FOI MEMO
Projekt
Mediemyndigheten

Sidnr
1 (35)

Projektnummer
E13979

Uppdragsgivare
Mediemyndigheten

FoT-område
Inget FoT-område

Författare
Sofiya Voytiv, Mattias Svahn

Datum
2026-01-27

Memo nummer
FOI Memo 9187

AI-genererat innehåll och desinformation på sociala medier – en systematisk forskningsöversikt.

Titel Memo nummer 9187

AI-genererat innehåll och desinformation på sociala medier – en systematisk forskningsöversikt. FOI

Sammanfattning

Denna systematiska litteraturöversikt kartlägger forskning om AI-genererat innehåll på sociala medier och analyserar tekniska, juridiska och samhälleliga konsekvenser. Studien omfattar peer-review-artiklar och grå litteratur publicerad mellan juni 2022 och juni 2025 och identifierar tre centrala teman: generativ AI som innehåll, användning av AI i desinformation samt metoder för autenticitetsverifiering.

Forskningen domineras av teknikorienterade ansatser, ofta experimentella och med fokus på text-, bild- och videodetektion, särskilt på Twitter/X. Samtidigt finns betydande kunskapsluckor kring barns och ungas exponering, multimodala innehåll på plattformar som TikTok och Roblox samt användarbeteenden i vardagliga miljöer. De rättsliga analyserna belyser bland annat ansvarsfördelning, upphovsrätt, GDPR, och konsekvenser av EU AI Act i relation till plattformars skyldigheter.

AI används både för legitima syften och i skadliga sammanhang som bedrägerier, utpressning, politisk påverkan och icke-samtyckt pornografi. Flera studier visar att användare har begränsad förmåga att identifiera syntetiskt innehåll, vilket förstärker riskerna för desinformation och minskat förtroende för journalistik. Policydiskussionen kretsar kring märkning av AI-innehåll och behovet av robustare reglering utan att hämma innovation.

Översikten visar behov av tvärvetenskaplig forskning, särskilt kring barns sårbarhet, nordiska mediekontexter, praktiskt fungerande detektion samt effekter av nya regelverk på digitala ekosystem.

Nyckelord: Generativ AI, Desinformation, Äkthetsverifiering, Sociala medier

Summary

This systematic literature review examines research on AI-generated content on social media, focusing on technological, legal, and societal implications. Covering peer-reviewed studies and relevant grey literature published between June 2022 and June 2025, it identifies three core areas: AI-generated content, the use of AI in disinformation, and methods for authenticity verification.

The field is dominated by technically oriented studies, typically experimental and centred on detecting synthetic text, images, and video, most frequently on Twitter/X. Significant gaps remain regarding children's and adolescents' exposure, platform diversity across TikTok, Snapchat, and Roblox, and real-world user behaviour. Legal analyses highlight issues such as accountability, copyright, GDPR, and the implications of the EU AI Act for platform governance.

AI is used both for legitimate purposes and harmful activities, including fraud, extortion, political influence, and non-consensual pornography. Research consistently shows that users struggle to distinguish genuine from synthetic content, reinforcing risks related to misinformation, societal trust, and journalistic credibility. The policy debate focuses on content labelling, responsibility allocation, and the challenge of regulating AI without inhibiting innovation.

Overall, the review identifies a need for more integrated, interdisciplinary research, especially on youth vulnerability, Nordic media contexts, real-world detection performance, and the practical impact of emerging regulatory frameworks on digital ecosystems.

Keywords: Generative AI, Disinformation, Authenticity verification, Social media

1 Inledning

Denna rapport presenterar en systematisk litteraturöversikt av aktuell forskning om generativ artificiell intelligens, med fokus på det medieinnehåll den kan skapa används som innehåll i sociala medier. Fokus ligger särskilt på policyrelevanta frågor kopplade till barn och ungas medievardag, desinformation, journalistik och innehållsverifiering.

Översikten omfattar både akademisk litteratur och för uppdraget relevant så kallad *grå litteratur*[1] publicerad mellan den 30 juni 2022 och den 30 juni 2025. Syftet är att kartlägga denna tidsperiods tematiska forskningsfält, identifiera kunskapsluckor och därigenom stödja Mediemyndighetens beslutsfattare i att orientera sitt fortsatta arbete under 2026 och framåt.

Under de senaste åren har AI-genererat innehåll på sociala medier fått ökad uppmärksamhet. Även om teknologin i sig inte är ny har användningen av generativ AI för att skapa innehåll i sociala medier för desinformation, och brottslighet ökat markant sedan 2019 [2]. Särskilt betydelsefullt har varit lanseringen av de nu väl kända stora språkmodellerna som till exempel. ChatGPT[3] eller Claude[4] för text, och även andra plattformar som gör det möjligt att snabbt, enkelt och ibland kostnadsfritt generera innehåll, exempelvis DALL-E[5], Stable Diffusion[6] eller Midjourney[7] för stillbilder eller Gen-2[8] och SORA[9] för video, bara för att nämna några exempel. Utvecklingen mot lättillgängliga modeller och tjänster har inneburit en demokratisering av AI-teknologi vilket är positivt men har samtidigt blottlagt en rad etiska, juridiska och sociala utmaningar, risker och problem av särskild betydelse vilket behöver tas ställning till.

Debatten och forskningen har i hög grad fokuserat på risker kopplade till kontroll, ansvar och spridning av felaktigt eller skadligt AI-genererat innehåll. Nuvarande regulatoriska ramverk, såsom till exempel. EU AI Act[10], Storbritanniens Online Safety Act[11] samt USA:s Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI[12], samt överenstämmelse mellan privata aktörer som Munich Accord[13], betonar i första hand riskbedömning och riskanalys av generativ AI i kommersiella, politiska och sociala sammanhang.

Denna rapport syftar till att sammanställa, analysera och syntetisera den forskningsbaserade kunskap som finns tillgänglig i dagsläget, med särskild inriktning på frågor där policyaktörer såsom Mediemyndigheten behöver ett uppdaterat kunskapsunderlag. Tematiskt ansluter detta arbete till tidigare FOI-projekt inom samhällssäkerhet och informationspåverkan[14].

1.1 Begrepp

Denna rapport behandlar ett forskningsområde som är föremål för intensiv akademisk och samhällelig debatt. För att skapa tydlighet redovisas här hur denna rapport väljer att se på de centrala begreppen.

