fig. 2.2.1) (*ivi*, sez. *Our contribution*). Gli autori mostrano che, dopo l'inversione, le immagini sintetiche conservano più caratteristiche dell'originale rispetto alle immagini reali, che invece risultano più distorte. In questo modo, è possibile distinguere reale e fake osservando la risposta dell'immagine al processo generativo. I risultati indicano che ZeroFake performa molto bene anche su modelli non visti in fase di addestramento, raggiungendo ad esempio un'accuracy di 0,957 nella detection di immagini create da Stable Diffusion, superando i metodi che sono basati su classificatori tradizionali (*ivi*, sez. *Results*).

Lo studio evidenzia come la detection possa essere affrontata anche sfruttando le proprietà interne dei modelli generativi. Nonostante ciò, il metodo conferma che la detection automatica dipende ancora dall'evoluzione delle tecniche generative, e che deve essere ripensata alla luce di questo fatto.

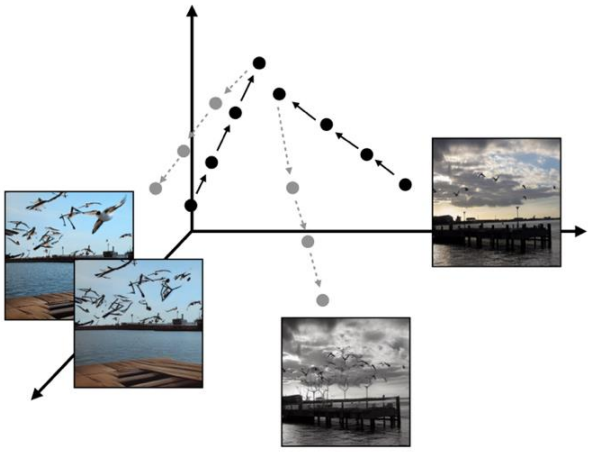


Figura 2.2.1 – Rappresentazione concettuale del metodo ZeroFake: durante il processo di inversione dei modelli di diffusione, le immagini generate dall'IA mostrano traiettorie più stabili nello spazio latente rispetto alle immagini reali, consentendo la detection in modalità zero-shot. (Sha et al., 2024).

Parallelamente, la ricerca più recente si è spostata verso detector in grado non solo di classificare un'immagine come reale o sintetica, ma anche di spiegare il perché della scelta. È il caso di FakeBench (Li et al., 2025) e del lavoro di Wen et al. (2025), che valutano la capacità dei modelli multimodali di fornire spiegazioni comprensibili riguardo alle proprie decisioni.

FakeBench introduce un benchmark che permette di analizzare le capacità dei Large Multimodal Models (LMM) in quattro diversi compiti legati al riconoscimento delle immagini IA: detection, interpretazione, reasoning (costruzione di una catena causale esplicita) e analisi fine-grained (capacità di individuare artefatti specifici) degli indizi visivi (fig. 2.2.2) (Li et al., 2025, p. 2).

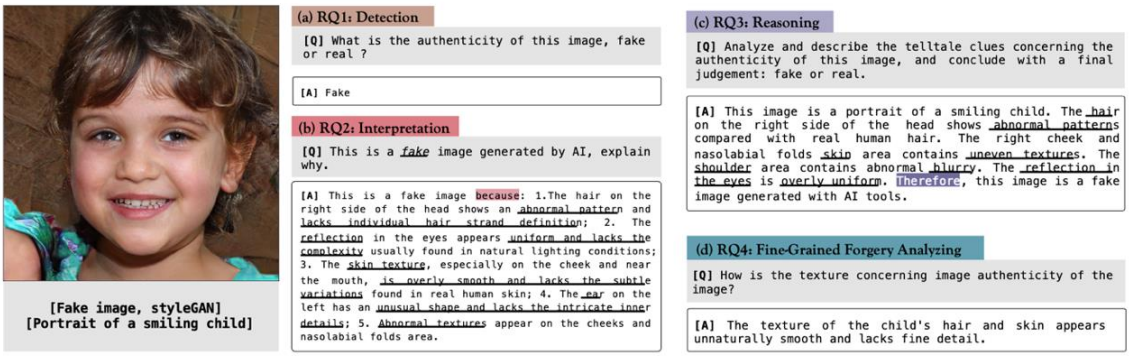


Figura 2.2.2 – Ambito delle domande di ricerca del dataset FakeBench, con esempi di coppie domanda-risposta per i quattro compiti principali: detection, interpretazione, reasoning e analisi fine-grained.
Le risposte mostrate sono fornite da annotatori umani (Li et al., 2025).

Confrontando i risultati con la performance umana, emerge che la maggior parte dei LMM performa peggio degli esseri umani, scendendo al di sotto della soglia minima umana (*ivi*, tab. V), e soltanto pochi modelli (come GPT-4V) si avvicinano o superano leggermente la performance umana. Inoltre, le difficoltà dei modelli aumentano quando viene chiesto loro di interpretare e fornire spiegazioni. Il reasoning risulta molto più difficile rispetto all'interpretazione (*ivi*, tab. VI, tab. VII). In generale, i modelli riescono più facilmente a descrivere indizi di autenticità quando sono guidati dall'immagine, ma trovano difficoltà nella costruzione di spiegazioni coerenti che colleghino tali indizi alla decisione finale, ovvero nell'interpretazione. Anche nell'analisi fine-grained emergono delle criticità: nel compito FakeQA, che richiede di analizzare aspetti specifici della falsificazione, nessun modello raggiunge punteggi elevati. Infatti, le prestazioni migliori restano al di sotto di 0,5 su una scala normalizzata tra 0 e 1, indicando una chiara difficoltà nell'analisi dettagliata degli artefatti (*ivi*, tab. VIII). Lo studio mette in risalto il dislivello tra accuratezza di detection e comprensione dell'autenticità, poiché anche quando i modelli multimodali riescono a distinguere immagini reali e sintetiche, hanno difficoltà nel fornire spiegazioni sui motivi della loro decisione.
Il lavoro di Wen et al. ottiene risultati simili. Gli autori si propongono di sviluppare un modello multimodale (FakeVLM) capace non solo di distinguere immagini reali e sintetiche, ma anche di spiegare quali artefatti visivi hanno portato alla classificazione. Dai risultati emerge che

FakeVLM supera sia modelli general-purpose che modelli specializzati nella detection binaria e, in alcuni casi, anche la performance umana, come nel dataset LOKI, in cui ottiene un'accuratezza pari all'84,3%, contro l'accuratezza delle performance umane pari all'80,1% (Wen et al., 2025, p. 7). Per quanto riguarda la spiegazione degli artefatti, il modello è in grado di identificare correttamente elementi come texture incoerenti, distorsioni geometriche, anomalie nell'illuminazione e composizioni innaturali, fornendo descrizioni che risultano più vicine alle annotazioni umane rispetto ai baseline. La figura 2.2.3 illustra il funzionamento di un detector multimodale con spiegazione degli artefatti: il modello è progettato per identificare un'immagine come reale o sintetica e fornire una descrizione degli indizi visivi che supportano la decisione, come incoerenza nella texture, nel colore o nello sfondo. Tuttavia, come evidenziato dagli autori, la presenza di spiegazioni non implica necessariamente che il modello abbia realmente compreso il processo decisionale. Le spiegazioni possono risultare plausibili ma generiche, o non direttamente collegate ai fattori causali che hanno guidato la decisione.
Questo aspetto rafforza l'importanza che ricopre l'interpretabilità all'interno della detection: spiegare perché qualcosa è sintetico è cruciale sia per i modelli che per le persone.

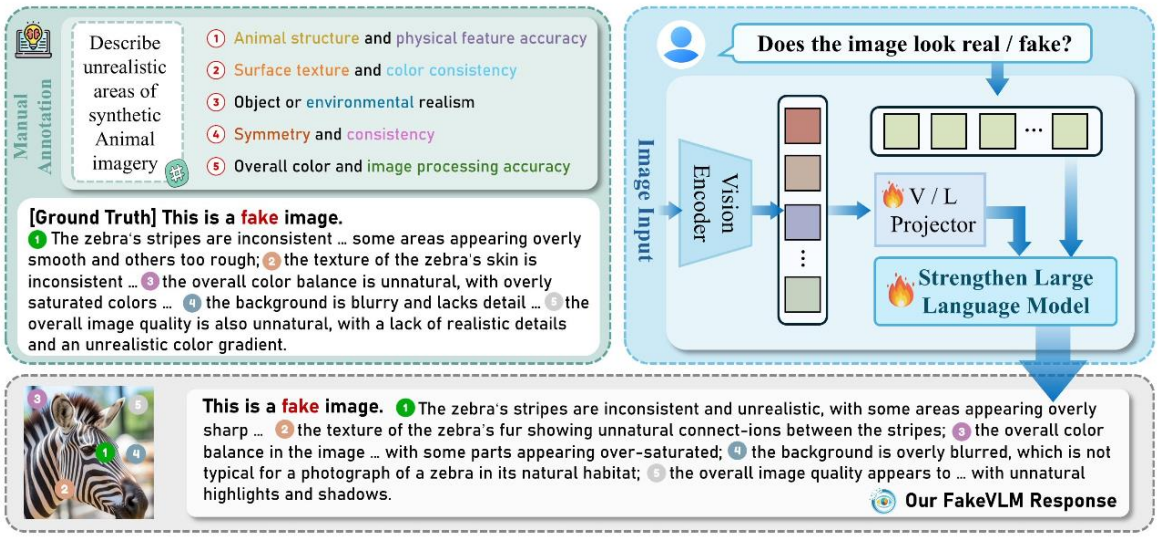


Figura 2.2.3 – Architettura del modello FakeVLM per la detection di immagini sintetiche e la spiegazione degli artefatti visivi. Il sistema combina rappresentazioni visive e linguistiche per supportare la classificazione e la descrizione degli indizi di falsificazione (Wen et al., 2025).

Di conseguenza, la detection automatica non può sostituire la valutazione umana, ma è necessario che sia affiancata da competenze critiche e interventi di media literacy.
Infine, alcuni studi mostrano che anche i modelli più sofisticati possono basarsi su segnali fuorvianti. Tanfoni et al. (2024) evidenziano che molti di questi non rilevano il fake basandosi sul volto, ma analizzando soprattutto informazioni presenti nello sfondo dell'immagine. In

particolare, per il modello StyleGAN2 le regioni più importanti sono spesso nello sfondo o ai bordi del viso: quando il background viene rimosso, l'accuratezza cala e le prestazioni peggiorano. Al contrario, per StyleGAN3 l'attenzione si sposta sul volto (occhi-naso-bocca), infatti rimuovere lo sfondo, che risulta distrattivo per il modello, migliora o comunque non peggiora le sue performance (fig. 2.2.4) (Tanfoni et al., 2024, p. 9). Questo risultato è rilevante perché dimostra che anche i sistemi automatici possono utilizzare "scorciatoie" non generalizzabili, simili a quelle che gli esseri umani usano inconsapevolmente.

Generator	Background	Accuracy
StyleGAN2	Yes	**0.9495**
	No	0.9022
StyleGAN3	Yes	0.8373
	No	**0.8499**

Figura 2.2.4 – Accuratezza della detection per immagini generate con StyleGAN2 e StyleGAN3 in presenza o assenza dello sfondo. I risultati mostrano come il background influisca in modo significativo sulle prestazioni di riconoscimento, soprattutto per le immagini generate con StyleGAN2 (Tanfoni et al., 2024).

Nell'insieme, la letteratura mostra che i detector automatici hanno compiuto progressi significativi nel riconoscimento delle immagini generate dall'intelligenza artificiale, ma evidenzia anche limiti strutturali che ne riducono l'affidabilità in contesti reali e dinamici. Da una parte, alcuni modelli raggiungono livelli di accuratezza elevati, talvolta superiori alla performance umana, soprattutto in compiti di classificazione binaria. Dall'altra, tali risultati risultano fortemente dipendenti dal tipo di modello generativo considerato, dal dataset di riferimento e dalla stabilità degli indizi utilizzati per la decisione.

Il confronto con la valutazione umana mette in luce un'importante tensione: se gli esseri umani faticano a riconoscere immagini sintetiche a causa di limiti percettivi e strategie di giudizio poco affidabili, anche i sistemi automatici possono basarsi su scorciatoie non generalizzabili, come elementi di sfondo o pattern specifici appresi in fase di addestramento. Inoltre, come evidenziato dagli studi sui modelli multimodali, un'elevata accuratezza di detection non implica necessariamente una reale comprensione del contenuto visivo né la capacità di spiegare in modo coerente i fattori che hanno guidato la decisione.

Dai risultati emerge che la detection automatica, pur rappresentando uno strumento utile, non può essere considerata una soluzione definitiva al problema del riconoscimento delle immagini sintetiche. In particolare, nei contesti di fruizione quotidiana delle immagini – come i social media – l'affidamento esclusivo a sistemi automatici appare insufficiente, sia per i limiti di generalizzazione, sia per la difficoltà di integrare tali strumenti nelle pratiche ordinarie degli

utenti. Di conseguenza, emerge la necessità di affiancare alle soluzioni tecniche un'attenzione specifica alle competenze critiche delle persone, aprendo lo spazio per una riflessione sul ruolo della media literacy come possibile strategia complementare.

2.3 Media literacy e interventi: cosa funziona

Negli ultimi anni la diffusione di contenuti sintetici ha reso sempre più difficile orientarsi nell'ecosistema informativo. L'istinto o la percezione visiva umani non sono più sufficienti, ma anche i detector automatici hanno dei limiti significativi, come visto nella sezione precedente.
Per questo motivo la media literacy è considerata una delle strategie più efficaci per aiutare le persone a riconoscere contenuti manipolati o generati dall'IA. La domanda centrale è se interventi brevi, facilmente replicabili e accessibili a tutti possano davvero migliorare la capacità di distinguere il reale dal sintetico.
Uno degli studi più completi su questo tema è quello di Geissler et al. (2025), che ha confrontato cinque diversi tipi di intervento: testuale, visivo, gamificato, basato sul feedback e conoscitivo. Il primo è basato su descrizioni scritte degli errori tipici dei deepfake, il secondo abbina le spiegazioni a esempi concreti, nel terzo gli utenti individuano attivamente gli errori, il quarto si fonda sull'apprendimento implicito tramite ripetizione e l'ultimo spiega come funzionano i modelli generativi (Geissler et al., 2025, sez. 2.4). I risultati mostrano che l'intervento più efficace è quello visivo, composto da esempi concreti di errori tipici del deepfake accompagnati da brevi spiegazioni. In particolare, questo metodo aumenta l'accuratezza nel riconoscimento dei contenuti sintetici di circa 13 punti percentuali (Geissler et al., 2025, p. 25) e riduce anche l'intenzione di condividere le immagini false (–5,18%) (*ivi*, p. 29). Tuttavia, gli effetti risultano immediati ma limitati nel tempo: nel follow-up a due settimane dall'intervento, le differenze rispetto al gruppo di controllo tendono a scomparire e nessuno dei cinque interventi mantiene un miglioramento statisticamente significativo (fig. 2.3.1). Un risultato importante dello studio è che gli interventi di media literacy non riducono la fiducia nelle immagini reali, evitando così il rischio di un eccesso di scetticismo (*ivi*, p. 27).

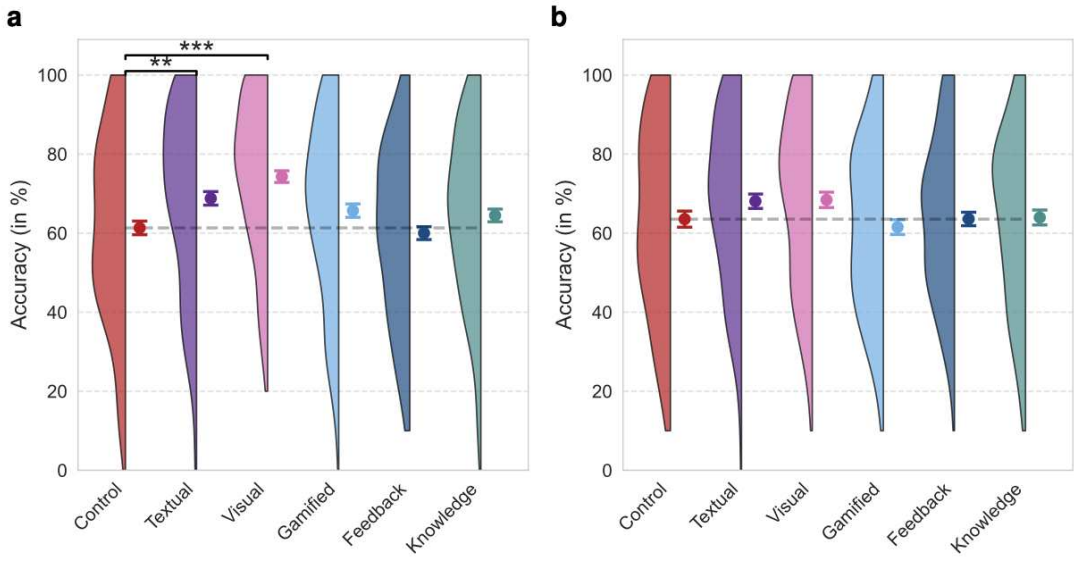


Figura 2.3.1 – Accuratezza nel riconoscimento di immagini deepfake nei diversi interventi di media literacy, misurata immediatamente dopo l'intervento e a due settimane di distanza. L'intervento visivo mostra un miglioramento significativo nel breve periodo, che tende a scomparire nel follow-up (Geissler et al., 2025).

Anche la ricerca di Hwang et al. (2021) mette in evidenza l'efficacia della media literacy nel contrastare la disinformazione basata sul deepfake. Gli autori mostrano che i messaggi di disinformazione accompagnati da un deepfake video sono più vividi, persuasivi e credibili rispetto alle versioni con il solo testo, aumentando anche l'intenzione di condivisione del contenuto (Hwang et al., 2021, sez. *Results*). Come stimolo viene utilizzata una fake news su Mark Zuckerberg, presentata come articolo di giornale di 69 parole, attribuito ad una fonte fittizia ("Cheil News"), in due versioni: una testuale e una accompagnata da un deepfake video di 90 secondi (*ivi*, sez. *Experimental materials*). Prima dell'esposizione al messaggio, i partecipanti ricevono un intervento di media literacy della durata di circa sette minuti, in due varianti: una generale sulla definizione di disinformazione, con esempi e conseguenze sociali, e una specifica sui deepfake, con la definizione di deepfake, esempi e relativi impatti sociali (*ivi*). Dai risultati emerge che entrambi i tipi di intervento riducono la vividezza percepita, la credibilità del messaggio e l'intenzione di condividerlo. È interessante notare che la media literacy generale è stata spesso più efficace rispetto a quella specifica sui deepfake, suggerendo che non è necessario fornire spiegazioni complesse o dettagliate per ottenere dei buoni risultati. Secondo gli autori "ciò potrebbe essere dovuto al fatto che l'educazione generale alla disinformazione è più efficace nel rafforzare le difese attitudinali, che sono state suggerite come una dimensione importante dell'alfabetizzazione mediatica" (*ivi*, sez. *Discussion*).
Anche in questo caso, l'intervento richiede pochi minuti ed è facilmente replicabile.

Un contributo simile arriva dallo studio di El Mokadem (2023), che analizza l'effetto di brevi lezioni di media literacy sulla capacità di riconoscere contenuti di disinformazione testuale e audiovisiva. In particolare, si utilizzano due tipi di messaggi: un post Facebook falso accompagnato da un'immagine manipolata e un deepfake video della durata di circa 16 secondi, entrambi riferiti a Mark Zuckerberg (fig. 2.3.2). Con un disegno quasi-sperimentale, lo studio coinvolge 204 partecipanti altamente istruiti, suddividendoli in tre gruppi: il primo viene sottoposto ad una lezione di media literacy generale sulla disinformazione, il secondo ad una lezione specifica sui deepfake, e il terzo a nessun intervento come gruppo di controllo (El Mokadem, 2023, sez. *Methodology*).

Figura 2.3.2 – Post Facebook falso utilizzato come stimolo sperimentale nello studio di El Mokadem (El Mokadem, 2023).

L'intervento, della durata di 10-15 minuti, migliora in modo significativo la capacità dei partecipanti di riconoscere i contenuti falsi, riduce la credibilità attribuita ad essi e l'intenzione di condividerli sui social media (*ivi*, sez. *Results*). In particolare, i partecipanti che hanno ricevuto una lezione di media literacy valutano i messaggi come meno credibili, meno accurati e meno persuasivi rispetto al gruppo di controllo, che non riceve nessun tipo di intervento educativo. Un risultato particolarmente interessante è che non ci sono differenze significative legate all'età o al livello di istruzione: studenti universitari e membri dello staff più anziani rispondono nello stesso modo, mostrando difficoltà simili nella capacità di detection di contenuti falsi (*ivi*). Inoltre, gli interventi specifici sui deepfake risultano particolarmente efficaci nel caso dei video, confermando che per i contenuti audiovisivi sono necessarie strategie di educazione mirate (*ivi*, sez. *Discussion*).

Anche questa ricerca mostra come interventi brevi e facilmente replicabili possano migliorare la capacità di detection, evidenziando l'importanza del ricorso alla media literacy per poter valutare correttamente e con sguardo critico i contenuti digitali.

Accanto agli studi che mostrano l'efficacia degli interventi di media literacy, una parte della letteratura mette in discussione l'assunto secondo cui il possesso di competenze di information literacy si traduca automaticamente in una capacità efficace di valutare contenuti visivi. In particolare, ricerche recenti evidenziano una significativa incongruenza tra le competenze dichiarate dagli utenti e le loro performance reali nel riconoscere immagini fuorvianti o manipolate. Analizzando il caso della Generazione Z, Zak (2024) mostra come, pur essendo cresciuti in un ecosistema digitale e avendo ricevuto una formazione di base sull'information literacy, gli studenti universitari incontrino difficoltà specifiche nell'identificare la visual misinformation, un ambito che risulta ancora sottovalutato nei programmi educativi tradizionali (*ivi*, p. v). Questa discordanza tra percezione delle proprie abilità e capacità effettive è coerente con risultati emersi in studi precedenti sull'information literacy. Come riportato dall'autrice, ricerche antecedenti mostrano che studenti universitari e laureandi possiedono una comprensione di base delle competenze informative, ma necessitano di ulteriore formazione per applicarle in modo efficace (Zhao, 2019, cit. in Zak, 2024, p. 32). Risultati analoghi emergono anche in studi che evidenziano un'elevata fiducia dichiarata nelle proprie capacità di valutare le informazioni online, pur applicando tali competenze in modo limitato e spesso riduttivo, concentrandosi su strumenti familiari come motori di ricerca e social media (Geary, 2021, cit. in Zak, 2024, p. 32). Questo insieme di studi suggerisce l'esistenza di un gap strutturale tra autovalutazione e pratica reale, particolarmente evidente quando il giudizio riguarda contenuti visivi.