---

[1] För definition av grå litteratur se 2.1 denna rapport.
[2] Sumsub, *Identity Fraud Report 2024–2025*, 2025.
[3] https://chatgpt.com/ [hämtad 2025-11-10]
[4] https://claude.com/product/overview [hämtad 2025-11-10]
[5] https://chat.ai-pro.org/chat/dall-e/ [hämtad 2025-11-10]
[6] https://stability.ai/ [hämtad 2025-11-10]
[7] https://www.midjourney.com/home [hämtad 2025-11-10]
[88] https://runwayml.com/ [hämtad 2025-11-10]
[9] https://openai.com/sora/ [hämtad 2025-11-10]
[10] EU, 'EU AI Act', EU AI Act, 2024.
[11] GOV.UK, 'Online Safety Act: Explainer', Guidance Online Safety Act: Explainer, 2025.
[12] The White House, 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence', *The White House*, 30 October 2023.
[13] Security Conference, 'Accord - Munich Security Conference', A Tech Accord to Combat Deceptive Use of AI in 2024 Elections, 2024, https://securityconference.org/en/aielectionsaccord/accord/.
[14] Ola Svenonius, *Varning – desinformation! – Allmänhetens syn på psykologiskt försvar*, R R--5264—SE. Stockholm: FOI, 2022.

- Generativ artificiell intelligens (GenAI): AI-system med kapacitet att skapa nytt innehåll i form av text, bild, ljud eller video, vanligtvis genom moderna språkmodeller, bildgeneratorer eller andra transformer- och neuronnätsbaserade tjänster (exempelvis GPT-4, DALL·E).
- AI-genererat innehåll: Material framställt helt eller delvis av AI, ofta med en autentisk framtoning som försvårar verifiering och mottagarens källkritiska bedömning.
- AI-manipulerat innehåll: Befintligt material som ändrats med hjälp av AI, exempelvis genom röstsyntes, bild- och videoredigering ofta i syfte att vilseleda eller förstärka budskap.
- General Purpose AI (GPAI): En kategori definierad i EU:s AI Act (2024), avseende AI-system med bred funktionalitet och mångsidig tillämpbarhet. GPAI medför särskilda regulatoriska utmaningar.
- AI-modeller: De datamodeller som ligger till grund för generering, manipulation eller analys av innehåll, och som därmed är centrala för att förstå kapacitet och begränsningar.
- Deepfake: Teknik som använder generativ AI för att syntetiskt skapa stillbilder-, ljud- eller videomaterial med hög realism, ofta i syfte att imitera verkliga personer.

1.2 Forskningsfrågor

Syftet med denna systematiska litteraturöversikt är att kartlägga och analysera forskning om AI-genererat innehåll på sociala medier, med fokus på desinformation, autenticitetsverifiering och rättsliga implikationer. Följande forskningsfrågor har väglett arbetet:

1. Hur karakteriseras AI-genererat innehåll, inklusive sådant som används för desinformation, på sociala medieplattformar inom OECD-området i peer-review-granskad forskning och relevant grå litteratur publicerad 2022 till juni 2025?
2. Vilka teknologiska (till exempel. maskininlärning, bild- och ljudanalys) respektive samhällsvetenskapliga (till exempel. diskursanalys, policyanalys) metoder har empiriskt utvärderats för att identifiera, verifiera och moderera AI-genererat innehåll i digitala miljöer?
3. Vilka rättsliga ramar och etiska diskussioner framträder i litteraturen kring plattformsansvar, innehållsreglering, barns rättigheter och journalistisk trovärdighet i relation till AI-genererat innehåll?

Frågorna är utformade för att möjliggöra en tvärvetenskaplig analys som integrerar teknologiska, juridiska och sociala perspektiv, och som identifierar kunskapsluckor av relevans för Mediemyndighetens framtida arbete.

2 Metod

Det metodologiska arbetet är inspirerat av två centrala metodreferenser: PRISMA 2020 [15], samt Tranfield m.fl., (2003). Den bygger även vidare på FOI:s tidigare systematiska litteraturöversikter[16]. PRISMA 2020 tillför en internationellt erkänd ram för transparens och rapporteringsstandard, där särskild vikt läggs vid dokumentation av sökstrategier och urvalsprocessen i form av flödesscheman. Tranfield m.fl., (2003) kompletterar genom att introducera en samhällsvetenskaplig ansats till systematiska översikter, med fokus på tematisering, kodning och syntes tvärvetenskapligt. Tillsammans ger dessa två den vägledande ram som studien relaterar till och anpassar efter sin egen kontext. Resultat, urval och kodning hanterades via analysmjukvaran Covidence[17], referenshantering skedde i Zotero[18]. För att säkra träffsäkerhet och bredd i källmaterialet kombinerades flera databaser (se avsnitt 2.2).

Metoden har tillämpats på så sätt att funna artiklar först sorterades in i fördefinierade huvudkategorier, härledda ur studiens forskningsfrågor, såsom till exempel AI-genererat innehåll, desinformation och autenticitetsverifiering. Därefter genomfördes en induktiv analys, där återkommande mönster, nya underkategorier och tematiska samband identifierades i materialet utan att dessa var förutbestämda. På detta sätt möjliggjordes både en systematisk strukturering av den publicerade forskningslitteraturen och en öppenhet för att fånga upp framväxande forskningsfält och oväntade resultat som inte var förutspådda vid arbetets inledning.

Denna studie har inte utvecklat någon egen definition av begreppet *sociala medier*, utan utgår i stället från de definitioner som förekommer i den vetenskapliga litteratur som har inkluderats i översikten. Detta val är metodologiskt motiverat av översiktens syfte att spegla och syntetisera rådande begreppsanvändning i fältet. Vilka specifika sociala medieplattformar som förekommer mest frekvent i materialet redovisas och diskuteras närmare i avsnitt 4.1 *Dominanta teman.*

2.1 Inklusions- och exklusionskriterier

För att säkerställa både relevans och kvalitet i urvalet av källmaterial har studien skapat och tillämpat tydligt definierade inklusions- och exklusionskriterier för både akademiskt forskningsmaterial och för grå litteratur. Till inklusionskriterierna för det studien kallar akademiskt forskningsmaterial hör att granskat material ska vara publicerat inom det aktuella tidsspannet och vara vetenskapligt granskade artiklar publicerade i *peer-reviewed* vetenskapliga tidskrifter, samt även konferensartiklar förutsatt att fulltext varit tillgänglig.

Exklusionskriterierna för akademiskt forskningsmaterial har varit att studien uteslutit publikationer med otillräcklig vetenskaplig eller analytisk tyngd och/eller som på annat sätt inte bidragit till en översikt över forskningsläget. Detta är till exempel artiklar som inte innehåller empiriska resultat och/eller saknar teoretisk relevans i relation till studiens forskningsfrågor. Publikationer som inte genomgått peer-review-process, eller som härrör från känt tveksamma utgivare[19], har också exklude-

---

[15] Matthew J. Page m.fl., *The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews*, British Medical Journal, 2021.

[16] Magdalena Granåsen m.fl., *Tvärsektoriell Krishantering: Värdering Av Förmåga Och Modellering Av System En Systematisk Litteraturöversikt*, R FOI-R--5022—SE. Stockholm: FOI, 2021; Per-Erik Nilsson et al., *Våldsbejakande extremism och digitala medier: En forskningsöversikt*, R FOI-R--5500—SE. Stockholm: FOI, 2024.