All'interno di questo quadro, la visual literacy emerge come una componente distinta e ancora poco sviluppata dell'information literacy. Sebbene diversi autori sottolineino che l'uso di competenze informative possa ridurre la diffusione della disinformazione, la maggior parte della ricerca si concentra su contenuti testuali, mentre il dominio visivo rimane relativamente inesplorato (*ivi*, p. 21). Come osserva l'autrice, molti studenti ritengono di essere visivamente competenti, ma faticano a riconoscere segnali di manipolazione nelle immagini, soprattutto quando queste appaiono realistiche o coerenti con le aspettative contestuali. Tale mancanza di visual literacy si riflette non solo nella valutazione delle immagini online, ma anche nell'uso non critico di materiali visivi in presentazioni e contesti pubblici (*ivi*, p. 32).

In sintesi, questi risultati suggeriscono che la media literacy non possa essere considerata un insieme uniforme di competenze, ma piuttosto una variabile individuale che incide in modo

differenziato sulla capacità di valutare contenuti visivi, sulla sicurezza nel giudizio e sui comportamenti di condivisione online. Come sottolineato nella letteratura, la persistenza di questo divario evidenzia la necessità di integrare esplicitamente la visual literacy all'interno dei modelli di information literacy, affinché gli utenti possano sviluppare strumenti critici adeguati ad affrontare la crescente diffusione di immagini sintetiche e fuorvianti nei contesti digitali contemporanei (*ivi*, p. 38).

Un altro elemento rilevante per comprendere il ruolo delle immagini nella disinformazione riguarda il rapporto tra immagine e contesto testuale. Come mostrano Newman e Schwarz (2024), le immagini influenzano in modo significativo il modo in cui le persone valutano la veridicità di un contenuto, anche quando non sono false o manipolate. Infatti, gli autori sottolineano che una fotografia non deve necessariamente essere alterata o generata artificialmente per risultare fuorviante: anche immagini reali possono orientare l'interpretazione di un messaggio in modo distorto, ad esempio quando vengono riutilizzate fuori contesto o associate a un testo che suggerisce un significato diverso da quello originario (Newman & Schwarz, 2024, p. 1).

Secondo gli autori, il punto centrale non è l'immagine in sé, ma la relazione che si crea tra immagine e testo. Le persone tendono a dare per scontato che gli elementi di un messaggio siano collegati in modo coerente e intenzionale: se un'immagine accompagna un testo, viene percepita come rilevante e informativa rispetto a ciò che viene affermato. Questo porta gli utenti a trarre inferenze che collegano immagine e contenuto testuale, anche quando l'immagine non fornisce alcuna prova reale a supporto dell'affermazione (*ivi*, p. 2). In questo senso, il contesto testuale – come una caption o un titolo – può guidare l'interpretazione dell'immagine e contribuire a rendere un messaggio più credibile.

Un aspetto particolarmente interessante riguarda il ruolo delle immagini cosiddette "decorative", ovvero immagini di repertorio o stock photo che sono solo semanticamente collegate al testo. Anche se queste immagini non hanno valore probatorio, la loro presenza può aumentare l'accettazione del messaggio attraverso un meccanismo di maggiore facilità di elaborazione cognitiva, definito come *processing fluency*. Quando un contenuto risulta più facile da comprendere o da immaginare, le persone tendono a giudicarlo come più vero, indipendentemente dalla sua accuratezza (*ivi*, p. 3). Questo effetto, noto come *truthiness effect*, è stato osservato in diversi ambiti e mostra come anche immagini apparentemente innocue possano influenzare il giudizio di veridicità (fig. 2.3.4).

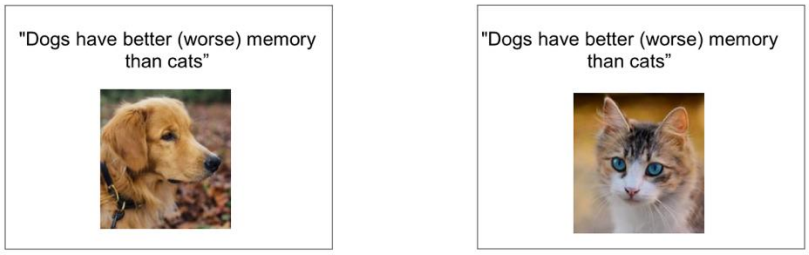


Figura 2.3.4 – La presenza dell'immagine del soggetto del confronto (cane) facilità l'elaborazione dell'affermazione e ne aumenta l'accettazione come vera (truthiness effect), mentre l'immagine del termine di riferimento (gatto) riduce la facilità di elaborazione e la probabilità che l'affermazione venga giudicata vera (falsiness effect). L'effetto è dovuto all'influenza delle immagini sulla processing fluency del confronto (Newman & Schwarz, 2024, p. 3).

Inoltre, le immagini catturano maggiormente l'attenzione e aumentano la probabilità che un contenuto venga visto e condiviso. Studi basati su *eye-tracking* mostrano che i messaggi accompagnati da immagini attirano l'attenzione più rapidamente e la mantengono più a lungo rispetto a quelli solo testuali (*ivi*, p. 2). Questo contribuisce a spiegare perché i post con immagini abbiano una maggiore diffusione sui social media, indipendentemente dal fatto che il contenuto sia vero o falso. In questo senso, le immagini non solo influenzano la percezione di credibilità, ma incidono anche sui comportamenti di condivisione.

Un ulteriore effetto rilevante riguarda la memoria. Newman e Schwarz mostrano che le immagini possono rafforzare la comprensione di un messaggio, ma anche favorire la formazione di ricordi errati. Quando testo e immagine vengono elaborati come un'unica unità, le persone possono ricordare informazioni che non erano esplicitamente presenti nel testo, ma che erano suggerite visivamente dall'immagine (*ivi*, p. 3). Questo vale sia per immagini chiaramente fuorvianti, sia per immagini reali che offrono una rappresentazione parziale o "sbilanciata" di un evento.

Per concludere, gli autori evidenziano un aspetto particolarmente critico: molti di questi effetti avvengono al di fuori della consapevolezza degli utenti. Anche quando le persone sono informate sui rischi della disinformazione visiva, l'influenza delle immagini può persistere perché agisce su processi automatici come l'attenzione, la fluidità di elaborazione e la memoria (*ivi*, p. 4). Questo rende difficile contrastare l'effetto delle immagini attraverso semplici interventi di educazione critica o avvisi, e aiuta a spiegare perché la media literacy tradizionale, focalizzata soprattutto sul testo, possa risultare meno efficace nel contesto visivo.

In sintesi, questo contributo evidenzia come il contesto testuale giochi un ruolo fondamentale nel modo in cui le immagini vengono interpretate e valutate. La combinazione di immagine e testo può aumentare la credibilità percepita di un messaggio e la sua probabilità di condivisione, anche in assenza di manipolazioni evidenti. Questi risultati sono particolarmente rilevanti per

l’analisi dei contenuti sui social media, dove immagini e caption sono quasi sempre presentate insieme, e forniscono una base teorica solida per indagare sperimentalmente l’effetto del contesto testuale sulla valutazione di immagini reali e generate dall’intelligenza artificiale.

3
METODOLOGIA

3.1 Obiettivi, ipotesi e disegno della ricerca

Alla luce della letteratura analizzata nel capitolo precedente, emerge come la capacità di distinguere immagini reali da immagini generate dall'intelligenza artificiale sia spesso limitata, anche in presenza di utenti giovani e con competenze digitali (Nightingale & Farid, 2022; Lu et al., 2023; Roca et al., 2025). Inoltre, diversi studi hanno mostrato che il contesto in cui un'immagine viene presentata può influenzare la percezione di credibilità e la sua accettazione come vera (Newman & Schwarz, 2024). Parallelamente, la ricerca sulla media literacy suggerisce che competenze critiche più sviluppate possano migliorare la capacità di valutare contenuti digitali, anche se tale effetto non è sempre stabile nel tempo (Geissler et al., 2025; Zak, 2024).

Sulla base di queste evidenze, l'obiettivo principale di questa ricerca è analizzare la capacità degli utenti di distinguere immagini reali da immagini generate dall'intelligenza artificiale in un contesto simile a quello dei social media, valutando il ruolo della caption testuale e di alcune differenze individuali. In particolare, si intende verificare:

- se la presenza di una caption influenzi l'accuratezza del giudizio;
- se incida sul livello di sicurezza dichiarato nella risposta;
- quale sia la propensione degli utenti a condividere sui social media immagini presentate in formato post;
- se il livello di media literacy e la riflessività cognitiva (misurata tramite CRT) siano collegati a una migliore capacità di riconoscimento.

Accanto a questi obiettivi principali, lo studio esplora anche alcune dinamiche legate al comportamento di condivisione e alla percezione di autenticità delle immagini.

Vengono formulate le seguenti ipotesi:

H1. L'accuratezza nel distinguere immagini reali da immagini generate dall'IA non sarà significativamente superiore al 70%, in linea con quanto emerso nella letteratura precedente.

H2. La presenza di una caption influenzerà l'accuratezza della classificazione rispetto alla condizione senza caption.

H3. La presenza di una caption influenzerà il livello di sicurezza dichiarato nella risposta.

H4. Le immagini percepite come reali saranno associate a una maggiore intenzione di condivisione rispetto a quelle percepite come generate da IA.
H5. Un livello più alto di media literacy sarà associato a una maggiore accuratezza nel riconoscimento delle immagini generate artificialmente.
H6. Un punteggio più alto al Cognitive Reflection Test (CRT) sarà associato a una maggiore accuratezza nel distinguere immagini reali e sintetiche.
Lo studio considera inoltre due quesiti di ricerca esplorativi:
RQ1. La propensione alla condivisione è associata al livello di sicurezza percepita nella classificazione delle immagini?
RQ2. I partecipanti mostrano un bias sistematico nel classificare le immagini come reali piuttosto che come generate da IA?
Dal punto di vista metodologico, lo studio adotta un disegno sperimentale con due versioni del questionario. Ogni versione comprende 12 immagini totali: 6 presentate senza caption e 6 accompagnate da caption. All'interno di ciascuna condizione (con caption / senza caption) sono presenti 3 immagini reali e 3 immagini generate dall'IA.
Il campione è suddiviso in due gruppi: nella prima versione alcune immagini sono presentate con caption e altre senza, mentre nella seconda versione la presenza della caption è invertita per le stesse immagini. In questo modo è possibile isolare l'effetto della caption mantenendo costanti gli stimoli visivi e controllando eventuali differenze legate al contenuto specifico delle immagini.

3.2 Campione

Il campione dello studio è composto da partecipanti reclutati tramite una strategia di campionamento di convenienza, coerente con la natura esplorativa della ricerca e con la somministrazione online del questionario.
La partecipazione è stata volontaria e anonima. Il questionario è stato diffuso principalmente attraverso gruppi WhatsApp (universitari e personali) e tramite social network (in particolare, tramite condivisione sul profilo Instagram dell'autrice). Per partecipare era necessario avere almeno 18 anni, come indicato nella sezione introduttiva del questionario.
In totale, sono state raccolte 202 risposte, distribuite tra le due versioni del questionario: 100 risposte per la versione A e 102 risposte per la versione B.
Per quanto riguarda le caratteristiche socio-demografiche, l'età viene rilevata attraverso una domanda a scelta multipla (18–24; 25–34; 35–44; 45+). Ulteriori informazioni sul profilo dei

partecipanti sono state rilevate tramite una sezione specifica del questionario, descritta nel paragrafo dedicato agli strumenti.
Il campione non è probabilistico e non consente generalizzazioni statistiche alla popolazione generale.

3.3 Strumenti

3.3.1 Struttura generale del questionario

Il questionario è stato costruito e somministrato tramite la piattaforma Google Forms. La partecipazione è avvenuta in modalità online e anonima. Prima di accedere alle domande, ai partecipanti veniva presentata una breve informativa sullo scopo della ricerca e veniva richiesto il consenso alla partecipazione, con l'indicazione che era necessario avere almeno 18 anni.
Il questionario era composto da più sezioni, organizzate in modo sequenziale. In primo luogo, veniva somministrato il Cognitive Reflection Test (CRT) composto da tre domande a risposta numerica aperta. Successivamente, era presente una sezione dedicata alla media literacy, con cinque affermazioni valutate su scala Likert a 7 punti.
La parte centrale del questionario era costituita dalla valutazione delle immagini. In totale, ogni partecipante visualizzava 12 immagini: 6 presentate senza caption e 6 accompagnate da una caption testuale. All'interno di ciascuna condizione erano presenti 3 immagini reali e 3 immagini generate dall'intelligenza artificiale. Le immagini venivano mostrate una alla volta, ciascuna in una pagina separata, per evitare confronti diretti tra stimoli.
Infine, il questionario si concludeva con una breve sezione socio-demografica, composta da quattro domande relative all'età, al titolo di studio, alla frequenza di utilizzo dei social media e alla familiarità con immagini generate dall'IA.
Il tempo medio di compilazione stimato è di circa 8-10 minuti.

3.3.2 Valutazione delle immagini

La sezione centrale del questionario era dedicata alla valutazione di immagini fotografiche presentate in un formato simile a quello dei post sui social media. Ogni partecipante visualizzava un totale di 12 immagini, mostrate una alla volta in pagine separate, per evitare confronti diretti tra stimoli.
Le immagini erano suddivise in due condizioni:

- 6 immagini presentate senza caption;
- 6 immagini accompagnate da una caption testuale in stile social (con hashtag e tono informativo).

All'interno di ciascuna condizione erano presenti 3 immagini reali e 3 immagini generate tramite intelligenza artificiale. Le immagini coprivano macro-temi coerenti con contenuti tipici di cronaca e attualità, in modo da risultare plausibili e familiari nel contesto dei social media. I principali temi selezionati sono stati:

- Disastro naturale (alluvione, tornado);
- Incendio in contesto urbano;
- Evento meteorologico estremo (grandine, allagamenti);
- Protesta o manifestazione pubblica;
- Cronaca locale;
- Politica o geopolitica.

Per ciascun macro-tema era presente un'immagine reale e una generata da IA, in modo da mantenere un bilanciamento tematico tra le due tipologie di stimolo.

Le immagini reali sono state selezionate principalmente da archivi open source (quattro da Wikimedia Commons e due da Unsplash), privilegiando contenuti fotografici realistici, non professionali e coerenti con contesti di cronaca quotidiana. Le immagini sintetiche sono state generate tramite Stable Diffusion, seguendo criteri analoghi per soggetto e ambientazione, così da mantenere coerenza tra le versioni reali e quelle artificiali.

L'ordine di presentazione delle immagini è stato definito in modo da alternare contenuti reali e sintetici e da evitare la successione immediata di immagini appartenenti allo stesso tema. L'ordine era identico per tutti i partecipanti della stessa versione del questionario.

Per ciascuna immagine, ai partecipanti veniva richiesto di rispondere alle seguenti domande:

1. Classificazione dell'immagine

   "Secondo te, questa immagine è:"

   - Foto reale
   - Immagine generata da intelligenza artificiale

2. Livello di sicurezza nella risposta

   "Quanto sei sicuro/a della tua risposta?"

   Scala Likert a 7 punti (1 = per nulla sicuro/a; 7 = totalmente sicuro/a).

Per le immagini con caption era presente una terza domanda:

3. Intenzione di condivisione

"Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?"
Scala Likert a 7 punti (1 = per nulla probabile; 7 = molto probabile).

La domanda relativa all'intenzione di condivisione è stata presentata esclusivamente per le immagini accompagnate da caption. Di conseguenza, le analisi relative alla propensione alla condivisione sono state condotte solo su queste immagini.

3.3.3 Media literacy

Per misurare il livello di media literacy dei partecipanti è stata inserita una breve sezione composta da cinque affermazioni relative all'uso dei social media e alla valutazione delle informazioni online.

Le istruzioni erano le seguenti:

"Di seguito troverai alcune affermazioni relative al modo in cui utilizzi i social media e valuti le informazioni online.

Indica quanto sei d'accordo con ciascuna affermazione."

Le risposte sono state raccolte tramite una scala Likert a 7 punti (1 = per niente d'accordo; 7 = totalmente d'accordo).

Le affermazioni proposte erano:

1. "Prima di condividere un post sui social, controllo se la fonte è affidabile."
2. "Sono consapevole che molte immagini online possono essere generate o modificate con l'intelligenza artificiale."
3. "Ritengo di saper riconoscere alcuni segnali tipici di immagini generate artificialmente."
4. "Quando vedo una notizia online, cerco conferme da più fonti."
5. "Penso sia importante verificare le informazioni prima di condividerle sui social."

Questa sezione misura una media literacy percepita, cioè autodichiarata dai partecipanti, e non una competenza oggettivamente verificata tramite prove pratiche. Gli item sono stati costruiti sulla base della letteratura sulla media e information literacy analizzata nel capitolo teorico, con l'obiettivo di rilevare quanto gli individui si percepiscono consapevoli e attenti nel valutare le informazioni online e le immagini potenzialmente generate dall'IA.

Per ciascun partecipante è stato calcolato un indice medio di media literacy, ottenuto dalla media delle cinque risposte. Valori più alti indicano un livello più elevato di media literacy percepita.

3.3.4 Cognitive Reflection Test (CRT)

Per valutare la dimensione cognitiva legata al ragionamento riflessivo è stato inserito nel questionario il Cognitive Reflection Test (CRT) nella versione breve composta da tre item.
Il CRT è uno strumento ampiamente utilizzato in letteratura per misurare la tendenza degli individui a inibire una risposta intuitiva ma errata e ad attivare un processo di riflessione più analitico[2].
I partecipanti hanno risposto a tre problemi a risposta aperta, ciascuno dei quali presenta una soluzione intuitiva immediata che risulta però sbagliata. Il punteggio è stato calcolato assegnando:

- 1 punto per ogni risposta corretta
- 0 punti per ogni risposta errata

Il punteggio totale varia quindi da 0 a 3, dove valori più alti indicano una maggiore tendenza al ragionamento riflessivo.
L'inserimento del CRT è coerente con l'obiettivo dello studio di esplorare se una maggiore capacità di riflessione cognitiva sia associata a una migliore capacità di distinguere immagini reali da immagini generate dall'intelligenza artificiale.

3.3.5 Sezione socio-demografica

Al termine del questionario è stata inserita una breve sezione socio-demografica con l'obiettivo di raccogliere alcune informazioni descrittive sui partecipanti e alcune variabili potenzialmente rilevanti per l'analisi dei risultati.
Le domande riguardavano:

1. Età, suddivisa in quattro fasce:
    - 18–24
    - 25–34
    - 35–44
    - 45+
2. Titolo di studio, con le seguenti opzioni:
    - Licenza media
    - Scuola superiore

[2] Per un esempio divulgativo delle domande del Cognitive Reflection Test si veda: https://lamenteemeravigliosa.it/test-di-riflessione-cognitiva-crt-il-test-di-sole-3-domande/

- Laurea triennale
- Laurea magistrale / post-laurea

3. Frequenza di utilizzo dei social media:
    - Più volte al giorno
    - Una volta al giorno
    - Qualche volta a settimana
    - Raramente
4. Familiarità percepita con immagini generate da Intelligenza Artificiale:
    - Molto
    - Abbastanza
    - Poco
    - Per nulla

Questa sezione ha una funzione principalmente descrittiva, ma alcune variabili (in particolare la frequenza di utilizzo dei social media e la familiarità con immagini generate dall'IA) possono essere utilizzate anche come variabili esplorative per verificare eventuali differenze nell'accuratezza di riconoscimento o nell'intenzione di condivisione.

3.4 Procedura

Il questionario è stato somministrato online tramite la piattaforma Google Forms. La partecipazione è avvenuta in forma volontaria e anonima. All'inizio del questionario era presente la seguente informativa:
"Il seguente questionario ha finalità di ricerca accademica. Le risposte sono anonime e verranno analizzate in forma aggregata. Non esistono risposte giuste o sbagliate: ci interessa esclusivamente la tua percezione. Per partecipare è necessario avere almeno 18 anni."
La raccolta dati è avvenuta nel mese di febbraio 2026, diffondendo il questionario attraverso canali online, in particolare gruppi WhatsApp universitari e personali e tramite condivisione sul profilo Instagram della ricercatrice.
Sono state predisposte due versioni del questionario (Versione A e Versione B), distribuite tramite link differenti. Le due versioni differivano per la distribuzione delle immagini con e senza caption, con l'obiettivo di controllare l'effetto del contesto testuale sull'interpretazione visiva. In particolare, le stesse immagini sono state presentate con caption in una versione e senza caption nell'altra, secondo uno schema controbilanciato. Questo approccio consente di

isolare l'effetto della caption mantenendo costanti gli stimoli visivi e riducendo il rischio che eventuali differenze nei risultati siano dovute alle caratteristiche specifiche delle immagini piuttosto che alla presenza del testo. Le due versioni complete del questionario sono riportare in Appendice A (A.1 Questionario A; A.2 Questionario B).

Non era previsto un limite di tempo per la compilazione. Al termine della raccolta dati, le risposte sono state esportate dalla piattaforma Google Forms in formato .csv e successivamente analizzate tramite il linguaggio di programmazione Python, utilizzando l'ambiente Google Colab e librerie statistiche dedicate all'analisi dei dati (in particolare Pandas, NumPy, SciPy e scikit-learn). Il codice completo utilizzato per l'analisi dei dati è disponibile in Appendice C.

3.5 Piano di analisi

3.5.1 Analisi descrittive e inferenziali

In una prima fase sono state effettuate analisi descrittive, al fine di fornire una panoramica generale del campione e delle principali variabili considerate. In particolare, sono state calcolate:

- Frequenze e percentuali per le variabili socio-demografiche;
- Media e deviazione standard per il livello di sicurezza nella risposta, l'intenzione di condivisione, l'indice di media literacy e il punteggio al Cognitive Reflection Test (CRT);
- Percentuale complessiva di risposte corrette nella classificazione delle immagini (accuratezza).

Per analizzare l'accuratezza, è stata creata una variabile dicotomica (0 = risposta errata; 1 = risposta corretta), che ha consentito di calcolare l'accuratezza media per ciascun partecipante e per ciascuna immagine. A ciascuna immagine è stata associata una condizione di verità (immagine reale vs immagine generata da IA) sulla base della struttura delle due versioni del questionario. Confrontando la risposta del partecipante con la condizione corretta dell'immagine, è stata determinata la correttezza della classificazione.

Successivamente, sono state condotte analisi inferenziali per verificare le ipotesi formulate.

In primo luogo, è stata valutata l'accuratezza complessiva nel riconoscimento delle immagini, confrontando la performance media del campione con la soglia del 70%, indicata in letteratura come livello minimo di riconoscimento relativamente affidabile.

Per analizzare l'effetto della caption, sono stati effettuati confronti tra le condizioni sperimentali utilizzando le stesse immagini presentate con e senza caption nelle due versioni del questionario. In particolare, sono stati analizzati:

- Il confronto dell'accuratezza tra immagini presentate con caption e senza caption;
- Il confronto del livello medio di sicurezza dichiarato nelle due condizioni.

Inoltre, è stato effettuato un confronto tra l'accuratezza nel riconoscimento di immagini reali e immagini generate da IA, al fine di verificare eventuali differenze nella difficoltà di classificazione tra le due tipologie di stimolo.

Per quanto riguarda le differenze individuali, sono state analizzate le relazioni tra:

- Indice di media literacy percepita e accuratezza nella classificazione delle immagini;
- Punteggio al CRT e accuratezza;
- Livello medio di sicurezza dichiarato e accuratezza.

Le analisi relative alla propensione alla condivisione sono state condotte esclusivamente sulle immagini accompagnate da caption, poiché solo per queste era presente nel questionario la relativa domanda. In questa sezione sono state analizzate:

- La propensione media alla condivisione nel campione;
- La relazione tra percezione dell'immagine (reale vs generate da IA) e intenzione di condivisione;
- La relazione tra livello di sicurezza percepita (confidence) e intenzione di condivisione;
- Le differenze nella propensione alla condivisione tra immagini percepite come reali e immagini percepite come generate da IA.

Infine, è stata analizzata la distribuzione delle classificazioni dei partecipanti per verificare l'eventuale presenza di un bias sistematico verso la classificazione delle immagini come reali.

A seconda della natura delle variabili considerate, sono stati utilizzati t-test a un campione, t-test per campioni indipendenti o appaiati, test non parametrici (Wilcoxon) e analisi di correlazione Pearson.

L'obiettivo complessivo dell'analisi è verificare le ipotesi formulate e esplorare il ruolo della caption, delle differenze individuali e delle percezioni soggettive nel processo di valutazione e condivisione delle immagini.

3.5.2 Analisi di clustering

Al fine di approfondire ulteriormente i pattern comportamentali e cognitivi emersi dai dati, è stata condotta un'analisi di clustering con l'obiettivo di individuare gruppi omogenei di partecipanti sulla base di specifiche caratteristiche individuali.

In primo luogo, è stato effettuato un clustering basato sul punteggio al Cognitive Reflection Test (CRT), al fine di segmentare i partecipanti in funzione del loro livello di pensiero analitico. A tale scopo, è stato utilizzato l'algoritmo K-means, previa standardizzazione della variabile, e il numero di cluster è stato definito sulla base della struttura dei dati e supportato dall'analisi del metodo del gomito (elbow method). Considerata la natura discreta del punteggio CRT (valori compresi tra 0 e 3), il clustering ha assunto principalmente una funzione di segmentazione per livelli, più che di individuazione di strutture latenti nei dati.

Successivamente, i cluster ottenuti sono stati utilizzati per analizzare eventuali differenze nella performance di riconoscimento delle immagini, con particolare attenzione al confronto tra condizioni con e senza caption. Per ciascun cluster sono state calcolate l'accuratezza media e la relativa variabilità, e sono stati effettuati confronti statistici tra le due condizioni sperimentali mediante t-test per campioni appaiati. Inoltre, al fine di verificare eventuali differenze complessive tra i cluster CRT, è stata condotta un'analisi della varianza (ANOVA) sull'accuratezza media.

In una seconda fase, è stata condotta un'ulteriore analisi di clustering su due variabili chiave: l'accuratezza media nella classificazione delle immagini e il livello medio di sicurezza dichiarato (confidence). Anche in questo caso è stato applicato l'algoritmo K-means, con standardizzazione delle variabili e determinazione del numero ottimale di cluster tramite elbow method.

Questo secondo clustering ha permesso di identificare profili comportamentali differenziati, caratterizzati da diverse combinazioni di performance e sicurezza percepita. In particolare, sono stati individuati gruppi distinti quali partecipanti accurati e sicuri, partecipanti inaccurati ma fiduciosi e partecipanti cauti, consentendo una lettura più articolata del rapporto tra competenza effettiva e percezione soggettiva. Per verificare la significatività delle differenze tra i gruppi individuati, sono state condotte analisi della varianza (ANOVA) sulle variabili utilizzate per il clustering.

Infine, i cluster ottenuti sono stati utilizzati per esplorare la relazione tra caratteristiche individuali e comportamento di condivisione dei contenuti. In particolare, è stata analizzata la propensione media alla condivisione nei diversi cluster e la sua variazione in funzione della

percezione dell'immagine (reale vs generata da IA). Per queste analisi sono stati effettuati confronti statistici tra le condizioni considerate e sono state calcolate, ove opportuno, le dimensioni dell'effetto tramite il coefficiente Cohen's d.

L'analisi di clustering è stata dunque impiegata come strumento esplorativo complementare alle analisi statistiche tradizionali, con l'obiettivo di individuare profili differenziati nei dati e fornire una comprensione più approfondita dei meccanismi cognitivi e comportamentali alla base del riconoscimento e della diffusione delle immagini.

4
RISULTATI

Il presente capitolo presenta i risultati delle analisi statistiche condotte sui dati raccolti tramite il questionario online. Le analisi sono state effettuate utilizzando Python e le librerie Pandas, SciPy e scikit-learn, seguendo il piano di analisi descritto nel capitolo metodologico.

In una prima fase sono state effettuate analisi descrittive al fine di fornire una panoramica generale delle principali variabili considerate, tra cui l'accuratezza nella classificazione delle immagini, il livello di sicurezza dichiarato nelle risposte (confidence), l'indice di media literacy percepita e il punteggio al Cognitive Reflection Test (CRT).

Successivamente sono state condotte analisi inferenziali per verificare le ipotesi di ricerca formulate. In particolare, sono stati analizzati:

- Il livello complessivo di accuratezza nel distinguere immagini reali da immagini generate da intelligenza artificiale;
- Le differenze nel riconoscimento tra immagini reali e immagini generate artificialmente;
- L'effetto della presenza della caption sull'accuratezza e sul livello di sicurezza nelle risposte;
- La relazione tra accuratezza e alcune variabili individuali, tra cui media literacy percepita, punteggio al CRT e livello di sicurezza dichiarato;
- La relazione tra percezione dell'immagine e intenzione di condivisione sui social media;
- L'eventuale presenza di un bias sistematico verso la classificazione delle immagini come reali.

Infine, è stata condotta un'analisi di clustering, finalizzata a individuare gruppi di partecipanti omogenei sulla base del punteggio al CRT e, successivamente, sulla base dell'accuratezza media e del livello medio di confidence, al fine di approfondire i profili cognitivi e comportamentali emersi dai dati.

Le analisi sono state condotte utilizzando t-test per campioni indipendenti o appaiati, test t a un campione, analisi della varianza (ANOVA), test non parametrici (Wilcoxon), correlazioni di Pearson e analisi di clustering K-means, a seconda della natura delle variabili considerate.

4.1 Accuratezza complessiva nel riconoscimento delle immagini

La prima analisi è stata finalizzata a valutare il livello complessivo di accuratezza dei partecipanti nel distinguere tra immagini reali e immagini generate tramite intelligenza artificiale. A questo scopo è stata calcolata, per ciascun partecipante, la percentuale media di risposte corrette nella classificazione delle immagini.

L'accuratezza media del campione risulta pari a 0,52 (52%), con una deviazione standard di 0,13, indicando che i partecipanti hanno identificato correttamente poco più della metà delle immagini presentate.

Per verificare se tale valore fosse significativamente diverso dalla soglia del 70%, indicata in letteratura come livello minimo di riconoscimento relativamente affidabile, è stato condotto un t-test a un campione. I risultati mostrano che l'accuratezza osservata nel campione è significativamente inferiore a tale soglia ($t(201) = -19.09$, $p < .001$).

Questo risultato indica che, nel complesso, i partecipanti hanno mostrato una capacità limitata nel distinguere tra immagini autentiche e immagini generate artificialmente, con una performance solo leggermente superiore al livello casuale.

Al fine di verificare eventuali differenze legate alla versione del questionario compilata, è stato inoltre effettuato un confronto tra l'accuratezza media dei partecipanti che hanno completato la versione A e quelli che hanno completato la versione B. Il confronto tramite t-test per campioni indipendenti non ha evidenziato differenze statisticamente significative tra le due condizioni ($t(200) = -1.11$, $p = .267$), suggerendo che le due versioni del questionario risultano sostanzialmente equivalenti in termini di difficoltà complessiva del compito (fig. 4.1).

Il risultato conferma la comparabilità delle due versioni sperimentali e consente di considerare il campione complessivo nelle successive analisi.

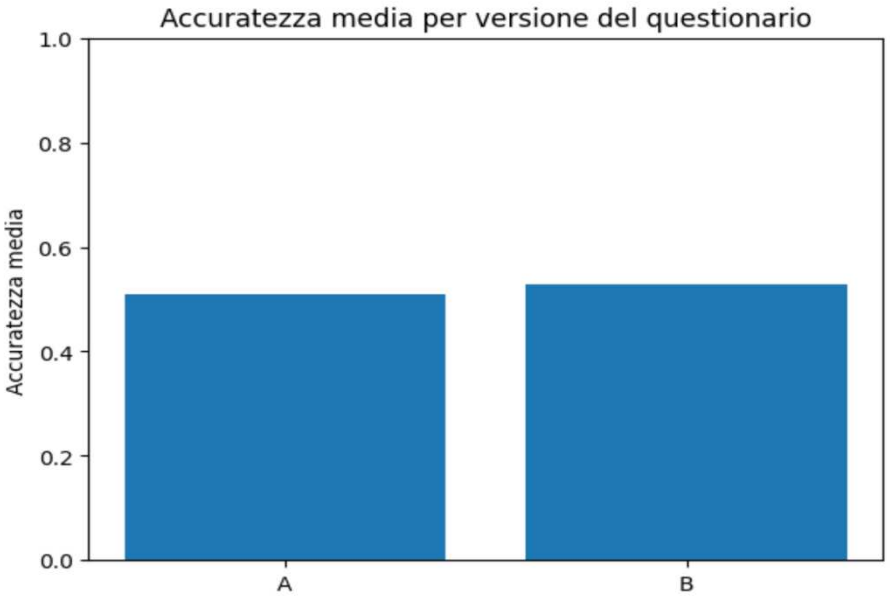


Figura 4.1 – Accuratezza media nella classificazione delle immagini per versione del questionario (A e B).

4.2 Differenza nel riconoscimento tra immagini reali e immagini generate da IA

Al fine di approfondire la capacità dei partecipanti di distinguere tra contenuti autentici e contenuti generati artificialmente, è stata analizzata separatamente l'accuratezza nella classificazione delle immagini reali e delle immagini generate tramite intelligenza artificiale.
Per ciascun partecipante è stata quindi calcolata la percentuale di risposte corrette nelle due categorie di immagini. I risultati mostrano che le immagini reali sono state riconosciute correttamente con una frequenza maggiore rispetto alle immagini generate da IA. In particolare, l'accuratezza media nel riconoscimento delle immagini reali risulta pari a 0,56 (56%), mentre per le immagini generate artificialmente l'accuratezza media è pari a 0,48 (48%) (fig. 4.2).
Per verificare se questa differenza fosse statisticamente significativa è stato condotto un t-test per campioni appaiati. I risultati indicano una differenza significativa tra le due condizioni ($t(201) = 3.56$, $p < .001$), suggerendo che i partecipanti hanno riscontrato maggiori difficoltà nell'identificare correttamente le immagini generate da intelligenza artificiale rispetto alle immagini reali. L'analisi della dimensione dell'effetto mostra un Cohen's d pari a 0,25, indicando un effetto di entità piccola ma significativa.
A livello descrittivo, l'analisi per singola immagine ha inoltre evidenziato una certa variabilità nelle percentuali di riconoscimento. In particolare, le immagini A11 e A8 sono risultate tra le

più difficili da classificare correttamente, mentre A10, B1 e B3 hanno mostrato livelli di accuratezza più elevati.

Sostanzialmente, questi risultati evidenziano come le immagini generate artificialmente tendano ad essere più facilmente scambiate per contenuti autentici, confermando la difficoltà dei partecipanti nel riconoscere correttamente i contenuti visivi sintetici.

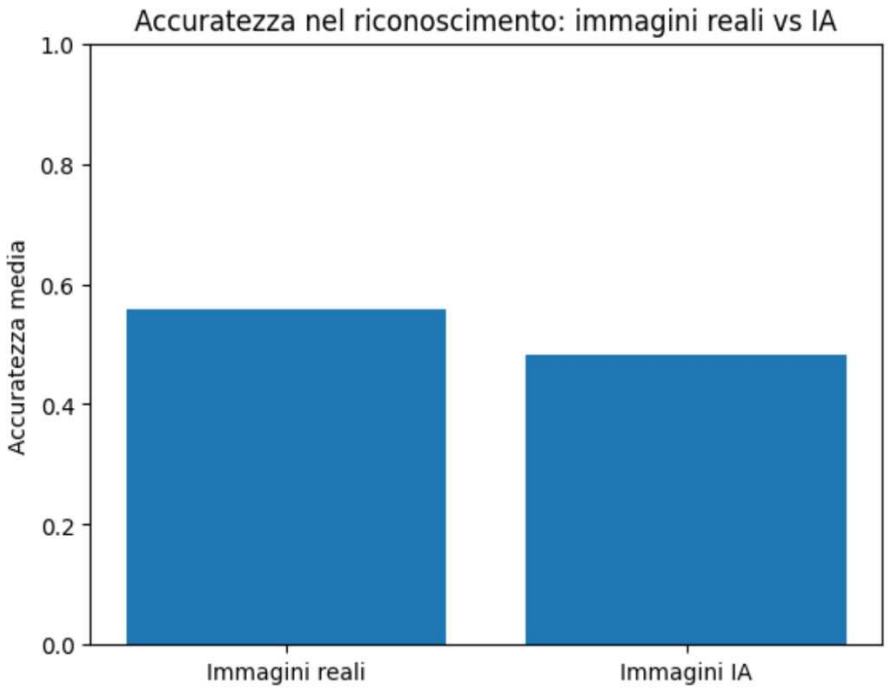


Figura 4.2 – Accuratezza media nel riconoscimento delle immagini reali e generate da intelligenza artificiale.

4.3 Effetto della caption

Uno degli obiettivi principali dello studio era verificare se la presenza di una caption testuale associata all'immagine potesse influenzare la capacità dei partecipanti di distinguere tra immagini reali e immagini IA.

Per analizzare questo aspetto sono state confrontate le performance di riconoscimento delle immagini presentate con caption e presentate senza caption. L'accuratezza è stata calcolata confrontando, per ciascuna immagine, la percentuale di classificazioni corrette nelle due condizioni sperimentali.

La figura 4.3.1 mostra l'accuratezza di classificazione per ciascuna immagine nelle condizioni con e senza caption. Come si può osservare, le differenze tra le due condizioni risultano generalmente contenute e non mostrano un andamento sistematico.

Per verificare se la presenza della caption producesse un effetto significativo sull'accuratezza è stato condotto un t-test per campioni appaiati, che non ha evidenziato differenze statisticamente significative tra le due condizioni ($t(11) = -0.23$, $p = .821$).

In termini descrittivi, l'accuratezza media risulta pari a 0,515 per le immagini con caption e a 0,523 per quelle senza caption.

Questo risultato indica che, nel contesto dell'esperimento, la presenza di un testo associato all'immagine non ha influenzato in modo significativo la capacità dei partecipanti di distinguere tra immagini reali e immagini generate artificialmente.

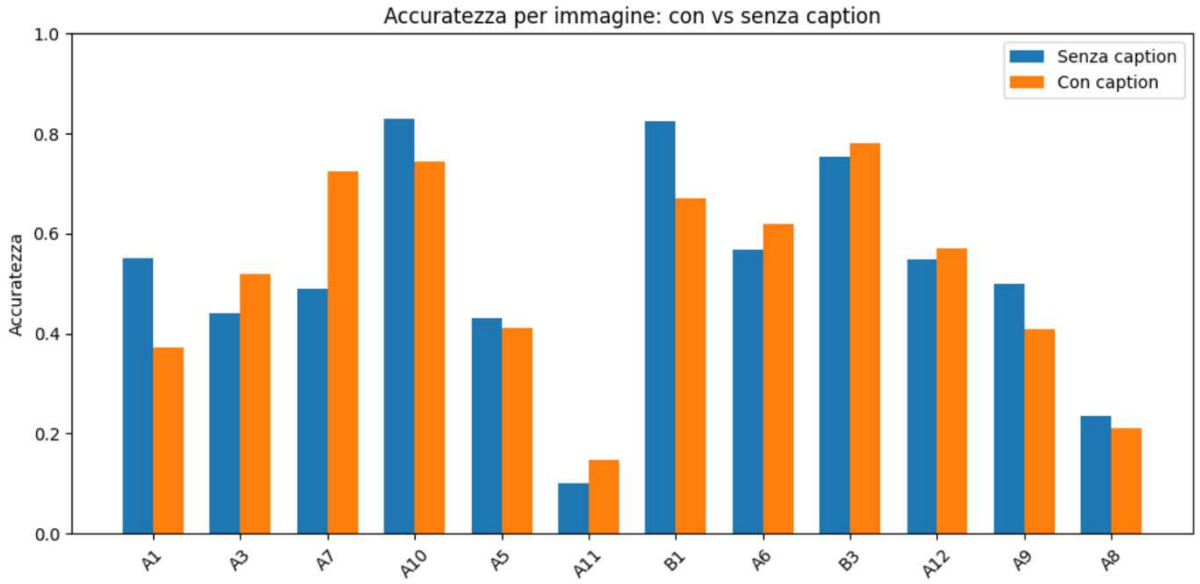


Figura 4.3.1 – Accuratezza di classificazione per ciascuna immagine nelle condizioni con caption e senza caption.

Inoltre, è stato analizzato il possibile effetto della caption sul livello di sicurezza dichiarato nelle risposte (confidence). Anche in questo caso è stato effettuato un confronto tra le condizioni con e senza caption.

La figura 4.3.2 mostra il livello medio di confidence associato a ciascuna immagine nelle due condizioni sperimentali. Analogamente a quanto osservato per l'accuratezza, le differenze tra le due condizioni risultano limitate e non evidenziano un pattern sistematico.

L'analisi statistica non ha evidenziato differenze significative nel livello di sicurezza tra le immagini presentate con caption e quelle prive di testo ($p = .424$). In termini descrittivi, il livello medio di confidence è pari a 4,83 e 4,90 rispettivamente.

In generale, questi risultati indicano che la presenza di un testo descrittivo associato all'immagine non sembra aver avuto un impatto rilevante né sull'accuratezza delle classificazioni né sul livello di sicurezza dichiarato dai partecipanti.

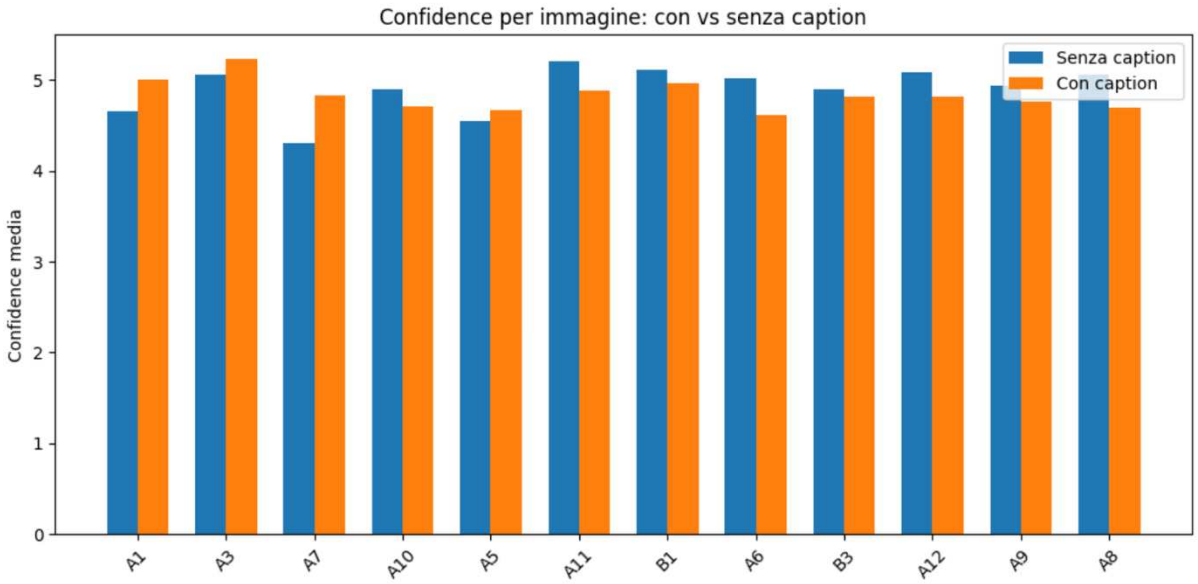


Figura 4.3.2 – Livello medio di confidence associato a ciascuna immagine nelle condizioni con caption e senza caption.

4.4 Relazione tra variabili individuali e accuratezza

Un ulteriore obiettivo dello studio era verificare se alcune caratteristiche individuali dei partecipanti fossero associate a una maggiore capacità di detection. In particolare, sono state analizzate tre variabili: media literacy percepita, pensiero analitico – misurato tramite il Cognitive Reflection Test (CRT) – e livello di sicurezza dichiarato nelle risposte (confidence).