[17] https://www.covidence.org/ [hämtad 2025-12-11]

[18] https://www.zotero.org/ [hämtad 2025-12-11]

[19] För att verifiera konferenser, tidskrifter och artiklar, använde vi flera källor. För konferenser var processen den följande:

1. Finns det en review process för inlämnade artiklar och är tiden för att utföra den processen rimlig (minst en månad)?
2. Finns det information om organisatörer och deras kontakter – vilken typ av e-postadress har dem? Återkommer samma kontakter för flera konferenser med olika teman/fokus?
3. Google sökning på konferensens titel och recensioner

För tidskrifter/artiklar:

1. Finns det peer-review process och för en viss artikel – finns det information om hur lång tid peer-review processesen tog
2. I Scimago databas (Scimago Journal & Country Rank) - är en viss tidskrift rank inom Q1-Q2? Om tidskriften tillhör Q3 och Q4, är den inte med i denna studie.

rats. Vidare har icke-granskade källor såsom blogginlägg, debattartiklar eller annan populärvetenskaplig kommunikation uteslutits från den akademiska kategorin.

Studien har även i viss mån beaktat vad som kallas för ”grå litteratur”. Det definieras i detta memo som material som produceras och distribueras utanför traditionella, akademiska publiceringskanaler. Grå litteratur omfattar dokument som till exempel, men inte uteslutande, tekniska rapporter, policydokument, ”white papers”, strategidokument, förhandsutgåvor, ”preprints”, myndighetsrapporter, tekniska översikter, statliga utredningar, arbetsmaterial och rapporter från exempelvis myndigheter, forskningsinstitut, universitet, ideella organisationer och NGO:er. Grå litteratur kännetecknas av att den inte har gått igenom den formella akademiska peer-review processen inom ramen för etablerade akademiska tidskrifter. De kan ha gått genom annan peer-review process. De kan hålla hög kvalitet och bidra med relevant kunskap, särskilt inom tillämpade och policyinriktade forskningsområden [20]. Till exempel. kan sägas att publikationer från FOI är att betrakta som grå litteratur. I denna studie har grå litteratur inkluderats under förutsättning att källan är tydligt identifierad och innehållet relevant i förhållande till studiens forskningsfrågor. Särskild vikt har lagts vid grå litteratur som har metodologisk transparens, empiriskt underlag och/eller ger analytisk fördjupning till forskningsfrågorna. Grå litteratur har exkluderats när materialet saknar spårbarhet, inte redovisar upphov eller utgivare, eller då det rör sig om icke-analytiska texter såsom blogginlägg, opinionsmaterial eller icke-granskad populärvetenskap.

För båda kategorierna har en geografisk avgränsning tillämpats i syfte att säkerställa relevans för Mediemyndigheten. Till övervägande del har studien fokuserat på publikationer med övervägande ursprung inom och med fokus på länder inom Organisationen för ekonomiskt samarbete och utveckling (OECD).

I praktiken har så gott som uteslutande engelskspråkiga publikationer beaktats, vilket är motiverat utifrån att den systematiska litteraturstudien ska ge en bred bild av forskningsläget. Ur ett tematiskt perspektiv har litteraturen valts ut baserat på dess relevans för minst ett av följande tre delområden: i) AI-genererat innehåll på sociala medier i allmänhet, ii) AI-genererat innehåll i sociala medier, för desinformation iii) metoder för autenticitetsverifiering i digitala miljöer, i relation till AI i sociala medier. Tidsperioden för all inkluderad litteratur sträcker sig som tidigare sagts från juni 2022 till juni 2025.

2.2 Datainsamling

Datainsamling skedde genom sökningar i databaserna Web of Science, IEEE Xplore, Scopus[21], ArXiv och Google Scholar, samt den juridiska databasen HeinOnline för rättsvetenskaplig forskning, då framför allt för att finna publicerad forskning om relevanta rättsakter från EU. Att kombinera olika databaser har visat sig ge optimala resultat för systematiska litteraturöversikter [22]. Referenser som inte står att finna i en vetenskaplig databas, kanske finns i en annan. Därför är det god metod att kombinera olika databaser beroende på forskningens tema. En kombination av Web of Science och Google Scholar ses som den mest heltäckande kombinationen[23]. Däremot har Google Scholars söksystem som enskilt verktyg vissa brister så som till exempel otydliga sätt att presentera resultat samt

3. Finns tidskriften i Bealls lista (Beall's List – of Potential Predatory Journals and Publishers) och andra databaser med listade bedrägliga publikationer (Predatory Journals - Journals)?

[20] GreyNet, ‘Grey Literature -’, Greynet International, 2025, https://www.greynet.org/home/aboutgreynet.html [hämtad 2025-12-11]; SFU Library, ‘Grey Literature: What It Is & How to Find It’, 2025.

[21] IEEE Xplore är indexerad av Scopus, men det finns oklarheter om alla IEEE Xplore tidskrifter (specifikt de nyare tidskrifter, från 2020, se Literature research: Are IEEE and ACM covered in Scopus? - Academia Stack Exchange) kan även hittas på Scopus. Covidence tar bort dubbletter automatiskt, så i valet mellan att få många dubbletter, jämfört med risken att missa viktiga artiklar, väljer vi ändå att kombinera Scopus och IEEE Xplore.

[22] Wichor M. Bramer m.fl., ‘Optimal Database Combinations for Literature Searches in Systematic Reviews: A Prospective Exploratory Study’, *Systematic Reviews* 6, no. 1 (2017): 245.

[23] Wichor M. Bramer m.fl., ‘Optimal Database Combinations for Literature Searches in Systematic Reviews: A Prospective Exploratory Study’, Systematic Reviews 6, no. 1 (2017): 245.

alltför stor variation i fråga om bredd och noggrannhet på resultaten [24]. I denna studie har vi kombinerat flera databaser för att undvika snedfördelning beroende på möjlig överregistrering av artiklar från större institutioner, samt för att täcka tidskrifter och artiklar som är inte registrerade i varje databas.