4.4.1 Media literacy percepita

Per valutare la relazione tra competenze percepite e performance effettiva nel compito di riconoscimento delle immagini, è stata analizzata la correlazione tra il punteggio medio di media literacy percepita e l'accuratezza media nella classificazione delle immagini.

L'analisi di correlazione di Pearson non ha evidenziato una relazione significativa tra le due variabili ($r = .03$, $p = .622$). Questo risultato indica che i partecipanti che dichiarano livelli più elevati di competenze di media literacy non mostrano necessariamente una maggiore capacità di distinguere tra immagini reali e immagini generate artificialmente.

4.4.2 Pensiero analitico (CRT)

Inoltre, è stata studiata la possibile relazione tra pensiero analitico, misurato attraverso il punteggio al Cognitive Reflection Test, e accuratezza nella classificazione delle immagini.

Anche in questo caso l'analisi di correlazione di Pearson non ha evidenziato una relazione statisticamente significativa tra le due variabili ($r = .09$, $p = .197$). Il punteggio ottenuto al CRT non risulta quindi associato in modo rilevante alla capacità dei partecipanti di riconoscere correttamente le immagini generate da intelligenza artificiale.

4.4.3 Livello di sicurezza nelle risposte (confidence)

Infine, è stata analizzata la relazione tra livello di sicurezza dichiarato nelle risposte (confidence) e accuratezza nella classificazione delle immagini.

L'analisi di correlazione di Pearson non ha evidenziato una relazione statisticamente significativa tra le due variabili ($r = .08$, $p = .270$), indicando che i partecipanti che dichiarano livelli di sicurezza più elevati non risultano necessariamente più accurati nel riconoscimento delle immagini.

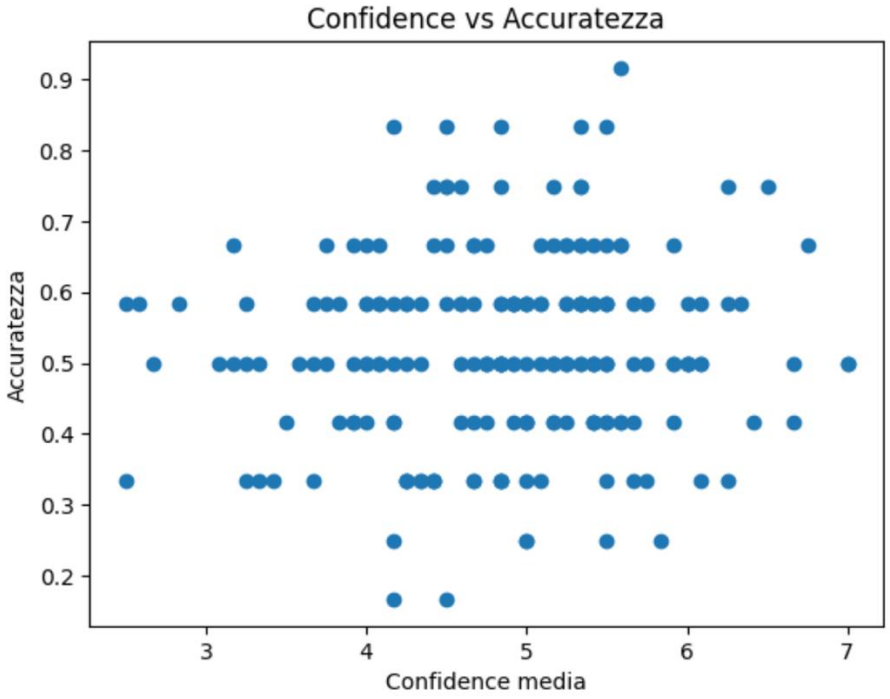


Figura 4.4 – Relazione tra livello medio di confidence e accuratezza nella classificazione delle immagini.

Complessivamente, questi risultati suggeriscono che media literacy percepita, pensiero analitico e livello di sicurezza dichiarato non risultano associati in modo statisticamente significativo alla capacità di distinguere tra immagini reali e immagini IA.

4.5 Relazione tra percezione dell'immagine e intenzione di condivisione

Un altro obiettivo dello studio era analizzare la propensione alla condivisione delle immagini sui social media e la sua relazione con diverse variabili percettive e cognitive.

In primo luogo, è stata analizzata la propensione media alla condivisione nel campione. I risultati mostrano un valore medio pari a 2,09 su una scala da 1 a 7, indicando una generale bassa probabilità dichiarata di condividere i contenuti presentati.

A livello descrittivo, l'analisi per singola immagine evidenzia valori medi di intenzione di condivisione generalmente bassi, compresi tra 1,78 e 2,75 su una scala da 1 a 7. In particolare, le immagini A8 (M = 2,75) e B1 (M = 2,37) risultano tra le più condivisibili, mentre A12 (M = 1,78) e B3 (M = 1,86) presentano i livelli più bassi di propensione alla condivisione.

Successivamente, è stata esaminata la relazione tra accuratezza di riconoscimento e intenzione di condivisione. L'analisi di correlazione non ha evidenziato una relazione statisticamente significativa tra le due variabili ($r = -.07$, $p = .308$), suggerendo che la capacità di distinguere correttamente tra immagini reali e immagini generate da IA non è associata in modo rilevante alla propensione alla condivisione.

Inoltre, è stata analizzata la relazione tra la percezione dell'autenticità delle immagini e la probabilità dichiarata di condivisione. In particolare, è stato esaminato se le immagini percepite come reali tendessero ad essere considerate più condivisibili rispetto a quelle percepite come generate da intelligenza artificiale.

Per questo scopo è stata confrontata la probabilità media di condivisione associata alle immagini classificate come reali (M = 2,44) e a quelle classificate come generate da IA (M = 1,71). Il confronto è stato effettuato tramite t-test per campioni indipendenti.

I risultati mostrano una differenza significativa tra le due condizioni: le immagini percepite come reali presentano una probabilità di condivisione significativamente più elevata rispetto alle immagini percepite come generate artificialmente ($t = 8.56$, $p < .001$) (fig. 4.5). Questo risultato suggerisce che la percezione di autenticità dell'immagine è associata a una maggiore propensione alla condivisione.

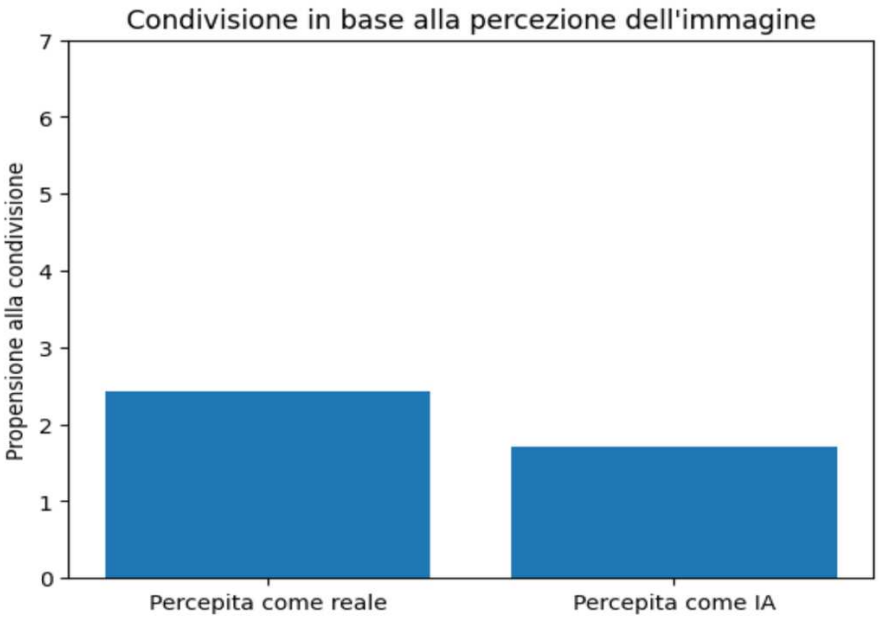


Figura 4.5 – Probabilità media di condivisione delle immagini in base alla percezione di autenticità (immagini percepite come reali vs generate da IA).

In aggiunta, è stata analizzata la relazione tra livello di sicurezza dichiarato nelle risposte (confidence) e probabilità di condivisione. L'analisi di correlazione ha evidenziato una relazione positiva tra le due variabili ($r = .247$, $p < .001$), indicando che i partecipanti che esprimono maggiore sicurezza nelle proprie valutazioni tendono anche a dichiarare una maggiore probabilità di condividere le immagini sui social media.

Dunque, i risultati indicano che sia la percezione di autenticità dell'immagine sia il livello di sicurezza nelle proprie valutazioni risultano associati a una maggiore propensione alla condivisione dei contenuti.

4.6 Bias di classificazione

Un'ulteriore analisi è stata condotta con lo scopo di verificare l'eventuale presenza di un bias sistematico nelle classificazioni delle immagini, ovvero una tendenza dei partecipanti a preferire una delle due categorie di risposta ("reale" o "generata da IA").

A questo proposito, è stata analizzata la distribuzione complessiva delle classificazioni effettuate dai partecipanti, confrontando la frequenza con cui le immagini venivano etichettate come reali rispetto alla frequenza con cui venivano classificate come generate artificialmente.

I risultati indicano una tendenza significativa a classificare le immagini come reali. Il confronto statistico ha evidenziato che le immagini vengono identificate come reali con una frequenza maggiore rispetto a quanto sarebbe atteso in assenza di bias ($t(201) = -4.68$, $p < .001$) (fig. 4.6).

Ciò suggerisce la presenza di una propensione generale a considerare autentici i contenuti visivi, anche quando questi possono essere generati artificialmente.

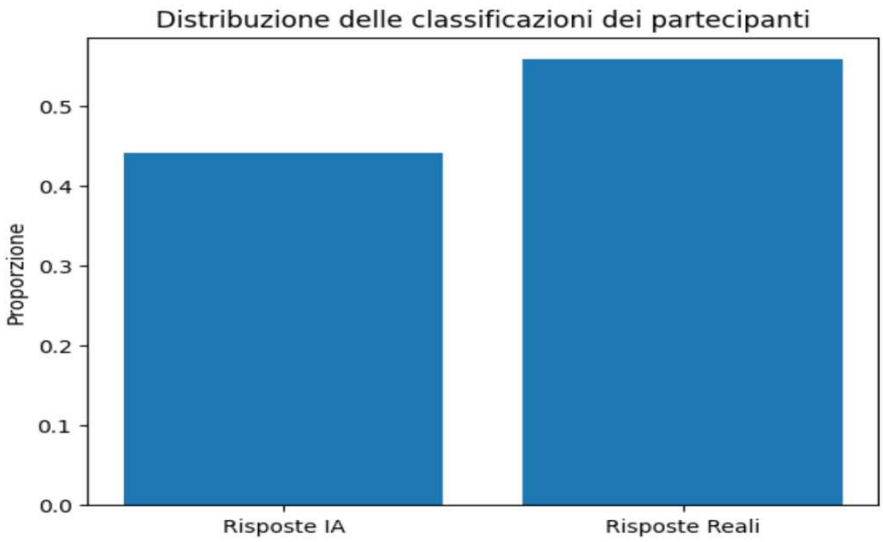


Figura 4.6 – Distribuzione complessiva delle classificazioni delle immagini (reali vs generate da IA).

Nell'insieme, i risultati evidenziano una difficoltà generale nel distinguere tra immagini reali e immagini generate da intelligenza artificiale, assieme alla presenza di alcuni fattori associati alla valutazione e alla condivisione dei contenuti visivi. Nel capitolo successivo tali risultati verranno discussi alla luce della letteratura esistente.

4.7 Analisi di clustering

4.7.1 Clustering basato sul punteggio al Cognitive Reflection Test (CRT)

Al fine di approfondire il ruolo del pensiero analitico nelle performance di riconoscimento, è stata condotta un'analisi di clustering basata sul punteggio ottenuto al Cognitive Reflection Test (CRT).

L'algoritmo K-means ha permesso di suddividere i partecipanti in quattro gruppi distinti, corrispondenti ai diversi livelli di punteggio CRT (da 0 a 3). Data la natura discreta della variabile, il clustering coincide di fatto con una segmentazione per livelli di riflessività cognitiva.

La distribuzione dei partecipanti nei cluster evidenzia una prevalenza di individui con punteggi elevati di pensiero analitico. In particolare, il gruppo con il punteggio massimo (CRT = 3) risulta il più numeroso (n = 89), seguito dai partecipanti con punteggio 2 (n = 43), 0 (n=39) e 1 (n = 31).

Per facilitare l'interpretazione, i cluster sono stati denominati come segue:

- *Gli Istintivi Puri* (CRT = 0);
- *I Riflessivi in Erba* (CRT = 1);
- *Le Menti Acute* (CRT = 2);
- *I Maestri Analitici* (CRT = 3).

La figura 4.7.1 (a) mostra la distribuzione dei partecipanti nei diversi cluster.

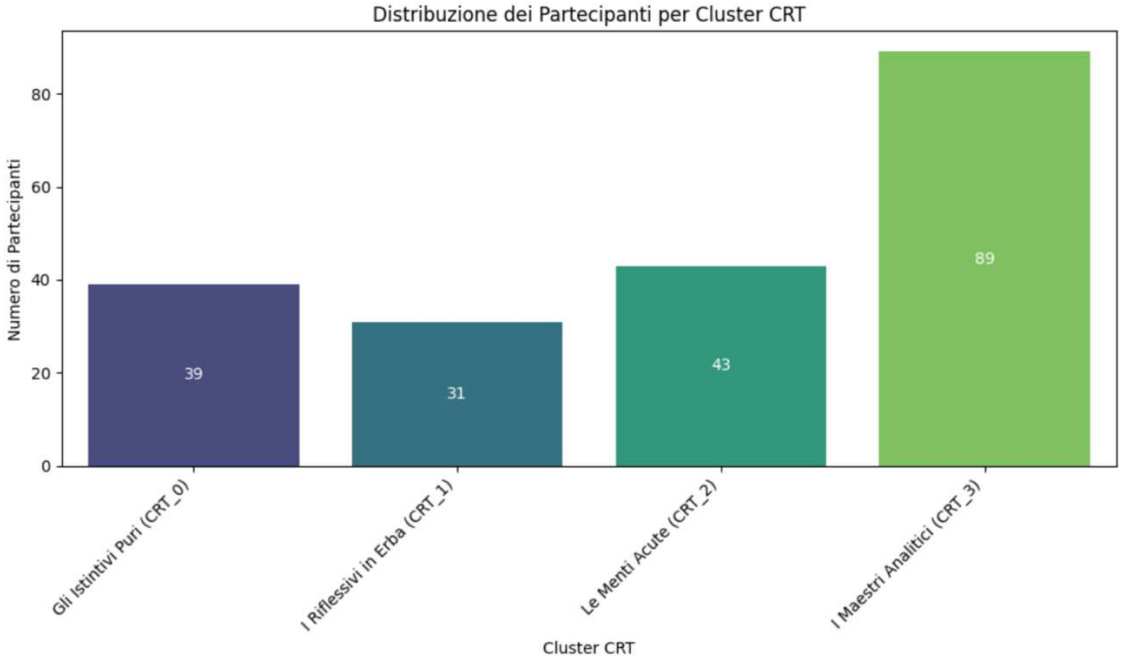


Figura 4.7.1 (a) – Distribuzione dei partecipanti nei cluster basati sul punteggio CRT.

Questa segmentazione consente di distinguere i partecipanti in base al loro livello di pensiero analitico, creando una base per le analisi successive volte a verificare se tali differenze individuali siano associate a variazioni nella capacità di riconoscimento delle immagini generate da intelligenza artificiale.

Inoltre, è stata analizzata la distribuzione dell'accuratezza all'interno dei diversi cluster CRT. Come mostrato in figura 4.7.1 (b), i valori medi (rappresentati dalle linee rosse tratteggiate) risultano molto simili tra i gruppi, con un'accuratezza pari a circa 0,50 per i partecipanti con punteggi CRT più bassi e pari a circa 0,53 per i gruppi con livelli più elevati.

Anche all'interno di ogni gruppo CRT, è presente una notevole variabilità nei punteggi di accuratezza. In altre parole, nonostante il raggruppamento per livello di pensiero analitico, non tutti i membri di un cluster si comportano allo stesso modo in termini di accuratezza. Ad esempio, nel gruppo dei *Maestri Analitici*, pur avendo la media più alta, si osservano comunque partecipanti con accuratezza inferiore, e viceversa.

Inoltre, sebbene la correlazione generale tra CRT e accuratezza non fosse significativa, osservando le medie evidenziate nel grafico, emerge una lieve differenza tra i gruppi:

- Gli *Istintivi Puri* (CRT_0) sembrano avere una distribuzione più ampia, con la media di accuratezza più bassa rispetto agli altri gruppi;
- I *Maestri Analitici* (CRT_3) mostrano una media di accuratezza leggermente più alta rispetto agli altri cluster, e la distribuzione sembra essere meno inclinata verso i valori bassi di accuratezza rispetto ai gruppi CRT più bassi. Tuttavia, è importante sottolineare che la differenza nelle medie tra i gruppi non è estremamente marcata e, come visto in precedenza, la correlazione complessiva non era significativa.

Il test ANOVA conferma che, nonostante alcune leggere differenze nelle medie visibili nei grafici, non c'è una differenza statisticamente significativa nell'accuratezza media tra i cluster CRT (F-statistic: 0.63, p = 0.595).

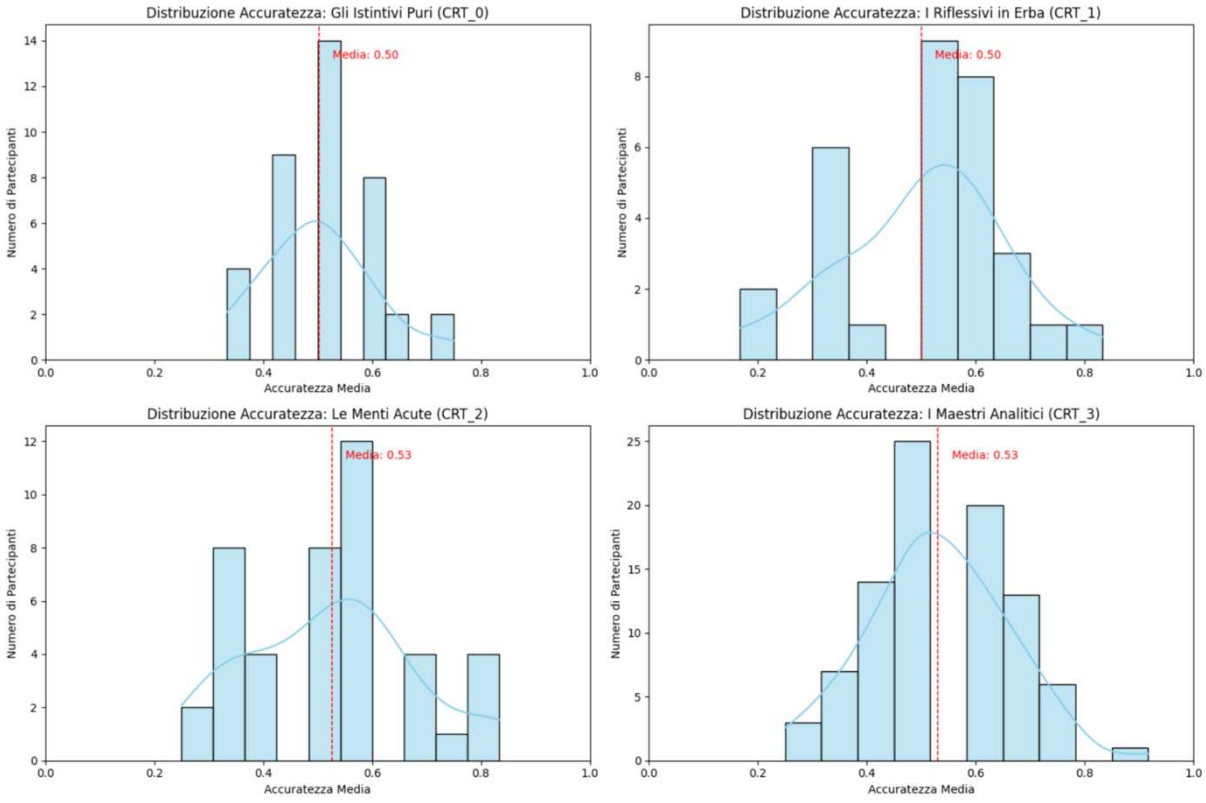


Figura 4.7.1 (b) – Distribuzione dell'accuratezza nei diversi cluster CRT, con indicazione della media per ciascun gruppo.

4.7.2 Accuratezza e caption nei cluster CRT

Per verificare se il livello di pensiero analitico potesse influenzare l'effetto della caption sulla capacità di riconoscimento, è stata analizzata l'accuratezza media nelle condizioni con e senza caption all'interno di ciascun cluster CRT.

Per ogni gruppo sono state calcolate l'accuratezza media e la variabilità nelle due condizioni sperimentali, e sono stati effettuati confronti statistici mediante t-test per campioni appaiati.

I risultati mostrano che, in tutti i cluster, le differenze tra le due condizioni sono contenute. In particolare, le medie di accuratezza con caption e senza caption sono molto simili per ciascun gruppo (fig. 4.7.2). Questa osservazione è in linea con i risultati dei test statistici, che non hanno mostrato differenze significative tra l'accuratezza con e senza caption in nessuno dei cluster:

- Nel cluster degli *Istintivi Puri* (CRT = 0), l'accuratezza media è pari a 0,52 con caption e 0,49 senza caption; il confronto non risulta statisticamente significativo (t = 0.51, p = .618);
- *I Riflessivi in Erba* (CRT = 1) mostrano un'accuratezza media pari a 0,50 in entrambe le condizioni; anche in questo caso non emerge una differenza significativa (t = -0.06, p = .955);
- Nel gruppo delle *Menti Acute* (CRT = 2), l'accuratezza media è pari a 0,51 con caption e 0,54 senza caption; la differenza non risulta significativa (t = -0.87, p = .404);
- Nel cluster dei *Maestri Analitici* (CRT = 3), i valori risultano rispettivamente pari a 0,52 e 0,54; anche in questo caso il confronto non è statisticamente significativo (t = -0.36, p = .728).

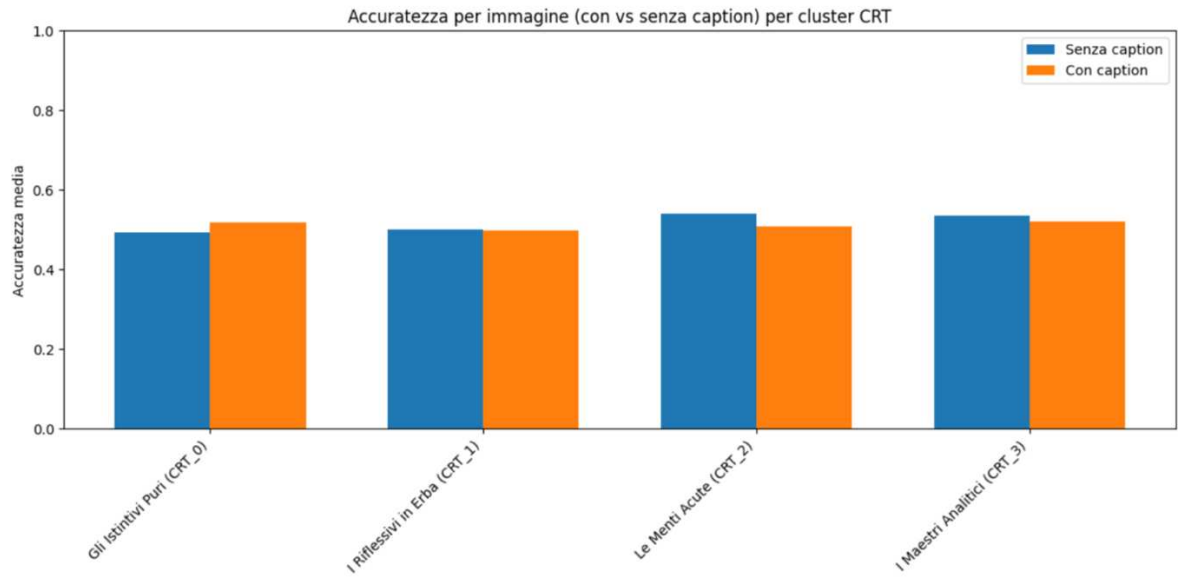


Figura 4.7.2 – Accuratezza media nelle condizioni con e senza caption nei diversi cluster CRT.

4.7.3 Clustering su accuratezza e livello di sicurezza (confidence)

Al fine di approfondire la relazione tra percezione soggettiva e performance effettiva, è stata condotta un'ulteriore analisi di clustering basata su due variabili chiave: l'accuratezza media nella classificazione delle immagini e il livello medio di sicurezza dichiarato nelle risposte (confidence).

L'algoritmo K-means ha permesso di individuare quattro gruppi distinti di partecipanti, caratterizzati da diverse combinazioni di accuratezza e sicurezza percepita. I cluster ottenuti evidenziano profili differenziati, che permettono di distinguere tra performance e sicurezza dichiarata.

In particolare, sono stati identificati i seguenti gruppi, visibili nella figura 4.7.3:

- *I Fiduciosi ma Inaccurati*: partecipanti caratterizzati da un basso livello di accuratezza (M = 0.35) e da un'elevata sicurezza nelle proprie risposte (M = 5.01);
- *I Mediocri ma Cauti*: partecipanti con livelli di accuratezza simili al gruppo precedente (M = 0.52), ma con un livello di sicurezza più basso (M = 3.71);
- *I Mediocri ma Molto Sicuri*: partecipanti con livelli di accuratezza intermedi (M = 0.53), ma con un'elevata confidence (M = 5.41);
- *I Competenti e Sicuri*: partecipanti che presentano sia un'elevata accuratezza (M = 0.73) sia un'elevata sicurezza nelle proprie valutazioni (M = 5.12).

Per verificare la presenza di differenze statisticamente significative tra i cluster, sono state condotte analisi della varianza (ANOVA) sulle due variabili considerate.

I risultati mostrano differenze significative sia per l'accuratezza ($F = 182.87$, $p < .001$) sia per il livello di confidence ($F = 97.44$, $p < .001$), indicando che le medie dei gruppi differiscono in modo significativo tra i cluster individuati.

In generale, questi risultati confermano che i cluster identificati rappresentano gruppi distinti di partecipanti in termini di accuratezza nella classificazione delle immagini e nella sicurezza dichiarata.

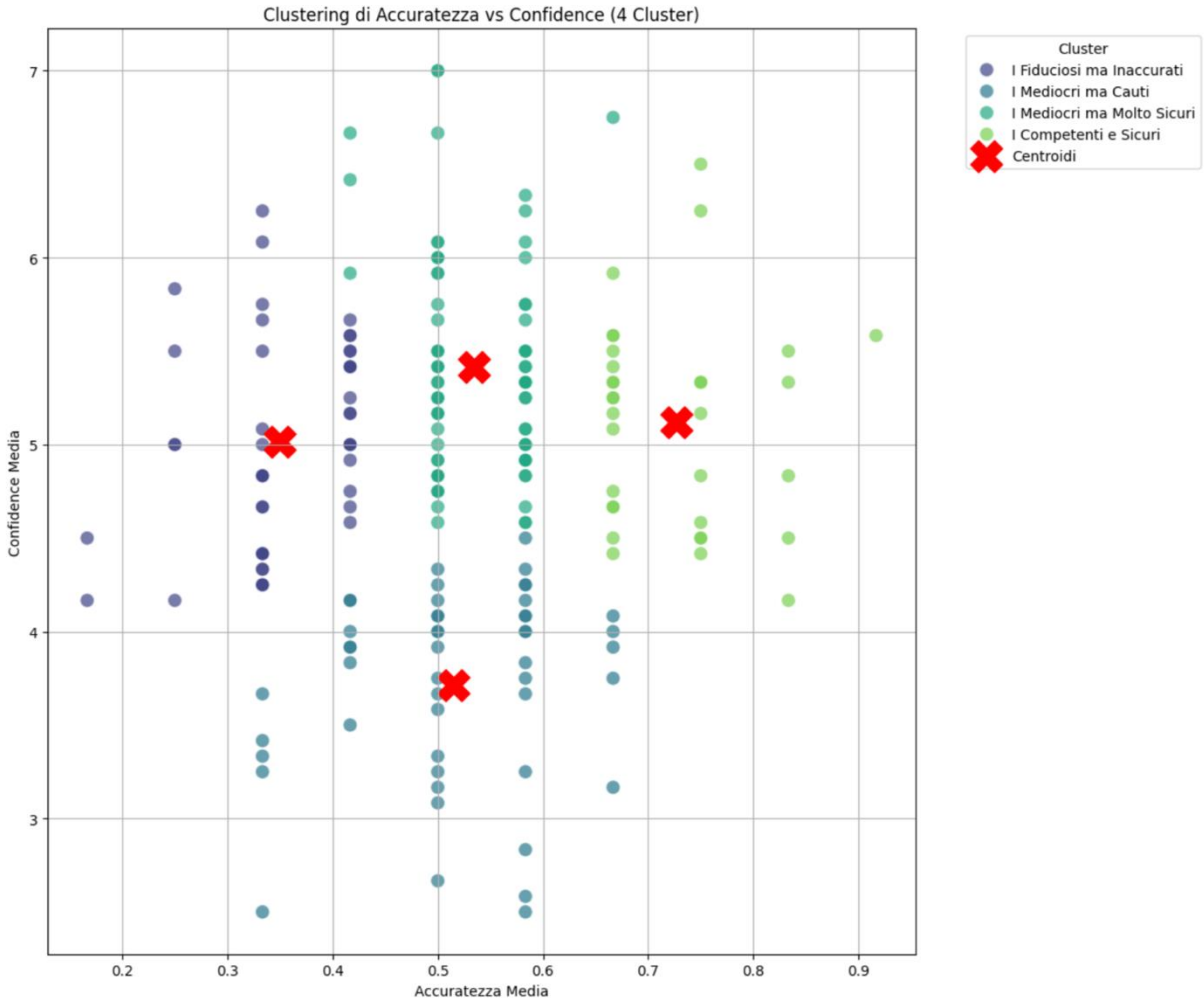


Figura 4.7.3 – Distribuzione dei partecipanti nei cluster individuati tramite analisi K-means sulla base dell'accuratezza e del livello di sicurezza (confidence).

4.7.4 Propensione alla condivisione nei cluster

Al fine di approfondire le differenze comportamentali tra i gruppi individuati tramite l'analisi di clustering, è stata analizzata la propensione alla condivisione delle immagini nei diversi cluster.

In primo luogo, è stata esaminata la probabilità media di condivisione per ciascun cluster, indipendentemente dalla percezione dell'immagine. I risultati evidenziano differenze tra i gruppi: il cluster dei *Mediocri ma Molto Sicuri* presenta il valore medio più elevato (M = 2.35), seguito dal cluster dei *Fiduciosi ma Inaccurati* (M = 2.13). Valori inferiori si osservano nel cluster dei *Mediocri ma Cauti* (M = 1.84) e nel cluster dei *Competenti e Sicuri* (M = 1.78) (fig. 4.7.4 a).

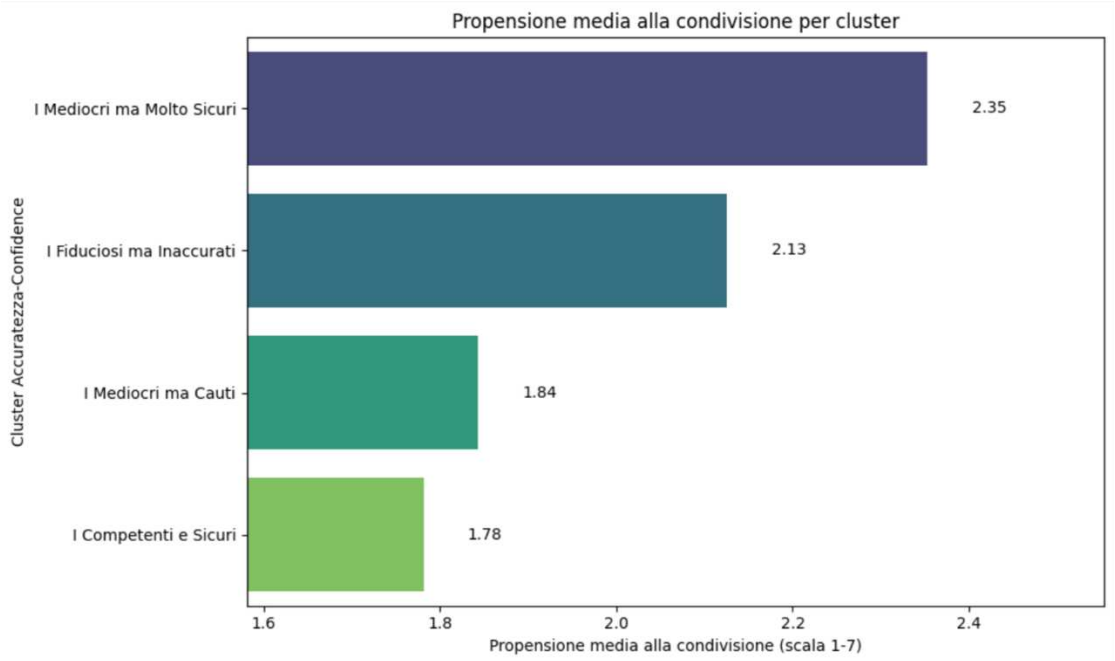


Figura 4.7.4 (a) – Propensione media alla condivisione per ciascun cluster.

Successivamente, è stata analizzata la propensione alla condivisione all'interno di ciascun cluster in funzione della percezione dell'immagine, distinguendo tra immagini classificate come reali e immagini classificate come generate da intelligenza artificiale. Per valutare la rilevanza delle differenze osservate, sono stati effettuati test per campioni appaiati e sono state calcolate le dimensioni dell'effetto tramite il coefficiente di Cohen's d.

I risultati mostrano che, in tutti i cluster, la probabilità di condivisione è più elevata per le immagini percepite come reali rispetto a quelle percepite come generate da IA. Tuttavia, tale differenza risulta statisticamente significativa solo in alcuni gruppi.

In particolare:

- Nel cluster dei *Fiduciosi ma Inaccurati,* la probabilità media di condivisione è pari a 2.50 per le immagini percepite come reali e a 1.78 per quelle percepite come IA;
- Nel cluster dei *Mediocri ma Cauti*, i valori risultano pari a 1.92 per le immagini percepite come reali e 1.76 per quelle percepite come IA;
- Nel cluster dei *Mediocri ma Molto Sicuri*, i valori sono pari a 2.84 per le immagini percepite come reali e 1.77 per quelle percepite come IA;
- Nel cluster dei *Competenti e Sicuri*, la probabilità media di condivisione è pari a 2.17 per le immagini percepite come reali e 1.36 per quelle percepite come IA (fig. 4.7.4 b).

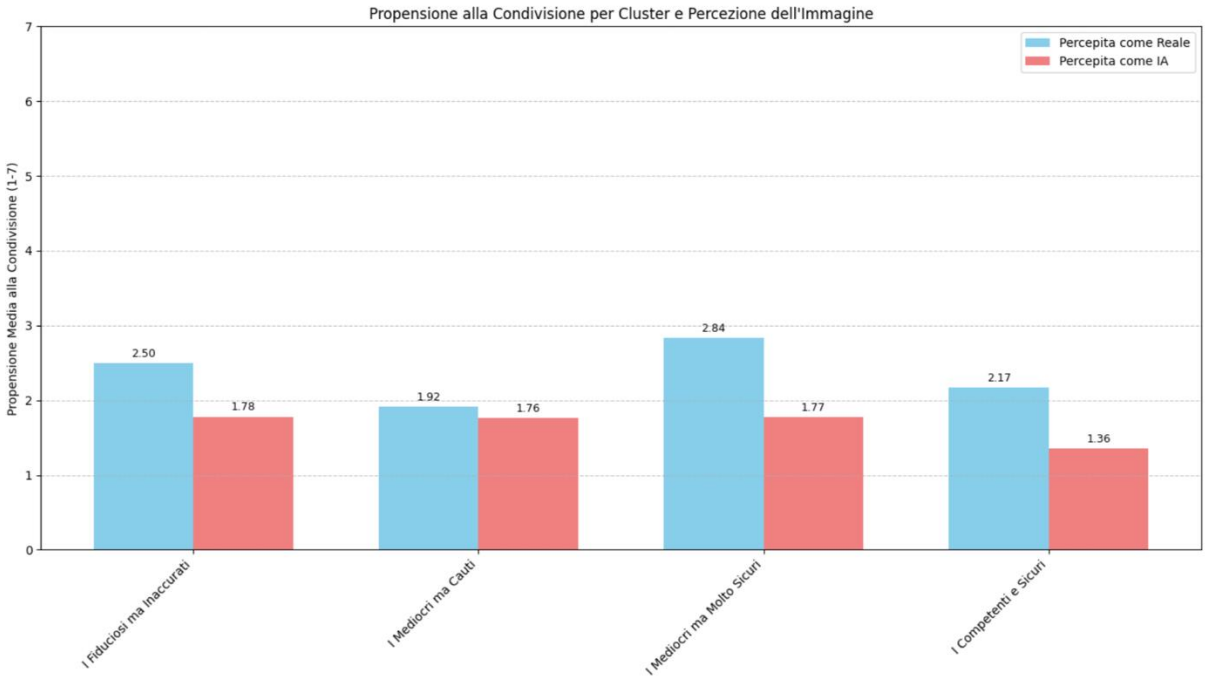


Figura 4.7.4 (b) – Propensione alla condivisione per cluster in base alla percezione dell'immagine (reale vs IA).

Per valutare la rilevanza statistica delle differenze osservate tra le due condizioni (immagini percepite come reali vs IA), sono state calcolate le dimensioni dell'effetto tramite il coefficiente Cohen's d e i relativi valori di significatività.

I risultati mostrano che:

- Il cluster dei *Fiduciosi ma Inaccurati* mostra una differenza significativa ($p < .001$) con un effetto medio ($d = 0.501$);
- Il cluster dei *Mediocri ma Cauti* non presenta una differenza statisticamente significativa ($p = .318$), con un effetto di entità molto ridotta ($d = 0.116$);
- Il cluster dei *Mediocri ma Molto Sicuri* evidenzia una differenza significativa ($p < .001$) con un effetto medio-grande ($d = 0.649$);
- Il cluster dei *Competenti e Sicuri* presenta una differenza significativa ($p < .001$) con un effetto di entità medio-grande ($d = 0.662$).

5
DISCUSSIONE

5.1 Sintesi dei risultati principali

Il presente studio aveva l'obiettivo di analizzare la capacità degli utenti di distinguere tra immagini reali e immagini generate da intelligenza artificiale, nonché il ruolo della caption testuale e di alcune differenze individuali nel processo di valutazione e condivisione dei contenuti visivi.

I risultati mostrano, in primo luogo, una generale difficoltà nel riconoscimento delle immagini, con un livello di accuratezza complessiva solo leggermente superiore al caso, confermando la prima ipotesi:

H1. *L'accuratezza nel distinguere immagini reali da immagini generate dall'IA non sarà significativamente superiore al 70%, in linea con quanto emerso dalla letteratura.*

Questo dato suggerisce che, nel contesto simulato dei social media, i partecipanti presentano una capacità limitata di distinguere in modo affidabile tra contenuti autentici e contenuti generati artificialmente.

Analizzando le diverse tipologie di immagini, emerge che quelle generate da intelligenza artificiale risultano più difficili da identificare correttamente rispetto alle immagini reali, indicando una maggiore capacità delle immagini sintetiche di apparire plausibili e credibili agli occhi degli utenti.

Per quanto riguarda il ruolo della caption, i risultati non evidenziano un effetto significativo né sull'accuratezza delle classificazioni né sul livello di sicurezza dichiarato nelle risposte, smentendo le ipotesi H2 e H3, secondo cui:

H2. *La presenza di una caption influenzerà l'accuratezza della classificazione rispetto alla condizione senza caption.*

H3. *La presenza di una caption influenzerà il livello di sicurezza dichiarato nella risposta.*

Ciò suggerisce che, nel contesto sperimentale considerato, la presenza di un testo associato all'immagine non influenza in modo rilevante il processo di valutazione visiva. Le analisi di clustering basate sul punteggio CRT rafforzano questo risultato, mostrando che l'assenza di un effetto della caption si mantiene stabile anche nei diversi gruppi di pensiero analitico.

Per quanto riguarda il comportamento di condivisione, emerge una generale bassa propensione a condividere i contenuti. Tuttavia, la probabilità di condivisione risulta significativamente

maggiore per le immagini percepite come reali rispetto a quelle percepite come generate da IA, confermando la quarta ipotesi:

H4. *Le immagini percepite come reali saranno associate a una maggiore intenzione di condivisione rispetto a quelle percepite come generate da IA.*

Ciò suggerisce che la percezione di autenticità giochi un ruolo centrale nelle dinamiche di diffusione dei contenuti.

Le analisi relative alle differenze individuali mostrano che né il livello di media literacy percepita né il punteggio al Cognitive Reflection Test risultano connessi in modo significativo alla capacità di riconoscimento, smentendo le ipotesi H5 e H6:

H5. *Un livello più alto di media literacy sarà associato a una maggiore accuratezza nel riconoscimento delle immagini generate artificialmente.*

H6. *Un punteggio più alto al Cognitive Reflection Test (CRT) sarà associato a una maggiore accuratezza nel distinguere immagini reali e sintetiche.*

In modo coerente, le analisi di clustering basate sul CRT non evidenziano differenze significative nell'accuratezza media tra i gruppi individuati.

Allo stesso modo, il livello di sicurezza dichiarato nelle risposte non risulta correlato in maniera significativa all'accuratezza, evidenziando una possibile discrepanza tra percezione soggettiva e performance reale. Coerentemente, le analisi di clustering evidenziano questa discrepanza in alcuni gruppi di partecipanti.

Per quanto riguarda i quesiti di ricerca esplorativi, il livello di sicurezza nelle proprie valutazioni risulta positivamente associato alla propensione alla condivisione: i partecipanti che si dichiarano più sicuri tendono anche a condividere di più. Le analisi di clustering mostrano che i gruppi caratterizzati da livelli più elevati di sicurezza soggettiva presentano anche, in media, una maggiore propensione alla condivisione. I risultati permettono dunque di rispondere positivamente al primo quesito:

RQ1. *La propensione alla condivisione è associata al livello di sicurezza percepita nella classificazione delle immagini?*

Infine, i risultati evidenziano la presenza di un bias sistematico nella classificazione delle immagini, con una tendenza dei partecipanti a considerare i contenuti come reali più frequentemente rispetto a quanto atteso. Questo dato suggerisce una generale tendenza a fidarsi delle immagini, anche in presenza di contenuti potenzialmente generati artificialmente, e permette di rispondere positivamente anche al secondo quesito di ricerca:

RQ2. *I partecipanti mostrano un bias sistematico nel classificare le immagini come reali piuttosto che come generate da IA?*

Nell'insieme, i risultati delineano un quadro in cui la difficoltà di riconoscimento, la percezione soggettiva di autenticità e il livello di sicurezza nelle proprie valutazioni emergono come fattori chiave nella valutazione e nella potenziale diffusione dei contenuti visivi online.

5.2 Capacità di distinguere immagini reali e immagini generate da IA

I risultati del presente studio evidenziano una generale difficoltà dei partecipanti nel distinguere tra immagini reali e immagini generate da intelligenza artificiale, con un grado di accuratezza pari circa al 52%, confermando la prima ipotesi:

H1. *L'accuratezza nel distinguere immagini reali da immagini generate dall'IA non sarà significativamente superiore al 70%.*

Questo dato risulta coerente con quanto emerso nella letteratura recente, che mostra come la crescente qualità delle immagini sintetiche renda sempre più complesso il riconoscimento anche per utenti motivati e abituati all'uso delle tecnologie digitali.

In particolare, i risultati sono in linea con lo studio di Nightingale e Farid (2022), secondo cui i volti generati dall'IA risultano spesso indistinguibili da quelli reali e vengono talvolta percepiti come più affidabili. Analogamente, Lu et al. (2023) evidenziano livelli di accuratezza relativamente bassi nella distinzione tra immagini reali e sintetiche, con un tasso significativo di errori di classificazione. Anche lo studio su larga scala di Roca et al. (2025) conferma che la capacità umana di riconoscere immagini generate dall'IA si mantiene solo leggermente al di sopra del livello casuale, nonostante l'ampiezza del campione e la varietà degli stimoli utilizzati.