Sökningen skedde genom tre steg enligt våra forskningsfrågor:

Steg 1. Generativ AI som innehåll i allmänhet på sociala medier

1.1 Samla in alla artiklar som nämner AI-genererat innehåll på sociala medier

Steg 2. Generativ AI innehåll för desinformation på sociala medier

2.1 Samla in alla artiklar som nämner AI-genererad desinformation på sociala medier

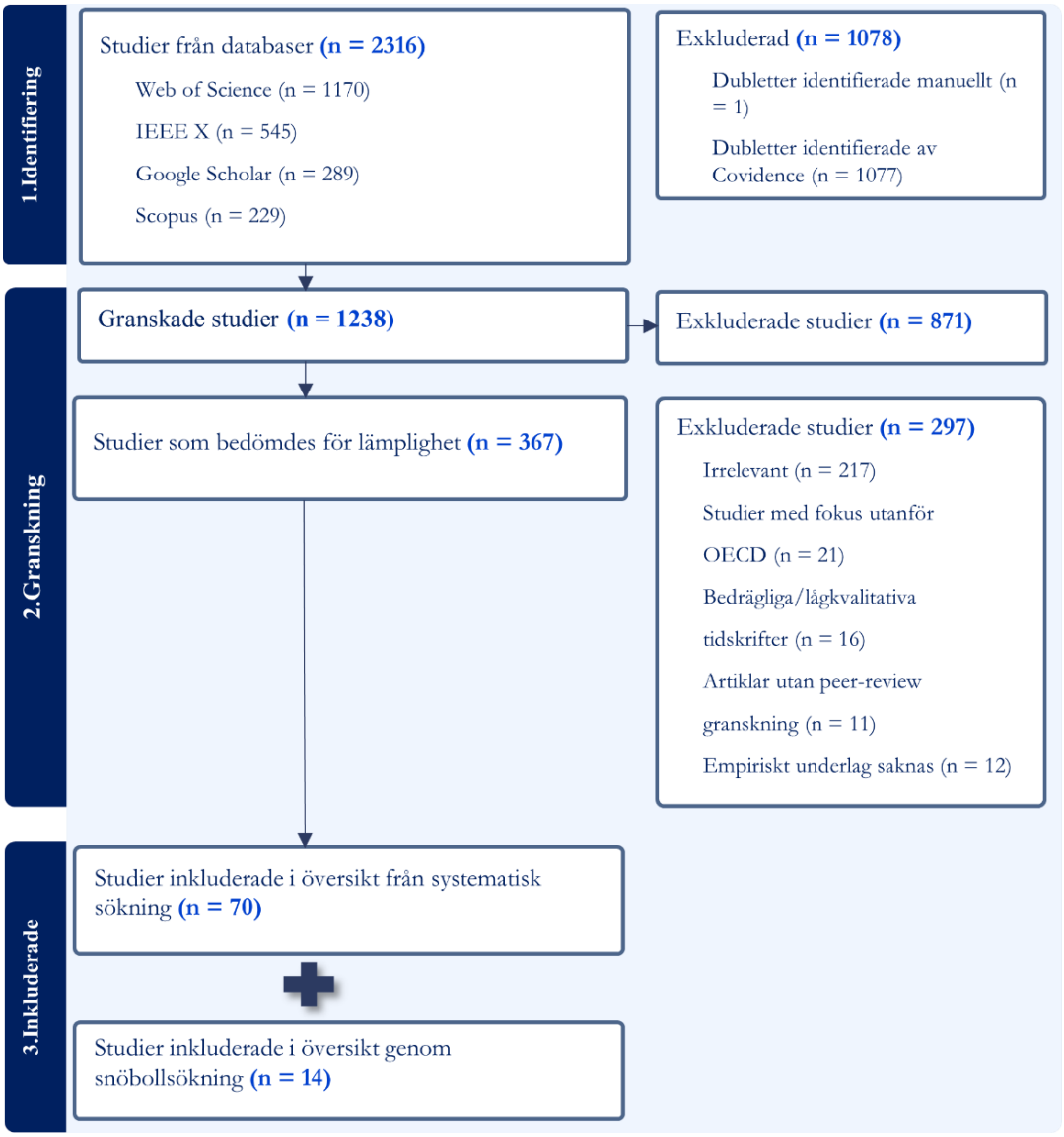
Steg 3. Äkthetsverifiering av AI genererat innehåll på sociala medier

3.1 Samla in alla artiklar som nämner äkthetsverifiering av AI genererat innehåll på sociala medier.

För alla databaser använde vi samma sökord (med lite variation beroende på regler för söksträngsyntax, se Appendix 1). När det gällde Google Scholar följde vi en rekommendation[25] att extrahera de första 50 artiklarna i varje sökning. Den processen resulterade i sammanlagt 2316 st. träffar (fig. 1).

Efter datainsamling granskades bearbetades artiklarna med hjälp av den nämnda Covidence plattformen och enligt PRISMA systemet (fig 2). Granskningen gick i två omgångar. Den första fokuserade på att upptäcka dubbletter (1078 st.) och enligt titel och abstract filtrera ut irrelevanta artiklar (vanligtvis de som inte passade tematiskt eller inte var skrivna på engelska, eller var renodlade litteratursammanställningar). Den andra omgången involverade läsning av hela texten mer på djupet enligt de förut nämnda exklusionskriterierna för att filtrera ut de artiklar som inte hade metodologisk eller teoretisk relevans till studiens forskningsfrågor.

[24] Isomi M. Miake-Lye m.fl., 'Using Google to Search for Evidence: How Much Is Enough? One Center's Experience', *Systematic Reviews* 14, no. 1 (2025): 92.
[25] Bramer m.fl., 'Optimal Database Combinations for Literature Searches in Systematic Reviews'.



Figur 1. PRISMA * ingen fulltext (n = 1), avhandling/bok (n = 1), studenttidskrift (n = 1), dubblett (n = 3), review (n = 9), utanför tidsram (n = 2), video transcript (n = 2)

Under granskningen, kodade vi artiklar med särskild fokus på forskningens målgrupper och teman, där barn och unga samt journalistisk trovärdighet var av speciellt intresse. Därför, upprepade vi både Steg 1 och Steg 2 med specificering av söksträng med orden ”children and youth”, och ”journalism”. De insamlade artiklarna granskades sedan igen utifrån samma logik.

Vi utförde en separat sökning för grå litteratur eftersom inte alla relevanta grå publiceringar inkluderas i de vetenskapliga databaserna. Vid sökningen av grå litteratur kombinerades att gå igenom rapporter från utvalda institutioner och myndigheter som författarna bedömt som relevanta till exempel. Rand[26], Nato Strategic Communications Centre of Excellence[27] Hybrid Centre of Excellence[28],

[26] https://www.rand.org/ [hämtad 2025-12-11]
[27] https://stratcomcoe.org/[hämtad 2025-12-11]
[28] https://www.hybridcoe.fi/ [hämtad 2025-12-11]

Stanford Internet Observatory, FOI och snöbollsurval[29] av grå litteratur som är nämnd eller återkommer i de vetenskapliga artiklarna från förut nämnda databaser. Den sökningen resulterade i 14 ytterligare resultat. Sammanlagt fick vi 70 st. artiklar från den systematiska sökningen, och ytterligare 14 st. till inom kategorin grå litteratur genom snöbollsmetod. Dokumentation av alla sökströngar finns bifogad i appendix 1.

2.2.1 Datainsamling av rättsvetenskaplig forskning

I denna studie har särskild hänsyn tagits till rättsvetenskaplig litteratur som fått vara en extra fördjupning inom de tre områdena, detta i syfte att synliggöra de rättsliga implikationer som uppstår när AI-genererat innehåll blir till innehåll i sociala medier. Vi gjorde detta med anledning av den särskilda betydelse vissa av EU:s rättsakter har för Mediemyndigheten. Det rättsliga källmaterialet identifierades huvudsakligen genom HeinOnline[30] samt genom tematiska sökningar i övriga databaser, med särskilt fokus på OECD-länders rättsordningar och EU:s rättsakter relaterade till ämnet.