Un elemento particolarmente rilevante emerso nel presente studio riguarda la differenza tra immagini reali e immagini generate da IA: queste ultime risultano significativamente più difficili da identificare correttamente. Tale risultato è coerente con quanto osservato da Lu et al. (2023), secondo cui le persone identificano meglio le immagini reali e tendono più facilmente a scambiare per autentiche quelle generate da IA, e da Roca et al. (2025), che mostrano come le immagini generate dai modelli più recenti risultino particolarmente ingannevoli quando presentano caratteristiche visive naturali e non eccessivamente rifinite.

Questa maggiore difficoltà nel riconoscere le immagini IA può essere interpretata alla luce dei meccanismi percettivi e cognitivi descritti nella letteratura. Come evidenziato da Li et al. (2025), gli utenti tendono a basare i propri giudizi su indizi visivi superficiali, come la texture, il colore o la chiarezza dell'immagine, trascurando elementi più complessi ma potenzialmente più rilevanti, come l'analisi delle ombre o elementi quali il layout e la coerenza del tema. Poiché

i modelli generativi contemporanei sono ormai particolarmente efficaci nel riprodurre questi segnali superficiali, le immagini sintetiche possono risultare visivamente plausibili anche in presenza di imperfezioni meno evidenti.

In generale, i risultati suggeriscono che la difficoltà di riconoscimento osservata non dipende esclusivamente dal livello di sofisticazione tecnica delle immagini generate dall'IA, ma potrebbe essere influenzata anche dalle modalità con cui gli individui valutano i contenuti visivi. In linea con la letteratura, è possibile ipotizzare che gli utenti tendano ad affidarsi a criteri percettivi che non sempre risultano affidabili, contribuendo così a rendere il processo di distinzione tra reale e sintetico intrinsecamente incerto. Questo dato rafforza l'idea che il riconoscimento delle immagini generate dall'intelligenza artificiale rappresenti non solo una sfida tecnologica, ma anche un problema cognitivo e interpretativo, particolarmente rilevante nei contesti di fruizione quotidiana dei contenuti visivi, come i social media.

5.3 Variabilità tra immagini e ruolo delle caratteristiche visive

Oltre alla difficoltà generale nel distinguere tra immagini reali e immagini generate da intelligenza artificiale, i risultati evidenziano una significativa variabilità nei livelli di accuratezza a seconda della singola immagine considerata. Alcuni stimoli si sono rivelati particolarmente difficili da classificare correttamente, mentre altri sono stati riconosciuti con maggiore facilità, suggerendo che la natura del contenuto visivo giochi un ruolo rilevante nel processo di valutazione.

Questo risultato è coerente con quanto emerso nella letteratura. In particolare, Lu et al. (2023) mostrano come la capacità di distinguere immagini reali e sintetiche vari significativamente in base alla categoria di contenuto, con alcune tipologie di immagini più facilmente riconoscibili e altre più ambigue. Allo stesso modo, Roca et al. (2025) evidenziano come le immagini più ingannevoli siano spesso quelle che rappresentano scene quotidiane, contesti naturali o situazioni apparentemente ordinarie, caratterizzate da uno stile visivo plausibile ma non troppo artificioso. In particolare, le immagini con accuratezza più bassa raffigurano contesti plausibili e visivamente coerenti, come scene istituzionali o eventi di cronaca e, in alcuni casi, includono figure umane, che possono risultare particolarmente convincenti per gli utenti (fig. 5.3).

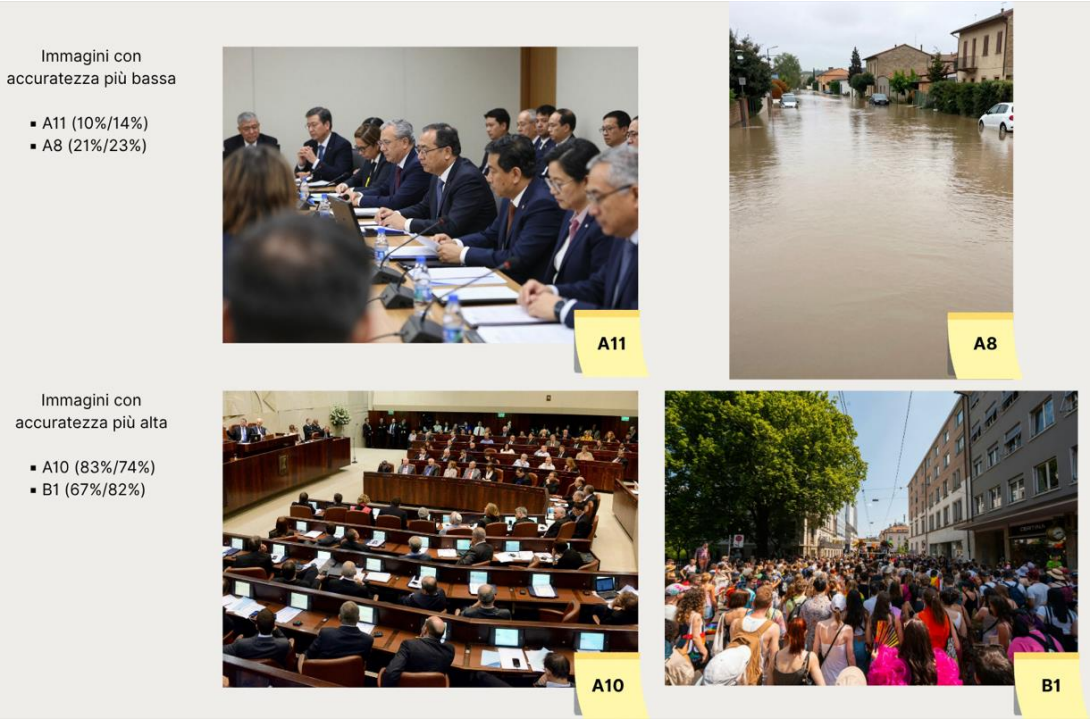


Figura 5.3 – Le immagini A11 e A8 (in alto) rappresentano gli stimoli con accuratezza più bassa, entrambe generate da IA, mentre le immagini A10 e B1 (in basso) corrispondono agli stimoli con accuratezza più elevata, entrambe reali. Per ciascuna immagine è riportata la percentuale di classificazione corretta nelle due versioni del questionario (A/B). Elaborazione grafica dell'autrice.

In questo senso, anche Nightingale e Farid (2022) evidenziano come i volti generati dall'IA possano risultare non solo indistinguibili da quelli reali, ma talvolta persino percepiti come più affidabili. Ciò suggerisce che la presenza di figure umane (in primo piano nell'immagine A11), soprattutto in contesti plausibili, possa contribuire a rendere le immagini sintetiche particolarmente convincenti e difficili da riconoscere.

I risultati del presente studio sembrano confermare questa tendenza: gli stimoli più difficili da classificare corrispondono a immagini generate da intelligenza artificiale, mentre quelli riconosciuti con maggiore facilità corrispondono a immagini reali. Questo dato è coerente con quanto emerso nelle analisi precedenti, che evidenziano una maggiore difficoltà nel riconoscimento delle immagini sintetiche. Tuttavia, anche all'interno di queste categorie emerge una certa variabilità nei livelli di accuratezza, suggerendo che il riconoscimento non dipenda unicamente dalla natura reale o sintetica dello stimolo, ma anche da caratteristiche visive e contestuali più specifiche.

Questo aspetto può essere ulteriormente interpretato alla luce del lavoro di Li et al. (2025), secondo cui gli utenti tendono a basare i propri giudizi su indizi visivi superficiali, come texture, colore e chiarezza. Se tali indizi risultano coerenti e plausibili, anche un'immagine generata

artificialmente può essere percepita come autentica. Al contrario, la presenza di anomalie visive, anche minime, può facilitare il riconoscimento dell'artificialità.
Nel complesso, questi risultati suggeriscono che la difficoltà di distinguere tra immagini reali e immagini generate da IA non sia uniforme, ma vari in funzione delle caratteristiche specifiche degli stimoli. Questo elemento è particolarmente rilevante nel contesto dei social media, dove circolano prevalentemente immagini di tipo quotidiano e non professionale, spesso simili a quelle risultate più difficili da classificare nel presente studio. Di conseguenza, la capacità di riconoscimento degli utenti può risultare ancora più limitata in contesti reali di fruizione.

5.4 Il ruolo della caption nel processo di valutazione

Uno degli obiettivi del presente studio era analizzare il ruolo della caption testuale nel processo di valutazione delle immagini. In particolare, si intendeva verificare se la presenza di un testo associato all'immagine potesse influenzare la capacità dei partecipanti di distinguere tra contenuti reali e contenuti generati da intelligenza artificiale, oltre al livello di sicurezza dichiarato nelle risposte.
I risultati non evidenziano un effetto significativo della caption né sull'accuratezza delle classificazioni né sulla confidence, smentendo le ipotesi H2 e H3:
H2. *La presenza di una caption influenzerà l'accuratezza della classificazione rispetto alla condizione senza caption.*
H3. *La presenza di una caption influenzerà il livello di sicurezza dichiarato nella risposta.*
Le analisi di clustering basate sul punteggio CRT risultano coerenti con questo risultato, mostrando che l'assenza di un effetto della caption si mantiene stabile nei diversi gruppi di partecipanti, indipendentemente dal loro livello di pensiero analitico (fig. 4.7.2).
Ciò suggerisce che, nel contesto dello studio, il testo associato all'immagine non modifica in modo rilevante il processo di valutazione visiva. Questo dato indica che i partecipanti abbiano basato il proprio giudizio principalmente sulle caratteristiche visive dello stimolo, senza essere influenzati in modo sistematico dal contesto testuale.
Questo risultato appare, almeno in parte, in contrasto con quanto suggerito dalla letteratura. In particolare, Newman e Schwarz (2024) evidenziano come il rapporto tra immagine e testo possa influenzare significativamente la percezione di veridicità di un contenuto. Secondo gli autori, anche immagini non manipolate possono risultare fuorvianti quando associate a un testo che ne orienta l'interpretazione, contribuendo ad aumentare la credibilità del messaggio attraverso meccanismi di *processing fluency* e *truthiness effect*. In questo senso, la presenza di una caption

dovrebbe teoricamente facilitare l'elaborazione del contenuto e influenzare il giudizio degli utenti.

L'assenza di un effetto significativo nel presente studio può essere interpretata alla luce di diverse possibili spiegazioni. In primo luogo, è possibile che le caption utilizzate non fossero sufficientemente informative, persuasive o fuorvianti da influenzare in modo sistematico il processo di valutazione. Come evidenziato da Newman e Schwarz (2024), l'effetto del contesto testuale dipende fortemente dalla relazione tra immagine e testo e dalla capacità di quest'ultimo di guidare l'interpretazione: una caption neutra o poco rilevante potrebbe non attivare tali meccanismi.

In secondo luogo, i partecipanti potrebbero aver adottato una strategia di valutazione focalizzata prevalentemente sugli aspetti visivi, trascurando il contenuto testuale. Questo comportamento potrebbe essere legato alla natura del compito sperimentale, che richiedeva esplicitamente di classificare l'immagine come reale o generata da IA, orientando l'attenzione verso indizi percettivi piuttosto che contestuali.

Infine, è possibile che l'effetto della caption emerga più chiaramente in contesti comunicativi più complessi e realistici, come i social media, dove immagine, testo e contesto narrativo sono strettamente integrati. In ambienti di questo tipo, la caption può contribuire non solo alla valutazione di autenticità, ma anche alla costruzione di significato e alla credibilità complessiva del messaggio, influenzando indirettamente anche i comportamenti degli utenti.

Nell'insieme, i risultati suggeriscono che, almeno nelle condizioni sperimentali adottate, il contesto testuale non rappresenta un fattore determinante nel processo di riconoscimento delle immagini. Tuttavia, alla luce della letteratura esistente, il ruolo della caption non può essere escluso, ma richiede ulteriori approfondimenti che tengano conto della complessità dei contesti comunicativi reali e della relazione tra immagine e testo.

5.5 Differenze individuali e divario tra percezione e performance

Un ulteriore obiettivo dello studio era analizzare il ruolo di alcune differenze individuali nella capacità di distinguere tra immagini reali e immagini generate da IA. In particolare, sono stati considerati il livello di media literacy percepita, il punteggio al Cognitive Reflection Test (CRT) e il livello di sicurezza dichiarato nelle risposte (confidence).

I risultati non evidenziano alcuna relazione statisticamente significativa tra queste variabili e l'accuratezza di riconoscimento. Infatti, né il livello di media literacy percepita né il punteggio

al CRT risultano associati a una maggiore capacità di distinguere correttamente le immagini, smentendo le ipotesi H5 e H6:

H5. *Un livello più alto di media literacy sarà associato a una maggiore accuratezza nel riconoscimento delle immagini generate artificialmente.*

H6. *Un punteggio più alto al Cognitive Reflection Test (CRT) sarà associato a una maggiore accuratezza nel distinguere immagini reali e sintetiche.*

Coerentemente, le analisi di clustering sul CRT evidenziano che, indipendentemente dal gruppo CRT, la maggior parte delle distribuzioni tende a concentrarsi attorno al 50-60% di accuratezza, confermando la difficoltà generale dei partecipanti nel distinguere immagini reali da quelle generate dall'IA, anche per chi ha un pensiero analitico più sviluppato (fig. 4.7.1 b).

Allo stesso modo, la confidence non mostra una correlazione significativa con la performance. Per approfondire ulteriormente questa dinamica, è stata condotta un'analisi di clustering basata congiuntamente su accuratezza e confidence (fig. 4.7.3):

- Il cluster dei *Fiduciosi ma Inaccurati,* che rappresenta il terzo gruppo per dimensione (45 partecipanti), include individui che tendono a essere eccessivamente sicuri delle proprie capacità di giudizio, nonostante le loro risposte siano, in media, ben al di sotto del livello casuale di accuratezza. Questo profilo suggerisce un potenziale *bias di overconfidence* in presenza di bassa performance e questi partecipanti, sentendosi sicuri nonostante gli errori, potrebbero essere più inclini a condividere informazioni errate, ritenendole autentiche;
- *I Mediocri ma Cauti* rappresenta il secondo cluster per dimensione (49 partecipanti). L'accuratezza di questi partecipanti è vicina al caso, e mostrano una confidence significativamente bassa. Questo potrebbe indicare una maggiore consapevolezza delle proprie difficoltà o una generale tendenza alla cautela, evitando di esprimere un'alta sicurezza quando la performance non la giustifica;
- *I Mediocri ma Molto Sicuri*: questo è il cluster più numeroso (76 partecipanti). Questi individui mostrano un'accuratezza di riconoscimento che è appena superiore al lancio di una moneta, ma, al contempo, dichiarano una confidence molto elevata. Questo profilo è un esempio classico di *overconfidence*, dove la sicurezza percepita non corrisponde a una performance effettiva;
- *I Competenti e Sicuri* rappresenta il gruppo più piccolo (32 partecipanti) ma anche quello con le migliori prestazioni. I partecipanti in questo cluster sono molto accurati nel distinguere le immagini reali da quelle IA e mostrano un livello di sicurezza nelle

loro valutazioni che è allineato con la loro effettiva competenza. Rappresentano il profilo ideale di chi riconosce correttamente ed è consapevole di farlo.

In sintesi, emerge una chiara dissociazione tra percezione di competenza (confidence) e capacità effettiva di riconoscimento (accuratezza), particolarmente evidente nei cluster caratterizzati da elevata sicurezza soggettiva ma bassa o media performance.

Tutto ciò evidenzia che fattori legati all'autopercezione delle proprie competenze e allo stile cognitivo riflessivo non si traducono necessariamente in una migliore capacità di valutazione dei contenuti visivi.

Un aspetto particolarmente rilevante riguarda la discrepanza tra percezione soggettiva e performance reale. I partecipanti riportano infatti livelli medi elevati di media literacy percepita, ma tali livelli non si riflettono in un'accuratezza di riconoscimento altrettanto elevata. Questo risultato è coerente con quanto evidenziato da Zak (2024), secondo cui esiste un divario strutturale tra le competenze dichiarate dagli utenti e la loro effettiva capacità di valutare contenuti visivi, soprattutto nel dominio della *visual misinformation*. In questo senso, gli individui possono ritenersi competenti nel valutare le informazioni online, pur applicando strategie di giudizio limitate o poco efficaci.

Anche il mancato effetto del CRT può essere interpretato alla luce della letteratura. Sebbene il pensiero riflessivo sia spesso associato a una maggiore capacità di valutazione critica, i risultati suggeriscono che, nel caso delle immagini generate da IA, il riconoscimento possa dipendere meno da processi deliberativi e più da meccanismi di giudizio immediati.

In questo senso, il risultato appare coerente con quanto evidenziato da Totti et al. (2024), secondo cui la valutazione della veridicità dei contenuti visivi non dipende esclusivamente da fattori percettivi, ma anche da caratteristiche individuali e da modalità soggettive di interpretazione. In particolare, gli autori mostrano come gli individui, soprattutto i più giovani, tendano a sovrastimare le proprie capacità di valutazione, evidenziando una discrepanza tra sicurezza percepita e performance reale. Questo elemento risulta in linea con quanto osservato nel presente studio, dove livelli elevati di confidence e media literacy percepita non risultano associati ad una maggiore accuratezza, suggerendo che il processo di valutazione possa essere influenzato più da percezioni soggettive che da effettive competenze.

In generale, la capacità di distinguere tra immagini reali e immagini generate da intelligenza artificiale non appare facilmente spiegabile attraverso variabili individuali tradizionalmente associate alla competenza informativa o al pensiero critico. Piuttosto, il processo di valutazione sembra influenzato da fattori percettivi e da strategie di giudizio che operano in modo rapido e spesso inconsapevole.

Questa evidenza ha implicazioni rilevanti anche sul piano applicativo. In particolare, suggerisce che interventi basati esclusivamente sul rafforzamento delle competenze auto-percepite o sulla promozione del pensiero riflessivo potrebbero non essere sufficienti a migliorare in modo significativo la capacità di riconoscimento delle immagini sintetiche. Al contrario, potrebbe essere necessario sviluppare interventi più specifici e mirati al dominio visivo, in grado di guidare l'attenzione degli utenti verso indizi più affidabili e ridurre la dipendenza da criteri percettivi fuorvianti.

5.6 Intenzione di condivisione, percezione di autenticità e confidence

I risultati relativi al comportamento di condivisione evidenziano alcune dinamiche rilevanti per la comprensione della diffusione dei contenuti visivi online.

Innanzitutto, emerge una generale bassa propensione alla condivisione nel campione, con valori medi contenuti su tutta la scala considerata. Questo dato suggerisce che, nel contesto sperimentale proposto, i partecipanti tendono a adottare un atteggiamento relativamente prudente rispetto alla diffusione dei contenuti.

Tuttavia, la propensione media alla condivisione varia a seconda dei quattro cluster identificati in base all'accuratezza e alla confidence (fig. 4.7.4 a):

- *I Fiduciosi ma Inaccurati* mostrano una propensione alla condivisione moderata, simile alla media complessiva;
- *I Mediocri ma Cauti* mostrano una bassa propensione alla condivisione, forse a causa della loro bassa confidence generale;
- *I Mediocri ma Molto Sicuri*, caratterizzati da *overconfidence*, hanno una propensione alla condivisione più elevata, il che risulta particolarmente rilevante, considerando la loro accuratezza appena superiore al caso;
- *I Competenti e Sicuri*, pur essendo i più accurati, non sono quelli con la più alta propensione alla condivisione, suggerendo che la loro competenza non si traduce automaticamente in una maggiore diffusione.

Analizzando più nel dettaglio le variabili coinvolte, emerge un risultato centrale: la probabilità di condivisione risulta significativamente più elevata per le immagini percepite come reali rispetto a quelle percepite come generate da intelligenza artificiale, confermando la quarta ipotesi:

H4. *Le immagini percepite come reali saranno associate a una maggiore intenzione di condivisione rispetto a quelle percepite come generate da IA*.

Le analisi di clustering risultano coerenti con questa tendenza (fig. 4.7.4 b):

- In tutti i cluster, ad eccezione dei *Mediocri ma Cauti*, le persone tendono a condividere significativamente di più le immagini che percepiscono come reali rispetto a quelle percepite come IA. Questa differenza è altamente significativa ($p < .001$) per quasi tutti i gruppi;
- Il gruppo dei *Mediocri ma Molto Sicuri* mostra la più alta propensione alla condivisione quando percepisce un'immagine come reale ($M = 2.84$ su 7). Questo è particolarmente rilevante, poiché la loro accuratezza media è solo al livello del caso;
- Anche *I Fiduciosi ma Inaccurati* mostrano una forte tendenza a condividere ciò che percepiscono come reale ($M = 2.50$), nonostante la loro bassa accuratezza complessiva;
- Solo per il cluster dei *Mediocri ma Cauti*, la differenza tra la condivisione di immagini percepite come reali e quelle percepite come IA non è statisticamente significativa ($p = .318$), suggerendo una maggiore cautela nella condivisione in generale, indipendentemente dalla percezione di autenticità.

In sintesi, emerge che la percezione di autenticità rappresenta un driver fondamentale per la condivisione online. Questo effetto risulta particolarmente marcato nei cluster caratterizzati da elevata sicurezza soggettiva, suggerendo che la combinazione tra percezione di autenticità e *overconfidence* possa amplificare la diffusione di contenuti erroneamente ritenuti reali.

Questo risultato è in linea con quanto evidenziato da Newman & Schwarz (2024), secondo cui le immagini influenzano il giudizio di veridicità non tanto per il loro contenuto informativo, ma per il modo in cui facilitano l'elaborazione cognitiva del messaggio. In particolare, la presenza di elementi visivi coerenti con il testo aumenta la cosiddetta *processing fluency*, ovvero la facilità con cui un contenuto viene elaborato, portando gli individui a giudicarlo come più credibile. Questo meccanismo, noto come *truthiness effect*, suggerisce che ciò che appare più plausibile viene anche percepito come più vero e, di conseguenza, più degno di essere condiviso.

Nel presente studio, tale dinamica si riflette nel fatto che le immagini percepite come autentiche risultano più condivisibili, indipendentemente dalla loro reale natura. Questo aspetto è particolarmente rilevante nei contesti social, dove la decisione di condividere un contenuto avviene spesso in modo rapido e basato su valutazioni intuitive.