Analysen av det rättsvetenskapliga området har tagit sin utgångspunkt i funna artiklar som behandlar frågor om ansvarsfördelning och upphovsrätt i relation till AI-genererat innehåll, samt gränsdragningar mellan yttrandefrihet och plattformsansvar. Vidare har studien uppmärksammat litteratur som diskuterar reglering av desinformation, särskilt i valpåverkande sammanhang, samt i sammanhang av barns rätt till skydd från skadligt AI-innehåll enligt exempelvis Barnkonventionen och GDPR.

De rättsvetenskapliga texterna har granskats med tematiserande läsning. Särskild vikt har lagts vid att identifiera i vilken utsträckning begrepp som ansvar, transparens och integritet omförhandlas i ljuset av generativ AI och dess tillämpningar på sociala medier. De rättsvetenskapliga analyserna granskades och inkluderas i studien, utefter studiens inklusion- och exklusionskriterier.

2.3 Urval och kodning

Efter att det initiala källmaterialet samlats in genomfördes en granskning. Granskningen strukturerades i enlighet med PRISMA 2020:s riktlinjer, däribland ett flödesschema som vägledde arbetet med att filtrera bort dubbletter, bedöma titel och abstract, samt genomföra fulltextgranskningar (fig. 2). Urvalet grundades på de inklusions- och exklusionskriterier som tidigare redogjorts för.

Det material som bedömdes som relevant inkluderades i ett kodningssystem som utvecklades iterativt. Den initiala kodningen var deduktiv, där följande fördefinierade teman härledda ur forskningsfrågorna låg till grund för kategorisering: AI-genererat innehåll, desinformation, samt autenticitetsverifiering. Samtidigt tilläts kodsystemet utvecklas i takt med analysen, vilket gjorde det möjligt att fånga upp framväxande begrepp, forskningsluckor och metodvariationer och belyste teman som inte var initialt fastlagda i studien (se Appendix för en fullständig lista).

Kodningen inkluderade flera dimensioner. En central aspekt var tekniktyp, det vill säga om innehållet behandlade AI-generering av text, bild, ljud, video eller multimodala uttryck på sociala medier. Vidare noterades vilken plattform som studien avsåg (till exempel. TikTok, Reddit, Instagram), samt vilken målgrupp som stod i fokus. Särskilt uppmärksammades förekomsten av barn och unga som explicita forskningsobjekt eller forskningssubjekt i materialet, samt artiklar som tog upp journalistik eller mediernas roll. Dessutom kategoriserades artiklar efter metodansats (till exempel. fallstudie, experiment, kvantitativ analys, policyöversikt), och efter geografisk kontext med fokus på OECD-länder. Innehållsanalysen inriktade sig också på identifiering av rättsliga och regulatoriska teman såsom ansvarsfördelning, transparens, dataskydd, yttrandefrihet och upphovsrätt i relation till genAI-innehåll på sociala medier.

[29] Metoden är att vetenskapliga artiklar identifieras genom rekommendationer från redan identifierade artiklar och deras referenslistor i relevant forskningslitteratur.
[30] https://home.heinonline.org/ [hämtad 2025-12-11]

En särskild kodkategori av artiklar avsåg former av AI-genererat innehåll kopplade till intention eller syfte, exempelvis om AI-materialet var designat för desinformation, marknadsföring, politisk påverkan eller andra ändamål. Här identifierades även återkommande begrepp som "deepfake", "chatbots" och "AI-genererade personas". Slutligen noterades om artiklar behandlade frågor om verifiering eller detektion. Kodningen utfördes i analysmjukvaran Covidence, där samtliga beslut dokumenterades för att säkerställa spårbarhet. Kodsystemet vidareutvecklades abduktivt, och justerades i takt med att nya mönster identifierades i litteraturen. Denna kombination av struktur, öppenhet och iterativt arbete möjliggjorde en syntes som är anpassad till det snabbt föränderliga forskningsfält som generativ AI på sociala medier är.

2.4 Dokumentation

Arbetet har följt PRISMA 2020:s principer för rapportering av systematiska översikter vilket innebär att varje moment från datainsamling till urval och kodning har strukturerats och registrerats med spårbarhet i åtanke[31]. Urvals- och granskningsprocessen dokumenterades löpande i analysverktyget Covidence, som användes för att identifiera dubbletter, klassificera material enligt inklusions- och exklusionskriterier samt strukturera bedömningsprocessen i två steg: först efter titel och abstract, och därefter i fulltext. Den resulterande översikten över inkludering och exkludering visualiseras i enlighet med PRISMA:s flödesschema (se figur 2), vilket möjliggör överskådlig granskning av beslutslogiken bakom urvalet. Sökstrategierna, inklusive söksträngar och urval av databaser redovisas i appendix 1. På så sätt följer studien PRISMAs riktlinjer och stärker möjligheten till framtida uppdateringar.

[31] Matthew J. Page m.fl., The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews, British Medical Journal, 2021.

3 Resultat och Analys

Mot denna bakgrund presenterar följande kapitel de resultat som framkommit genom analysen av det inkluderade materialet. Med stöd i den tematiska syntesmodell som valts, redovisas centrala mönster, begrepp och teman som speglar forskningsfältets nuvarande inriktning och utmaningar.

3.1 Tematisk syntes

I detta avsnitt genomförs en tematisk syntes, vilket innebär att resultaten från de inkluderade studierna integreras och tolkas för att identifiera återkommande mönster och övergripande teman som belyser rapportens forskningsfrågor.

Analysen av det inkluderade materialet genomfördes med utgångspunkt i en tematisk syntesmodell inspirerad av Tranfield, Denyer och Smart (2003), med målet att identifiera återkommande mönster, begrepp och problemområden i forskningsfältet. Denna syntesmodell möjliggjorde en systematisering av ett mångfacetterat material där teknik, juridik, etik och samhälleliga aspekter ofta är tätt sammanvävda. Genom att kombinera deduktiv och induktiv kodning med utgångspunkt i studiens forskningsfrågor kunde både etablerade forskningsområden och framväxande teman identifieras.

Fyra övergripande teman utkristalliserades ur analysen, vilka också strukturerar presentationen av resultaten i de följande avsnitten:

1. Tekniker för att identifiera och verifiera AI-genererat innehåll.
2. Risker och skyddsfaktorer för barn och unga.
3. Journalistisk trovärdighet och medielogik i en AI-kontext.
4. Rättsliga implikationer av AI-genererat innehåll.

Dessa teman är inte ömsesidigt uteslutande utan överlappar ofta, särskilt i frågor som rör ansvar, etik och policy. Deras funktion är inte att kategorisera forskningen strikt, utan snarare att möjliggöra en tvärdisciplinär syntes där olika; perspektiv tekniska, rättsliga, sociala kan belysas parallellt.