Un ulteriore elemento significativo riguarda la relazione tra livello di sicurezza dichiarato e intenzione di condivisione. I risultati mostrano una correlazione positiva tra confidence e propensione alla condivisione, indicando che i partecipanti che si percepiscono più sicuri nelle proprie valutazioni tendono anche a condividere maggiormente i contenuti.
Questo dato permette di rispondere positivamente al primo quesito di ricerca esplorativo:
RQ1: *La propensione alla condivisione è associata al livello di sicurezza percepita nella classificazione delle immagini?*
Inoltre, suggerisce che il comportamento di condivisione è maggiormente associato alla fiducia soggettiva nel proprio giudizio che all'accuratezza effettiva della valutazione.
Il risultato è coerente con quanto emerso negli studi sulla disinformazione e sulla media literacy. In particolare, ricerche come quelle di Hwang et al. (2021) e El Mokadem (2023) mostrano che la percezione di credibilità e la sicurezza nel giudizio influenzano direttamente l'intenzione di condividere contenuti, anche quando questi risultano fuorvianti o inaccurati. In questi studi, la riduzione della credibilità percepita attraverso interventi di media literacy porta a una diminuzione della propensione alla condivisione, confermando il ruolo centrale delle variabili soggettive nei comportamenti di diffusione.
Al contrario, nel presente studio non emerge alcuna relazione significativa tra accuratezza di riconoscimento e intenzione di condivisione. Questo risultato indica che la capacità effettiva di distinguere tra immagini reali e immagini generate da IA non costituisce un fattore determinante nel comportamento di condivisione. In altre parole, gli individui non condividono necessariamente contenuti perché li riconoscono come autentici, ma perché credono che lo siano.
Nel complesso, la diffusione dei contenuti visivi online sembra essere fortemente influenzata da processi percettivi e metacognitivi, piuttosto che da una valutazione accurata dell'autenticità. La percezione di veridicità e il livello di sicurezza nel giudizio emergono come fattori chiave nel determinare il comportamento di condivisione, contribuendo a spiegare perché contenuti fuorvianti o generati artificialmente possano diffondersi con facilità nei contesti digitali, soprattutto nei profili caratterizzati da elevata sicurezza soggettiva.

5.7 Bias di classificazione

Un ulteriore risultato rilevante emerso dall'analisi riguarda la presenza di un bias sistematico nella classificazione delle immagini. In particolare, i partecipanti mostrano una tendenza significativa a classificare le immagini come reali più frequentemente rispetto a quanto atteso

in assenza di bias. Ciò consente di rispondere positivamente al secondo quesito di ricerca esplorativo:

RQ2: *I partecipanti mostrano un bias sistematico nel classificare le immagini come reali piuttosto che come generate da IA?*

Questo dato suggerisce l'esistenza di una predisposizione generale a considerare autentici i contenuti visivi, anche in un contesto in cui è esplicitamente richiesto di distinguere tra immagini reali e immagini generate da intelligenza artificiale. Tale tendenza può essere interpretata come una forma di bias di default verso l'autenticità percepita, che porta gli individui a fidarsi delle immagini in assenza di segnali chiari di manipolazione.

Questo risultato appare coerente con quanto evidenziato nella letteratura sul riconoscimento delle immagini sintetiche. Studi come quelli di Lu et al. (2023) e Roca et al. (2025) mostrano che le persone tendono a identificare più facilmente le immagini reali, mentre incontrano maggiori difficoltà nel riconoscere quelle generate da IA, spesso classificandole erroneamente come autentiche.

In questa direzione, anche Nightingale e Farid (2022) evidenziano come i volti generati dall'intelligenza artificiale possano risultare non solo indistinguibili da quelli reali, ma talvolta persino percepiti come più affidabili. Questo elemento suggerisce che le immagini sintetiche non solo sfuggano al riconoscimento, ma possano attivare una risposta di fiducia negli osservatori, contribuendo a rafforzare il bias verso l'autenticità.

Accanto agli aspetti percettivi, il bias osservato può essere interpretato anche alla luce di meccanismi cognitivi più generali. Come evidenziato da Totti et al. (2024), la valutazione della veridicità dei contenuti non dipende esclusivamente dalle caratteristiche dell'immagine, ma anche dalle aspettative e dalle strategie di giudizio adottate dagli individui. In particolare, gli utenti possono essere portati a considerare le immagini come autentiche perché coerenti con le proprie rappresentazioni della realtà o con il contesto in cui vengono presentate.

Inoltre, il fatto che le immagini utilizzate nello studio rappresentino scene plausibili e tipiche dei social media può aver rafforzato questa tendenza, rendendo più difficile attivare un atteggiamento critico.

In generale, la presenza di un bias verso la classificazione come "reale" evidenzia un limite importante nella valutazione dei contenuti visivi: gli individui non solo faticano a riconoscere le immagini generate da intelligenza artificiale, ma tendono anche a sottostimare la presenza di tali contenuti. Questo aspetto ha implicazioni rilevanti per la diffusione della disinformazione visiva, poiché una predisposizione alla fiducia può favorire l'accettazione e la condivisione di contenuti non autentici.

5.8 Implicazioni teoriche e pratiche

I risultati emersi dal presente studio offrono alcune implicazioni rilevanti sia sul piano teorico che su quello pratico, contribuendo ad ampliare la comprensione dei processi attraverso cui gli utenti valutano e condividono contenuti visivi nei contesti digitali contemporanei.

Dal punto di vista teorico, i dati confermano la complessità del riconoscimento delle immagini generate da intelligenza artificiale, evidenziando come tale processo non possa essere spiegato unicamente in termini percettivi o cognitivi isolati. In linea con la letteratura analizzata, la difficoltà di distinzione tra reale e sintetico emerge come il risultato dell'interazione tra caratteristiche delle immagini, modalità di presentazione e strategie di valutazione adottate dagli individui. In questa prospettiva, le analisi di clustering consentono di approfondire ulteriormente tali dinamiche, evidenziando l'esistenza di profili differenziati di utenti caratterizzati da diverse combinazioni di accuratezza e sicurezza percepita. In particolare, emerge come alcuni gruppi presentino una marcata discrepanza tra performance effettiva e percezione soggettiva, suggerendo che il processo di valutazione dei contenuti visivi non sia uniforme, ma vari significativamente tra individui con caratteristiche diverse.

Inoltre, il mancato effetto della media literacy percepita e del pensiero analitico sull'accuratezza suggerisce che le competenze percepite non si traducano automaticamente in una maggiore capacità di riconoscimento. Questi risultati risultano coerenti con quanto evidenziato da Zak (2024), che sottolinea l'esistenza di un divario strutturale tra competenze percepite e competenze effettive, soprattutto nel dominio visivo. La visual literacy emerge quindi come una dimensione distinta e ancora parzialmente sviluppata all'interno dei modelli tradizionali di media literacy. I risultati del clustering rafforzano questa evidenza, mostrando come livelli elevati di sicurezza soggettiva possano coesistere con performance modeste o basse, indicando che la percezione di competenza non rappresenta un indicatore affidabile della reale capacità di valutazione.

Allo stesso tempo, i risultati relativi alla propensione alla condivisione evidenziano il ruolo centrale della percezione soggettiva di autenticità e della sicurezza nel giudizio.

In particolare, le analisi di clustering mostrano che i profili caratterizzati da elevata confidence, anche in presenza di bassa accuratezza, tendono a manifestare una maggiore propensione alla condivisione, evidenziando come l'*overconfidence* possa rappresentare un fattore di rischio nella diffusione di contenuti fuorvianti.

In linea con quanto discusso da Newman e Schwarz (2024), le immagini influenzano la valutazione di veridicità attraverso meccanismi automatici legati alla facilità di elaborazione

cognitiva (*processing fluency*), piuttosto che attraverso un'analisi critica esplicita. Questo implica che la diffusione dei contenuti visivi online possa essere guidata da processi intuitivi e spesso inconsapevoli, che rendono gli utenti particolarmente vulnerabili a contenuti plausibili ma non autentici.

Un ulteriore elemento teorico rilevante riguarda l'assenza di un effetto significativo della caption sul riconoscimento delle immagini. Sebbene la letteratura suggerisca che il contesto testuale possa orientare l'interpretazione dei contenuti visivi, i risultati indicano che tale influenza non si traduce necessariamente in un miglioramento della capacità di detection. Questo dato suggerisce che il rapporto tra immagine e testo sia più complesso e che l'effetto della caption possa dipendere da variabili contestuali, come la coerenza semantica o il tipo di contenuto presentato.

Dal punto di vista pratico, i risultati evidenziano alcune criticità rilevanti per la progettazione di interventi educativi e strumenti di contrasto alla disinformazione visiva. In primo luogo, il fatto che né la media literacy percepita né il pensiero analitico risultino associati a una maggiore accuratezza suggerisce che gli interventi tradizionali potrebbero non essere sufficienti. In linea con quanto mostrato da Geissler et al. (2025), emerge la necessità di sviluppare strategie di media literacy più mirate al dominio visivo, basate su esempi concreti e su un'attenzione specifica agli indizi realmente diagnostici.

In secondo luogo, i risultati indicano che il comportamento di condivisione è influenzato principalmente da fattori soggettivi, come la percezione di autenticità e il livello di sicurezza nel giudizio, piuttosto che dall'accuratezza effettiva. Questo implica che gli interventi educativi dovrebbero non solo migliorare la capacità di riconoscimento, ma anche aiutare gli utenti a sviluppare una maggiore consapevolezza nei limiti del proprio giudizio. In questo senso, i risultati suggeriscono l'importanza di sviluppare interventi differenziati in funzione dei diversi profili di utenti, ad esempio intervenendo in modo specifico sui soggetti caratterizzati da elevata sicurezza soggettiva ma bassa accuratezza, al fine di ridurre il rischio di diffusione inconsapevole di contenuti errati.

Infine, i risultati suggeriscono che il problema del riconoscimento delle immagini generate da intelligenza artificiale non possa essere affrontato esclusivamente attraverso strumenti tecnici. Come evidenziato da Dehghani e Saberi (2025), i sistemi automatici di detection presentano limiti legati alla generalizzazione e all'evoluzione continua dei modelli generativi. Di conseguenza, appare necessario adottare un approccio integrato, che combini soluzioni tecnologiche e sviluppo di competenze critiche.

Complessivamente, lo studio evidenzia come la valutazione dei contenuti visivi nei contesti digitali sia il risultato di un equilibrio complesso tra processi percettivi, fattori cognitivi e dinamiche soggettive e metacognitive, che possono variare significativamente tra diversi profili di utenti. Le implicazioni emerse suggeriscono la necessità di ripensare il ruolo della media literacy in chiave più specifica e orientata al dominio visivo, al fine di affrontare in modo più efficace le sfide poste dalla crescente diffusione delle immagini generate dall'intelligenza artificiale.

5.9 Limiti dello studio

Il presente studio presenta alcuni limiti che è opportuno tenere in considerazione nell'interpretazione dei risultati.

In primo luogo, il campione è stato selezionato tramite una strategia di campionamento di convenienza, prevalentemente composto da utenti giovani e con un livello di istruzione medio-alto. Questo aspetto limita la generalizzabilità dei risultati alla popolazione generale, in quanto non è possibile escludere che gruppi con caratteristiche socio-demografiche differenti possano mostrare comportamenti e livelli di accuratezza diversi.

Un secondo limite riguarda la natura sperimentale e controllata del contesto di ricerca. Sebbene il questionario sia stato progettato per simulare un ambiente simile a quello dei social media, la situazione di valutazione rimane comunque artificiale. In particolare, i partecipanti erano consapevoli di essere coinvolti in uno studio e di dover distinguere tra immagini reali e generate da IA, condizione che potrebbe aver influenzato i giudizi dei partecipanti, rendendoli potenzialmente più riflessivi rispetto a quanto avviene in contesti di fruizione quotidiana.

Un ulteriore limite è legato al numero e alla tipologia di immagini utilizzate. Lo studio si basa su un insieme limitato di stimoli, selezionati per rappresentare specifiche categorie di contenuto (ad esempio eventi di cronaca, disastri naturali o contesti urbani). Sebbene tale scelta consenta di mantenere un buon controllo sperimentale, essa non riflette la varietà e la complessità dei contenuti visivi che circolano online, che includono formati, stili e contesti molto più eterogenei.

Per quanto riguarda la misura della media literacy, è importante sottolineare che l'indice utilizzato si basa su una valutazione auto-percepita delle competenze, e non su una misurazione oggettiva. Di conseguenza, i risultati relativi a questa variabile devono essere interpretati con cautela, poiché potrebbero riflettere più la percezione soggettiva degli individui che le loro effettive capacità.

Infine, alcune variabili potenzialmente rilevanti non sono state approfondite in modo sistematico. In particolare, sebbene siano state raccolte informazioni socio-demografiche (come età, titolo di studio e familiarità con l'IA), tali fattori sono stati analizzati solo in modo esplorativo e non sono stati inclusi in modelli più complessi. Inoltre, l'analisi di clustering, pur offrendo indicazioni utili per l'individuazione di profili comportamentali, ha una natura esplorativa e dipende dalle variabili selezionate e dalle scelte metodologiche adottate (ad esempio il numero di cluster). Pertanto, i risultati relativi ai cluster individuati devono essere interpretati con cautela e necessitano di ulteriori conferme.

Studi futuri potrebbero approfondire il ruolo di queste variabili, al fine di comprendere meglio come le differenze individuali influenzino la capacità di riconoscimento e i comportamenti di condivisione.

Nonostante questi limiti, lo studio fornisce indicazioni utili per comprendere le difficoltà legate al riconoscimento delle immagini generate da intelligenza artificiale e alcune dinamiche rilevanti nella valutazione e nella condivisione dei contenuti visivi nei contesti digitali.

5.10 Direzioni future di ricerca

Alla luce dei risultati emersi e dei limiti individuati, il presente studio apre diverse possibili direzioni per ricerche future.

In primo luogo, sarebbe utile ampliare la varietà e la numerosità degli stimoli utilizzati, includendo un numero maggiore di immagini e una gamma più ampia di categorie visive. In particolare, studi futuri potrebbero esplorare contesti differenti rispetto a quelli analizzati, come contenuti pubblicitari, di intrattenimento o immagini fortemente stilizzate, al fine di verificare se la capacità di riconoscimento vari in funzione del tipo di contenuto.

Inoltre, future ricerche potrebbero approfondire e validare i profili emersi dalle analisi di clustering, verificando la loro stabilità in campioni diversi e in contesti sperimentali diversi.

Un secondo ambito di approfondimento riguarda il ruolo del contesto in cui le immagini vengono presentate. Sebbene il presente studio abbia analizzato l'effetto della caption testuale, ulteriori ricerche potrebbero esaminare in modo più sistematico l'interazione tra immagine e testo, manipolando variabili come la coerenza semantica, il tono della caption o la presenza di segnali di credibilità (ad esempio fonte, numero di interazioni o commenti). Questo permetterebbe di comprendere meglio in che modo il contesto influenzi la percezione di autenticità e i comportamenti di condivisione.

Un ulteriore ambito di ricerca riguarda l'approfondimento del ruolo delle differenze individuali attraverso l'utilizzo di misure più articolate. In particolare, studi futuri potrebbero integrare indicatori oggettivi di media literacy e visual literacy, oltre ad analizzare in modo più sistematico variabili come l'età, il livello di istruzione, la frequenza di utilizzo dei social media e il grado di familiarità con i contenuti generati da intelligenza artificiale. Questo consentirebbe di comprendere in modo più approfondito quali fattori individuali influenzino la capacità di detection e le modalità di valutazione dei contenuti visivi.

Inoltre, ricerche future potrebbero esplorare in modo più approfondito la relazione tra profili di utenti e comportamento di condivisione, analizzando se e in che misura specifici cluster risultino più vulnerabili alla diffusione di contenuti fuorvianti. Questo permetterebbe di sviluppare modelli predittivi più accurati e interventi mirati.

Un ulteriore sviluppo riguarda la progettazione e la valutazione di interventi di media literacy specificamente orientati al dominio visivo. In linea con quanto suggerito dalla letteratura, sarebbe utile testare l'efficacia di interventi brevi basati su esempi concreti e feedback immediato, verificandone non solo l'impatto immediato, ma anche la durata nel tempo.

In questo senso, ulteriori ricerche potrebbero analizzare se e in che modo le competenze acquisite vengono mantenute nel tempo. In particolare, sarebbe rilevante indagare strategie volte a ridurre l'*overconfidence*, ad esempio attraverso feedback correttivi o meccanismi di consapevolezza metacognitiva, per verificare se una maggiore consapevolezza dei propri limiti possa influenzare positivamente i comportamenti di condivisione.

Infine, una direzione particolarmente rilevante riguarda il possibile utilizzo congiunto di capacità umane e strumenti tecnologici. Ulteriori ricerche potrebbero indagare in che modo i sistemi automatici di rilevazione possano supportare gli utenti nella valutazione dei contenuti visivi, ad esempio fornendo indicazioni utili o segnalando elementi potenzialmente sospetti. Questo approccio potrebbe contribuire a ridurre le difficoltà di riconoscimento, affiancando le capacità umane senza sostituirle, e favorendo una maggiore attenzione critica nei confronti dei contenuti visivi.

Nel complesso, le direzioni future individuate evidenziano la necessità di un approccio multidimensionale allo studio delle immagini generate da intelligenza artificiale, che tenga conto non solo delle caratteristiche tecniche dei contenuti, ma anche delle modalità con cui questi vengono valutati e utilizzati nei contesti sociali e comunicativi.

CONCLUSIONI

Il presente lavoro si inserisce nel crescente dibattito relativo alla diffusione di contenuti visivi generati artificialmente e alle difficoltà degli utenti nel valutarne l'autenticità. In questo contesto, la ricerca ha analizzato la capacità degli utenti di distinguere tra immagini reali e immagini generate da intelligenza artificiale, con particolare attenzione al ruolo della caption testuale e di alcune variabili individuali e percettive nel processo di valutazione.

I risultati evidenziano come, nel complesso, la capacità di riconoscimento risulti poco superiore al caso, confermando la difficoltà degli utenti nel discriminare tra contenuti autentici e contenuti sintetici. In particolare, le immagini generate da IA si sono rivelate più difficili da identificare rispetto a quelle reali, suggerendo come i progressi tecnologici abbiano reso tali contenuti sempre più plausibili e difficilmente distinguibili sulla base di indizi visivi.

Contrariamente a quanto ipotizzato, la presenza della caption non ha mostrato un effetto significativo sulla capacità di riconoscimento. Questo risultato suggerisce che, almeno nel contesto analizzato, l'informazione testuale non sia sufficiente a migliorare l'accuratezza delle valutazioni, evidenziando una prevalenza dell'elaborazione visiva rispetto a quella testuale.

Le analisi di clustering hanno inoltre consentito di approfondire il rapporto tra performance effettiva e percezione soggettiva, evidenziando la presenza di profili differenziati di partecipanti. In particolare, accanto ad un gruppo di utenti accurati e coerenti nella valutazione delle proprie capacità, sono emersi cluster caratterizzati da elevata sicurezza soggettiva a fronte di accuratezza moderata o bassa. Questo risultato rafforza l'idea che la confidence non rappresenti necessariamente un indicatore affidabile della reale capacità di riconoscimento.

Allo stesso tempo, emergono elementi rilevanti legati alla dimensione soggettiva del giudizio. In particolare, la percezione di autenticità e il livello di sicurezza dichiarato influenzano significativamente l'intenzione di condivisione, mentre non si osserva una relazione tra accuratezza effettiva e comportamento di diffusione. Le analisi di clustering mostrano inoltre che i profili caratterizzati da maggiore sicurezza soggettiva tendono anche a presentare livelli più alti di propensione alla condivisione, soprattutto quando le immagini sono percepite come reali. Ciò indica che gli individui tendono a condividere contenuti non necessariamente perché li riconoscono come autentici, ma perché credono che lo siano.

Un ulteriore risultato riguarda la presenza di un bias sistematico verso la classificazione delle immagini come reali, che suggerisce una predisposizione generale alla fiducia nei confronti dei contenuti visivi. Questo elemento, insieme alla difficoltà di riconoscimento delle immagini

sintetiche, evidenzia un rischio concreto nei contesti digitali, in cui contenuti non autentici possono essere percepiti come credibili e diffusi con facilità.

Nell'insieme, i risultati dello studio mettono in luce come la valutazione dei contenuti visivi non dipenda esclusivamente da competenze oggettive, ma sia fortemente influenzata da processi percettivi e metacognitivi. Tali evidenze sottolineano l'importanza di sviluppare strumenti e interventi in grado di supportare gli utenti non solo nel riconoscimento dei contenuti visivi sintetici, ma anche nella consapevolezza dei limiti del proprio giudizio.

In un ecosistema informativo sempre più caratterizzato dalla presenza di contenuti sintetici, comprendere i meccanismi che guidano la percezione e la diffusione delle immagini rappresenta una sfida centrale per la comunicazione digitale contemporanea.

Bibliografia

Fonti bibliografiche

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, *13*(1), 210-230.

Dehghani, A., & Saberi, H. (2025). *Generating and detecting various types of fake image and audio content: A review of modern deep learning technologies and tools*. *arXiv preprint arXiv:2501.06227*.

El Mokadem, S. S. (2023). The Effect of Media Literacy on Misinformation and Deep Fake Video Detection. *Journal of Arab Media & Society*, *35*, 53-78.

Eyal, N. (2015). Creare prodotti e servizi per catturare i clienti (Hooked). edizioni LSWR.

Geissler, D., Robertson, C., & Feuerriegel, S. (2025). Digital literacy interventions can boost humans in discerning deepfakes. *arXiv preprint arXiv:2507.23492*.

Harris, T. (2016). How technology hijacks people's minds—from a magician and Google's design ethicist. *Medium Magazine*, *18*.

Hwang, Y., Ryu, J. Y., & Jeong, S. H. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 188-193.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, *53*(1), 59-68.

Li, Y., Liu, X., Wang, X., Lee, B. S., Wang, S., Rocha, A., & Lin, W. (2025). *Fakebench: Probing explainable fake image detection via large multimodal models*. *IEEE Transactions on Information Forensics and Security*.

Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., & Ouyang, W. (2023). Seeing is not always believing: Benchmarking human and model perception of ai-generated images. *Advances in neural information processing systems*, *36*, 25435-25447.

Newman, E. J., & Schwarz, N. (2024). Misinformed by images: How images influence perceptions of truth and what can be done about it. *Current Opinion in Psychology*, *56*, 101778.