Avslutningsvis bör nämnas att förekomsten av studier som i hög grad fokuserar på Twitter/X som forskningsplattform, i kombination med avsaknad av forskning på plattformar populära bland barn och unga (till exempel. Roblox, Snapchat), indikerar en möjlig snedvridning i forskningsfältets empiri. Denna observation ligger till grund för diskussionen om forskningsluckor i kapitel 4.1.

I kommande avsnitt (3.2 - 3.7) fördjupas respektive tema, med särskild betoning på deras inbördes relationer och relevans för en policymiljö i förändring.

3.2 Tematisk översikt av den insamlade litteraturen

Här presenteras en tematisk översikt, där det insamlade materialet systematiskt kartläggs för att synliggöra hur olika teman, plattformar och forskningsinriktningar fördelar sig och därigenom tydliggör fältets struktur, tyngdpunkter och kanske dess luckor.

Den samlade forskningen om AI-genererat innehåll på sociala medier präglas av ett fokus på risker, särskilt risker relaterade till kontroll, ansvar och spridning av felaktigt, vilseledande eller skadligt material. Studierna tenderar att vara reaktiva till teknikens snabba framväxt, med särskild uppmärksamhet riktad mot fenomen som desinformation, deepfakes och autenticitetsproblematik. Tabell 2, visar olika typer av områden där AI används för innehåll på sociala medier. Tabellen visar att de flesta, men inte alla områden är olika former av skadeområden.

Samtidigt förekommer även ett mindre men tydligt delområde av studier framför allt inom marknadsföring, kommunikation och affärsutveckling som betonar teknikens möjligheter, ex-

empelvis genom kostnadseffektiv innehållsproduktion eller skräddarsydd kundkommunikation [32]. Denna dualitet mellan risk och potential är framträdande i fältet och återkommer också i de regulatoriska diskussionerna (se 3.4), där balansen mellan innovation och skydd är ett återkommande tema.

En tydlig tendens i materialet är att majoriteten av studierna är författade av forskare knutna till amerikanska eller brittiska universitet, vilket i sig påverkar de kontexter, plattformar och normativa referensramar som analyseras (figur 3). Endast enstaka studier är producerade i nordiska forskningsmiljöer, tre från Finland och en från Sverige [33]. Det nordiska perspektivet är ett underbeforskat område och därmed med särskild relevans för svenska policyaktörer som till exempel. Mediemyndigheten. Även om vi insamlingsskedet valde att låta OECD-området dominera urvalet framstår det som klart att det finns en västerländsk problemformulering, där digitala plattformar som Twitter/X, YouTube och Instagram per se antas vara de mest centrala arenorna (se 3.3).

Metodologiskt präglas fältet av en stark dragning mot kvantitativa metoder, där experimentell forskningsdesign, innehållsanalys och maskininlärningsbaserade forskningstyper är vanligt förekommande (figur 4). Dessa tillämpas ofta på enstaka plattformar särskilt Twitter/X vilket gör att generaliserbarheten över flera plattformstyper är begränsad. Det kan bero på att Twitter/X erbjöd forskare öppen tillgång till data fram till februari 2023, vilket faller inom denna studies tidsspann.[34] Eftersom akademisk publicering har mycket långa ledtider så kan forskning baserad på material från Twitter/X fortfarande publiceras 2024 och 2025. De få storskaliga studier som identifierats bygger ofta på proprietära dataset eller syntetiskt genererade material snarare än på datainsamling från faktiska sociala interaktioner. Kvalitativa eller policyorienterade studier förekommer, men är i minoritet.

En central observation från denna analys är att det tvärvetenskapliga greppet ofta är bristfälligt realiserat i praktiken. Teknikfokuserade artiklar tenderar att sakna normativ eller juridisk förankring, medan samhällsvetenskapliga analyser ibland utelämnar teknologisk förståelse. Undantag finns, och dessa utgör viktiga bidrag till fältets utveckling. Ett sådant exempel är en studie av Kearney m.fl. (2025), som i ett explorativt angreppssätt introducerar ett nytt analytiskt ramverk (ASCENT) för att integrera analyser av AI-system, mänsklig kognition och sociala dynamiker i studiet av echo chambers på sociala medier[35] Här kan även nämnas filosofisk forskning som kan kombinera olika perspektiv på risker med AI, men på ganska abstrakt nivå [36].

Sammantaget visar forskningsfältets inriktning att det finns ett etablerat intresse för de problem och risker som AI-genererat innehåll på sociala medier medför, men att det fortfarande är ett underskott på samordnade forskningsinsatser som integrerar teknisk, juridisk och samhällelig analys. Denna splittring försvårar möjligheten att dra generaliserbara slutsatser kring frågor som till exempel ansvarsfördelning, innehållsmoderering och målgruppseffekter, frågor som är av särskild betydelse för policyaktörer.

---

[32] Jad Abi-Rafeh m.fl., 'Artificial Intelligence–Generated Social Media Content Creation and Management Strategies for Plastic Surgeons', *Aesthetic Surgery Journal* 44, no. 7 (2024): 769–78; Nestor Maslej et al., *AI Index Report 2025*, AI Index Steering Committee (Institute for Human-Centered AI, Stanford University), 2025.

[33] Yucong Lao m.fl., 'AI and Authenticity: Young People's Practices of Information Credibility Assessment of AI-Generated Video Content', *Journal of Information Science*, 2025; Eleonora Rosati, 'Infringing AI: Liability for AI-Generated Outputs under International, EU, and UK Copyright Law', *European Journal of Risk Regulation* 16, no. 2 (2025): 603–27.

[34] Twitter Dev, 'Starting February 9, we will no longer support free access to the Twitter API…' https://web.archive.org/web/20230213222854/https://twitter.com/TwitterDev/status/1621026986784337922 [hämtad 2025-12-11]

[35] Ashley Kearney, Nihal Poredi, Joseph A. Shelton, Seden Akcinaroglu, m.fl., "Echoes amplified: a study of AI-generated content and digital echo chambers," Proc. SPIE 13480, Disruptive Technologies in Information Sciences IX, 134800L (2025)

[36] Bartlomiej Chomanski och Lode Lauwaert, 'Automated Propaganda: Labeling AI -Generated Political Content Should Not Be Required by Law', *Journal of Applied Philosophy* 42, no. 3 (2025): 994–1015; Sarah A. Fisher, 'Something AI Should Tell You - The Case for Labelling Synthetic Content', *Journal of Applied Philosophy* 42, no. 1 (2025): 272–86; Sarah A. Fisher m.fl., 'Moderating Synthetic Content: The Challenge of Generative AI', *Philosophy & Technology* 37, no. 4 (2024): 133,; Christian Tarsney, 'Deception and Manipulation in Generative AI', *Philosophical Studies* 182, no. 7 (2025): 1865–87.]