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), e2120481119.
Odell, J. (2021). *Come non fare niente: Resistere all'economia dell'attenzione*. HOEPLI EDITORE.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. penguin UK.

Roca, T., Roman, A. C., Vega, J. T., Duarte, M., Wang, P., White, K., ... & Ferres, J. L. (2025). How good are humans at detecting AI-generated images? Learnings from an experiment. *arXiv preprint arXiv:2507.18640*.

Sha, Z., Tan, Y., Li, M., Backes, M., & Zhang, Y. (2024, December). *Zerofake: Zero-shot detection of fake images generated and edited by text-to-image generation models*. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (pp. 4852-4866).

Tanfoni, M., Ceroni, E. G., Marziali, S., Pancino, N., Maggini, M., & Bianchini, M. (2024). *Generated or Not Generated (GNG): The Importance of Background in the Detection of Fake Images*. *Electronics*, *13*(16), 3161.

Totti, Cavicchioli, Furini, Tagliani (2024). *Seeing is Believing? Understanding Truth in Synthetic Media*.

Wen, S., Ye, J., Feng, P., Kang, H., Wen, Z., Chen, Y., ... & Li, W. (2025). *Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation*. *arXiv preprint arXiv:2503.14905*.

Zak, E. (2024). *Can I Trust This Image?: Evaluating the Relationship Between Information Literacy Skills and the Ability to Identify Visual Misinformation* (Doctoral dissertation, The University of Iowa).

Fonti iconografiche

Dell'Olio, V. (2020). Il modello del gancio di Nir Eyal (da Hooked).
Recuperato da https://vincenzodellolio.com/il-modello-del-gancio-di-nir-eyal-da-hooked/

Apple. (s.d.). Tasto "Controllo fotocamera" di iPhone 16.
Recuperato da https://support.apple.com/it-it/guide/iphone/iph0c397b154/ios

Appendice A – Questionari

A.1 Questionario A

Titolo del questionario: “Percezione delle immagini e dei contenuti sui social media”.

Il seguente questionario ha finalità di ricerca accademica.
Le risposte sono anonime e verranno analizzate in forma aggregata.
Non esistono risposte giuste o sbagliate: ci interessa esclusivamente la tua percezione.
Per partecipare è necessario avere almeno 18 anni.

CRT (Cognitive Reflection test)

1. Una mazza e una palla costano in totale €1,10. Sapendo che la mazza costa €1 in più della palla, quanto costa la palla?
2. Se 5 macchine impiegano 5 minuti per produrre 5 pezzi, quanto tempo impiegano 100 macchine per produrre 100 pezzi?
3. In un lago ci sono delle ninfee che raddoppiano di numero ogni giorno. Se ci vogliono 48 giorni perché coprano tutto il lago, in quanti giorni ne coprono metà?

Media literacy

Di seguito troverai alcune affermazioni relative al modo in cui utilizzi i social media e valuti le informazioni online.
Indica quanto sei d’accordo con ciascuna affermazione.

1. Prima di condividere un post sui social, controllo se la fonte è affidabile.
2. Sono consapevole che molte immagini online possono essere generate da intelligenza artificiale.
3. Ritengo di saper riconoscere alcuni segnali tipici di immagini generate artificialmente.
4. Quando vedo una notizia online, cerco conferme da più fonti.
5. Penso sia importante verificare le informazioni prima di condividerle sui social.

(Scala Likert 1-7, dove 1 = Per niente d’accordo e 7 = Totalmente d’accordo).

Valutazione delle immagini

Nella prossima sezione ti verrà chiesto di osservare 12 immagini e rispondere a brevi domande.

Foto 1/12

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

Foto 2/12

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

Foto 3/12

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

Foto 4/12

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

Foto 5/12

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

Foto 6/12

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

Foto 7/12

Corteo in centro città durante una manifestazione per i diritti civili. #manifestazione #città

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?
  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).
- Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?
  (Scala Likert 1-7, dove 1 = Per nulla probabile e 7 = Molto probabile).

Foto 8/12

Danni alle auto dopo la grandinata di oggi. #maltempo #bergamo

- Secondo te, questa immagine è:

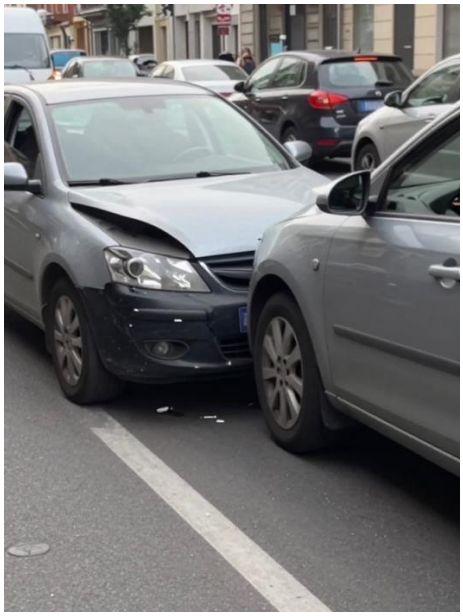
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?
  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).
- Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?
  (Scala Likert 1-7, dove 1 = Per nulla probabile e 7 = Molto probabile).

Foto 9/12

Intervento dei vigili del fuoco. Brutto incendio in un edificio residenziale periferico. #cronaca #incendio

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?
  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).
- Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?
  (Scala Likert 1-7, dove 1 = Per nulla probabile e 7 = Molto probabile).

Foto 10/12



Incidente in centro stamattina… traffico pazzesco. #incidenti #milano

- Secondo te, questa immagine è:
  - Foto reale
  - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?
  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).
- Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?
  (Scala Likert 1-7, dove 1 = Per nulla probabile e 7 = Molto probabile).

Foto 11/12

Intervento dei vigili del fuoco davanti a una chiesa dopo una segnalazione. #cronaca #intervento

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?
  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).
- Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?
  (Scala Likert 1-7, dove 1 = Per nulla probabile e 7 = Molto probabile).

Foto 12/12

Strade come fiumi… impressionante quello che sta succedendo in Toscana. #allerta #maltempo

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?
  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).
- Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?
  (Scala Likert 1-7, dove 1 = Per nulla probabile e 7 = Molto probabile).

Informazioni generali

Le seguenti domande sono anonime e verranno utilizzate solo a fini di ricerca.

- Età
    - 18-24
    - 25-34
    - 35-44
    - 45+
- Titolo di studio
    - Licenza media
    - Scuola superiore
    - Laurea triennale
    - Laurea magistrale / Post-laurea
- Frequenza di utilizzo dei social media
    - Più volte al giorno
    - Una volta al giorno
    - Qualche volta a settimana
    - Raramente
- Familiarità con immagini generate da intelligenza artificiale
    - Molto
    - Abbastanza
    - Poco
    - Per nulla

Alcune delle immagini mostrate potevano essere generate tramite intelligenza artificiale.
Il questionario fa parte di una ricerca accademica sul rapporto tra immagini digitali, credibilità e condivisione sui social media.
Grazie per il tempo dedicato.

A.2 Questionario B

Titolo del questionario: “Percezione delle immagini e dei contenuti sui social media”.

Il seguente questionario ha finalità di ricerca accademica.
Le risposte sono anonime e verranno analizzate in forma aggregata.

Non esistono risposte giuste o sbagliate: ci interessa esclusivamente la tua percezione. Per partecipare è necessario avere almeno 18 anni.

CRT (Cognitive Reflection test)

4. Una mazza e una palla costano in totale €1,10. Sapendo che la mazza costa €1 in più della palla, quanto costa la palla?
5. Se 5 macchine impiegano 5 minuti per produrre 5 pezzi, quanto tempo impiegano 100 macchine per produrre 100 pezzi?
6. In un lago ci sono delle ninfee che raddoppiano di numero ogni giorno. Se ci vogliono 48 giorni perché coprano tutto il lago, in quanti giorni ne coprono metà?

Media literacy

Di seguito troverai alcune affermazioni relative al modo in cui utilizzi i social media e valuti le informazioni online.

Indica quanto sei d'accordo con ciascuna affermazione.

6. Prima di condividere un post sui social, controllo se la fonte è affidabile.
7. Sono consapevole che molte immagini online possono essere generate da intelligenza artificiale.
8. Ritengo di saper riconoscere alcuni segnali tipici di immagini generate artificialmente.
9. Quando vedo una notizia online, cerco conferme da più fonti.
10. Penso sia importante verificare le informazioni prima di condividerle sui social.

(Scala Likert 1-7, dove 1 = Per niente d'accordo e 7 = Totalmente d'accordo).

Valutazione delle immagini

Nella prossima sezione ti verrà chiesto di osservare 12 immagini e rispondere a brevi domande.

Foto 1/12

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

Foto 2/12

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

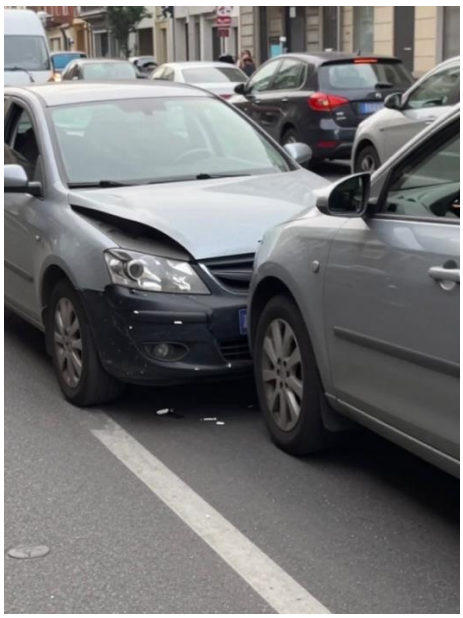
  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

Foto 3/12

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

Foto 4/12



- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

Foto 5/12

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?
  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

Foto 6/12

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale

- Quanto sei sicuro/a della tua risposta?

  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

Foto 7/12

Strada completamente allagata dopo le forti piogge delle ultime ore. #maltempo #allagamenti

- Secondo te, questa immagine è:
  - Foto reale
  - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).
- Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?

  (Scala Likert 1-7, dove 1 = Per nulla probabile e 7 = Molto probabile).

Foto 8/12

Manifestazione in centro città, centinaia di persone in piazza. #protesta #attualità

- Secondo te, questa immagine è:
  - Foto reale
  - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?
  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).
- Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?
  (Scala Likert 1-7, dove 1 = Per nulla probabile e 7 = Molto probabile).

Foto 9/12

Tornado colpisce un'area residenziale periferica, danni alle abitazioni. #meteoestremo #tornado

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?
  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).
- Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?
  (Scala Likert 1-7, dove 1 = Per nulla probabile e 7 = Molto probabile).

Foto 10/12

Incontro istituzionale tra rappresentanti politici durante una seduta parlamentare. #politica #istituzioni

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?
  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).
- Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?
  (Scala Likert 1-7, dove 1 = Per nulla probabile e 7 = Molto probabile).

Foto 11/12

Incendio ai piani alti di un condominio, sul posto i vigili del fuoco. #cronaca #incendio

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

  (Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).
- Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?

  (Scala Likert 1-7, dove 1 = Per nulla probabile e 7 = Molto probabile).

Foto 12/12

Vertice internazionale tra delegazioni politiche. #geopolitica #incontro

- Secondo te, questa immagine è:
    - Foto reale
    - Immagine generata da intelligenza artificiale
- Quanto sei sicuro/a della tua risposta?

(Scala Likert 1-7, dove 1 = Per nulla sicuro/a e 7 = Totalmente sicuro/a).

- Se vedessi questo post sui social, quanto sarebbe probabile che tu lo condividessi?
  (Scala Likert 1-7, dove 1 = Per nulla probabile e 7 = Molto probabile).

Informazioni generali

Le seguenti domande sono anonime e verranno utilizzate solo a fini di ricerca.

- Età
  - 18-24
  - 25-34
  - 35-44
  - 45+
- Titolo di studio
  - Licenza media
  - Scuola superiore
  - Laurea triennale
  - Laurea magistrale / Post-laurea
- Frequenza di utilizzo dei social media
  - Più volte al giorno
  - Una volta al giorno
  - Qualche volta a settimana
  - Raramente
- Familiarità con immagini generate da intelligenza artificiale
  - Molto
  - Abbastanza
  - Poco
  - Per nulla

Alcune delle immagini mostrate potevano essere generate tramite intelligenza artificiale.
Il questionario fa parte di una ricerca accademica sul rapporto tra immagini digitali, credibilità e condivisione sui social media.
Grazie per il tempo dedicato.

Appendice B –
Stimoli visivi utilizzati nel questionario

B.1 Immagini reali

A1 - Disastro naturale - strada allagata con auto mezze sommerse

- Autore: Jiří Sedláček.
- Titolo: *Cars and pathway in flooded Novodvorská street in Třebíč, Třebíč District*.
- Fonte: Wikimedia Commons.
- Licenza: Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA 4.0).
- URL: https://commons.wikimedia.org/wiki/File:Cars_and_pathway_in_flooded_Novodvorsk%C3%A1_street_in_T%C5%99eb%C3%AD%C4%8D,_T%C5%99eb%C3%AD%C4%8D_District.JPG
- Caption utilizzata nella condizione con testo:
  "Strada completamente allagata dopo le forti piogge delle ultime ore. #maltempo #allagamenti"

A5 - Incendio - edificio residenziale con fumo e intervento dei vigili del fuoco

- Autore: National Institute of Standards and Technology (NIST).
- Titolo: *High-Rise Fire Test; Positive Pressure Ventilation Fans*.
- Fonte: Wikimedia Commons.
- Licenza: Public Domain (CC0).
- URL: https://commons.wikimedia.org/wiki/File:High-Rise_Fire_Test;_Positive_Pressure_Ventilation_Fans_(5887634583)_(cropped).jpg
- Caption utilizzata nella condizione con testo:
  "Incendio ai piani alti di un condominio, sul posto i vigili del fuoco. #cronaca #incendio"

A6 - Evento meteorologico - auto danneggiata dalla grandine

- Autore: Simiprof.
- Titolo: *Hail damage car*.
- Fonte: Wikimedia Commons.
- Licenza: Public Domain (CC0).
- URL: https://commons.wikimedia.org/wiki/File:Hail_damage_car.jpg
- Caption utilizzata nella condizione con testo:

  “Danni alle auto dopo la grandinata di oggi. #maltempo #bergamo”

A9 - Cronaca locale - pompieri davanti ad una chiesa

- Autore: Dsns.gov.ua.
- Titolo: *Transfiguration Cathedral in Odesa after Russian missile attack, 2023-07-23*.
- Fonte: Wikimedia Commons.
- Licenza: Creative Commons Attribution 4.0 (CC BY 4.0).
- URL: https://commons.wikimedia.org/wiki/File:Transfiguration_Cathedral_in_Odesa_after_Russian_missile_attack,_2023-07-23_(11).jpg
- Caption utilizzata nella condizione con testo:
  "Intervento dei vigili del fuoco davanti a una chiesa dopo una segnalazione. #cronaca #intervento"

A10 - Politica - aula parlamentare reale

- Autore: Rafael Nir.
- Fonte: Unsplash.
- Licenza: Unsplash License.
- URL: https://unsplash.com/it/foto/gruppo-di-persone-sedute-sulle-sedie-dtP89KBWrxE
- Caption utilizzata nella condizione con testo:
  “Incontro istituzionale tra rappresentanti politici durante una seduta parlamentare. #politica #istituzioni”

B1 - Protesta - corteo reale

- Autore: Kajetan Sumila.
- Fonte: Unsplash.
- Licenza: Unsplash License.
- URL: https://unsplash.com/it/foto/una-folla-di-persone-che-camminano-lungo-una-strada-accanto-a-edifici-alti-nBK5vCrSJDE
- Caption utilizzata nella condizione con testo:
  “Corteo in centro città durante una manifestazione per i diritti civili. #manifestazione #città”

B.2 Immagini generate tramite IA

A3 - Protesta - folla con cartelli poco leggibili

- Strumento: Stable Diffusion
- Prompt: “A candid smartphone photo of a public protest in a city square. A large crowd of people holding protest signs and banners. The signs contain text, but the writing is partially unclear and hard to read. Daytime, natural lighting, overcast sky. Amateur photo taken from inside the crowd. Realistic colors, slightly imperfect framing. Social media style, non-professional. No artistic effects, no cinematic lighting, no dramatic composition.”
- Prompt negativo: “Artistic style, illustration, painting, hyperrealistic, cinematic lighting, perfect symmetry, sharp focus, studio photo, ultra detailed faces, clear readable text, slogans clearly readable, professional photography. ”
- Caption utilizzata nella condizione con testo:
  “Manifestazione in centro città, centinaia di persone in piazza. #protesta #attualità”

A7 - Disastro naturale - tornado vicino a case

- Strumento: Stable Diffusion
- Prompt: “A low-quality documentary photograph of a real tornado touching down in a suburban neighborhood. Shot from inside a parked car or from a front porch, partially sheltered from the rain. Uneven framing, slight tilt, imperfect composition. Heavy rain and wind reducing visibility, with streaks of rain visible across the image. The tornado funnel is irregular, partially obscured by rain and debris, not fully defined. Houses, trees, and street elements appear slightly blurred by motion and weather. Storm clouds form thick layered bands across the sky, with some cloud layers appearing unusually similar in shape and spacing, but only noticeable on closer inspection. Flat lighting, muted colors, visible noise, realistic atmospheric texture. Amateur disaster photo, not cinematic, not dramatic.”
- Prompt negativo: “illustration, painting, digital art, cinematic lighting, HDR, ultra sharp, perfectly centered tornado, clean edges, smooth gradients, stylized storm, fantasy disaster, professional stock photography, dramatic sky, extreme contrast, logos, text, watermarks.”
- Caption utilizzata nella condizione con testo:
  “Tornado colpisce un’area residenziale periferica, danni alle abitazioni. #meteoestremo #tornado”

A11 - Geopolitica - sala conferenze con figure politiche al tavolo

- Strumento: Stable Diffusion
- Prompt: “A realistic amateur-style photo of a political or geopolitical meeting inside a modern conference room. A large group of officials and delegates seated closely together around long tables with microphones, documents notebooks, water bottles. People are wearing formal or semi-formal suits, but the scene does not look staged or official. The photo is taken from a slightly off-center angle, from the back or side of the room, as if captured by a journalist or attendee. Some people are partially out of focus, some faces are only partially visible, others are obscured by heads in the foreground. The composition feels crowded and imperfect. Neutral institutional interior, plain walls, no flags, no logos, no banners, no visible text. Natural indoor lighting, slightly uneven exposure, mild motion blur in some areas. Medium to low sharpness, realistic noise and grain, smartphone or handheld camera look. Documentary news photography, unpolished, non-commercial, realistic perspective.”
- Prompt negativo: “flags, national symbols, banners, logos, readable text, official press conference, podium, speaker at lectern, perfectly centered composition, stock photo, promotional image, cinematic lighting, dramatic lighting, HDR, ultra sharp, studio lighting, cartoon, illustration, painting, 3D render, surreal, grotesque faces, horror, extreme deformation, extra limbs, perfect symmetry, clean background, empty space.”
- Caption utilizzata nella condizione con testo:
  “Vertice internazionale tra delegazioni politiche. #geopolitica #incontro”

B3 - Incendio - vigili del fuoco in azione davanti a un edificio in fiamme

- Strumento: Stable Diffusion
- Prompt: “Smartphone photo of a building fire in an urban street. A two-story residential or mixed-use building with flames visible from several upper windows, dark smoke rising into the sky. Firefighters and emergency vehicles present in the scene, positioned naturally and not symmetrically. Some bystanders visible at a distance, not individually recognizable. Wide horizontal framing, clear overall focus, slight depth variation but no motion blur. Natural daylight, overcast sky, realistic colors. Documentary news photography style, ordinary local incident, unpolished but realistic.”
- Prompt negativo: “Blurry image, out of focus, motion blur, cinematic lighting, dramatic fire, explosion, apocalyptic smoke, exaggerated flames, perfect composition, stock photo, professional studio lighting, HDR, ultra sharp, illustration, painting, digital art, CGI.”
- Caption utilizzata nella condizione con testo:
  “Intervento dei vigili del fuoco. Brutto incendio in un edificio residenziale periferico. #cronaca #incendio”

A8 - Evento meteorologico - alluvione con strada allagata

- Strumento: Stable Diffusion
- Prompt: “A realistic photo of a severe flood in a town in Tuscany, Italy. Streets completely flooded, water covering the road like a river, partially submerged cars and sidewalks, residential buildings typical of central Italy. Overcast rainy weather, natural daylight. The flood water appears unnaturally smooth and uniform, with a flat surface, minimal ripples or waves, slightly unrealistic consistency compared to real flood water. Documentary news photography style, realistic colors, medium sharpness, no artistic effects.”
- Prompt negativo: “Dramatic waves, splashing water, turbulent water, realistic water dynamics, foam, debris-filled water, strong reflections, cinematic lighting, HDR, ultra sharp, digital art, illustration, painting.”
- Caption utilizzata nella condizione con testo:
  “Strade come fiumi… impressionante quello che sta succedendo in Toscana. #allerta #maltempo”

A12 - Cronaca urbana - incidente con auto deformata

- Strumento: Stable Diffusion
- Prompt: “A low-quality realistic photo of a traffic accident on a ring road in Milan, Italy. Damaged car stopped on the road, visible deformation of the vehicle body after a crash. Urban road environment, traffic congestion in the background, other cars stopped. Amateur news photo, taken quickly, imperfect framing, slight blur, uneven focus, natural daylight. Moderate image noise, realistic colors, no dramatic effects. The damaged car shows slightly unnatural deformation patterns, subtle inconsistencies in metal bending typical of AI-generated images, while the rest of the scene appears normal.”
- Prompt negativo: “Cinematic lighting, dramatic scene, HDR, ultra sharp, hyper detailed, professional photography, studio lighting, surreal deformation, melted car, cartoon, illustration, digital art, perfectly clean image, stock photo, staged accident.”
- Caption utilizzata nella condizione con testo:
  “Incidente in centro stamattina… traffico pazzesco. #incidenti #milano

Appendice C –
Codice di analisi dei dati

Il codice utilizzato per l'analisi dei dati è disponibile al seguente link:
https://colab.research.google.com/drive/1dEyu143j_SXhVcvp5tBUSaKMBiRloGTv?usp=sharing
Il file include le procedure di pulizia del dataset, le analisi statistiche e di clustering, nonché le visualizzazioni presentate nel Capitolo 4.