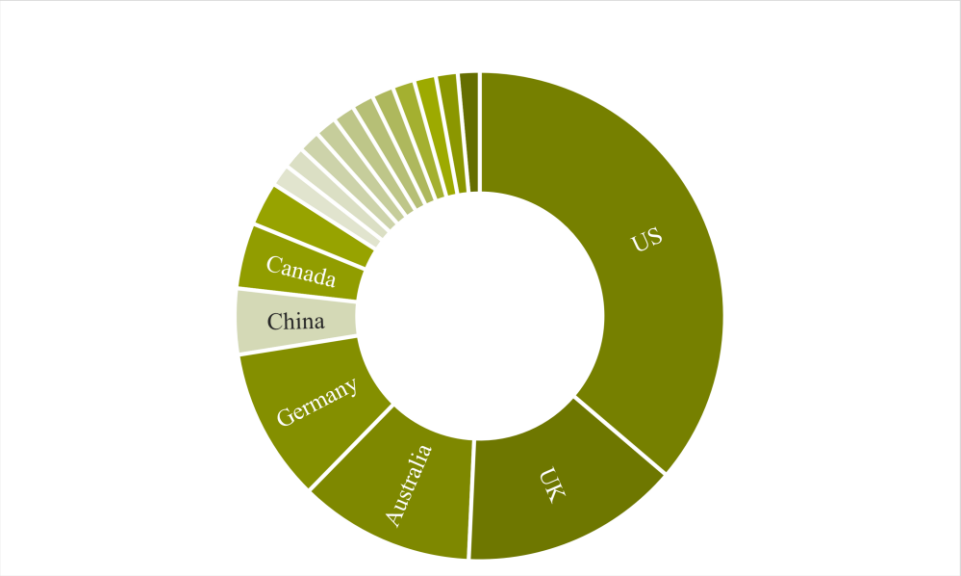


Fig 2. Författares geografiska affiliering.

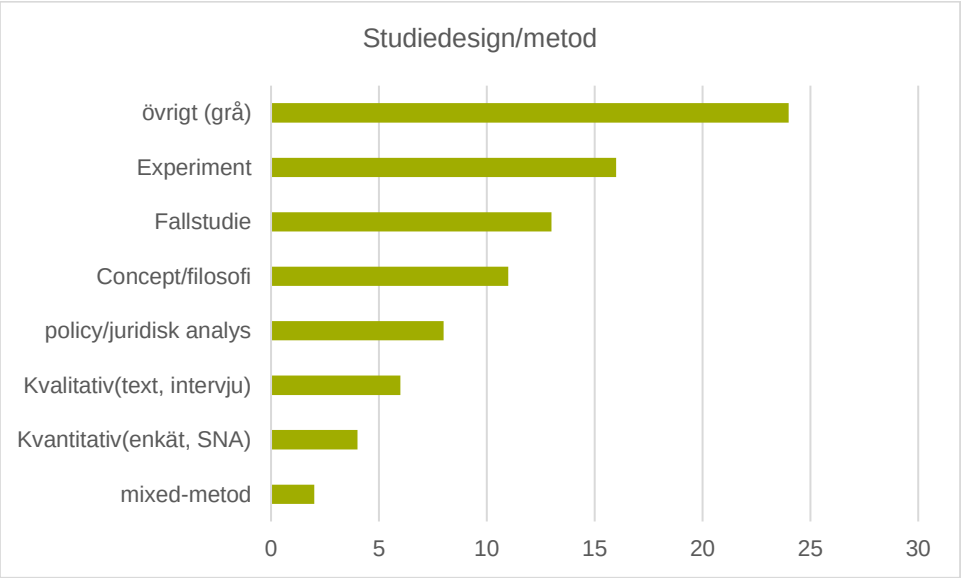


Fig. 3. Förekomster av metoder[37]

3.3 Plattformar och innehållstyper

Mer än hälften (65 %) av EU:s totalbefolkning var aktiva på sociala medier år 2024 [38]. Det är dock de yngre åldersgrupperna som uppvisar högst aktivitetsnivå 88 % av dem mellan 16 och 29 år var aktiva på sociala medier samma år. I Sverige är de mest frekvent använda plattformarna dagligen: Facebook, Instagram, YouTube, Snapchat och TikTok [39]. Användningen skiljer sig dock åt mellan åldersgrupper, och exempelvis barn i mellanstadieålder använder i hög grad Roblox för social interaktion online.

---

[37] SNA – social network analysis; övrigt (grå) - grålitteratur och peer-review artiklar som gör en viss typ av litteraturöversikt eller sammanfattningar av workshops/konferenser, samt olika källor som statistik, dokument och reglering som är inte empiriska; inga överlappar

[38] Eurostat, *Individuals - Internet Activities* (2025).

[39] Internetstiftelsen, *Kapitel 3. Sociala Medier*, Svenskarna Och Internet, 2025.

Vetenskapliga artiklar i denna studies urval tenderar att fokusera på Twitter/X, särskilt i storskaliga studier där syftet är att mäta förekomst av generativ AI och utveckla detektionsmodeller. Det finns en diskrepans mellan de sociala medier som faktiskt används av barn och vuxna, och de som förekommer i publicerad forskning om AI-genererat innehåll på sociala medier.

Tabell 1 visar fördelningen av studier med primärt fokus på specifika sociala medieplattformar. Andra plattformar, som till exempel. Snapchat, Discord, Roblox, Twitch och Patreon[40], nämns visserligen i artiklarna, ofta som bakgrund eller exempel, men analyseras sällan systematiskt utifrån empiriskt datamaterial.

Sociala medier	**# Artiklar**
Twitter/X	24
Facebook	15
Instagram	15
TikTok	11
YouTube	7
Reddit	5
LinkedIn	2
Pixiv	1
GitHub	1

Tabell 1 Fördelning av vetenskapliga artiklar och sociala mediaplattformers data (utan grå litteratur). OBS studier kan överlappa när de använder flera datakällor

Innehållstyper

Baserat på de funna artiklarna kan AI-genererat innehåll i studierna delas in i två analytiska kategorier: (1) tekniktyp och (2) funktion eller syfte i kontexten av sociala medier (se tabell 2). Tekniktyperna omfattar text, bild, ljud, video och multimodala kombinationer. Den andra kategorin relaterar till det avsedda målet för innehållet, exempelvis politisk påverkan, kriminella syften, ekonomiska intressen eller sociala effekter, se tabell 2. Som beskrivs i inledningen har demokratiseringen av generativa AI-modeller möjliggjort bred tillgång till tekniker för att generera olika former av digitalt innehåll. AI-genererade bilder förekommer frekvent och kan inkludera ansikten, människor, djur, landskap och mer surrealistiska motiv exempelvis inom genren ”body horror” [41]. Porträttbilder är vanliga, antingen som helt syntetiska ansikten [42] eller som manipulerade foton med FaceSwap[43] eller med filter och AI-effekter [44]. Flera plattformar erbjuder tjänster som till exempel att generera ansiktsbilder, ofta med StyleGAN2-träning på till exempel. Flickr-Faces-HQ Dataset [45].

Textgenerering sker i regel med ChatGPT, och används bland annat för att planera kampanjer, skapa marknadsföringsinnehåll och skriva journalistik online [46]. I vissa fall har AI-genererad text analyserats som potentiell resurs i extremistiska syften[47]. Ljudgenerering förekommer i mindre utsträckning i det vetenskapliga materialet, men uppmärksammas i grå litteratur i samband med

[40] Patreon tas som ett exempel på sociala medier för att användare kan ha följare och individer som sponsrar verksamheten.
[41] Jennifer O’Meara och Cáit Murphy, ‘Aberrant AI Creations: Co-Creating Surrealist Body Horror Using the DALL-E Mini Text-to-Image Generator’, *Convergence* 29, no. 4 (2023): 1070–96.
[42] Tingxuan Wu m.fl, ‘MSM-BD: Multimodal Social Media Bot Detection Using Heterogeneous Information’, arXiv:2501.00204, preprint, arXiv, 31 December 2024.
[43] FaceSwap är en process där ansikte på en person bytes med en annan persons ansikte
[44] Marissa Spada, ‘History “for the Gram”: AI Beauty and the Vintage, Analogue, and “Throwback” Celebrity’, *Celebrity Studies*, ahead of print, Routledge, 2025.
[45] Jonas Ricker m.fl, ‘AI-Generated Faces in the Real World: A Large-Scale Case Study of Twitter Profile Images’, *The 27th International Symposium on Research in Attacks, Intrusions and Defenses*, ACM, 30 September 2024, 513–30.
[46] Jasper David Brüns and Martin Meißner, ‘Do You Create Your Content Yourself? Using Generative Artificial Intelligence for Social Media Content Creation Diminishes Perceived Brand Authenticity’, *Journal of Retailing and Consumer Services* 79 (July 2024): 103790.
[47] Stephane J. Baele m.fl., ‘Is AI-Generated Extremism Credible? Experimental Evidence from an Expert Survey’, *Terrorism and Political Violence* 37 (8): 1060–76, 2024.

musik, konst eller kriminella tillämpningar [48]. Multimodala innehåll, såsom video med ljud och bild med text, behandlas ofta i relation till begreppet deepfakes[49].

En inte oväsentlig andel AI-genererat innehåll är enligt det funna materialet pornografiskt eller NSFW (not safe for work), inklusive icke-samtyckt material med barn och kvinnor ofta med koppling till cybermobbning, utpressning och digitalt våld [50]. Medan det är svårt att estimera precis vilken andel av genAI-innehåll är NSFW med data som är öppen, flera källor nämner det, inklusive officiella lagar/dokument. AI-genererat innehåll förekommer även inom konst, reklam, opinionsbildning och påverkan, hälsoinformation och utbildning exempelvis i syfte att påverka attityder till vaccinering eller rökning [51], eller genAI som potentiell resurs att för att motverka hatpropaganda [52].

Innehållstyper	**Exempel på artiklar**
Desinformation	Bontcheva m.fl., 2024; Hajli et al 2022; Ienca 2023; Insikt Group 2024; Kearney m.fl. 2025; Lao m.fl. 2025; Matich m.fl. 2025; Minici m.fl. 2024; Puczyńska och Djenouri 2024; Rand Corporation 2022; Wei m.fl. 2022
Pornografi/våld	Chawki 2025; Kira 2024; Romero Moreno 2024
Extremism	Baele m.fl. 2024
Dating	Mink m.fl. 2022
Cybermobbning	Alexander 2025; Milosevic m.fl. 2023; Laczi och Póser 2024; Yu m.fl. 2025
Bedrägeri/utpressning	Rand Corporation 2022; Security Hero 2023; Sumsub 2025
Medicin	Ayers m.fl. 2023
Journalistik	Arguedas and Simon 2024; Bayer 2024; Borchardt m.fl. 2024; Cools och Diakopoulos 2024; Matich m.fl. 2025; Thomson m.fl. 2025;
Konst	Ashton och Patel 2024; Tang och Liu 2025; U.S. Copyright Office 2023.
Marknadsföring	Abi-Rafeh m.fl. 2024; Brüns och Meißner 2024
Satir/humör/meme	Scheirer 2024
(Para)sociala interaktioner	Andrejevic och Volčič 2025; Leaver och Srdarov 2025; Robb och Mann 2025
Sociala frågor/utbildning	Leung m.fl. 2025; Lv m.fl. 2025

Tabell 2. Målen med användning av genAI på sociala medier enligt de insamlade artiklarna.

AI-genererade ”personas”/profiler

AI-genererade profiler på sociala medier kan enligt det funna materialet delas in i två typer. Den första utgörs av falska profiler ofta kombinerande AI-genererade porträtt och text som används för desinformation, kommersiella syften eller bedrägerier. Dessa “bots” eller “chatagents” förekommer enligt artiklarna särskilt på Twitter/X och LinkedIn. Studier visar att sådana profiler kan kringgå plattformars automatiska

[48] Insikt Group, *Targets, Objectives, and Emerging Tactics of Political Deepfakes*, threat analysis (2024); Security Hero, *2023 State Of Deepfakes*. 2023; U.S. Copyright Office, *Part 1: Digital Replicas*, Copyright and Artificial Intelligence (2024).

[49] Sahar Tahmasebi m.fl., ‘Multimodal Misinformation Detection Using Large Vision-Language Models’, version 1, preprint, arXiv, 2024; Tingxuan Wu et al., *MSM-BD: Multimodal Social Media Bot Detection Using Heterogeneous Information*, arXiv, 2024.

[50] Beatriz Kira, ‘When Non-Consensual Intimate Deepfakes Go Viral: The Insufficiency of the UK Online Safety Act’, *Computer Law and Security Review* 54 (2024); Security Hero, *2023 State Of Deepfakes*; Laura G. E. Smith m.fl., ‘How and Why Psychologists Should Respond to the Harms Associated with Generative AI’, *Communications Psychology* 2, no. 1 (2024): 60.

[51] Janni Leung m.fl., ‘Generative Artificial Intelligence With Youth Codesign to Create Vaping Awareness Advertisements’, *JAMA Network Open* 8, no. 7 (2025).

[52] Chuanhui Wu m.fl., ‘Confront Hate with AI: How AI-Generated Counter Speech Helps against Hate Speech on Social Media?’, *Telematics and Informatics* 101 (2025).

detektionssystem [53]. Det finns även exempel på "catfishing" och "deep phishing", där AI-profiler skapar relationer med verkliga användare i bedrägligt syfte ibland med hjälp av ljud, bilder och textgenerering[54].

Den andra typen är enligt artiklarna öppet AI-genererade och syftar till interaktion snarare än vilseledning. Exempel inkluderar AI-companions som används för chatt och spel, särskilt på plattformar som till exempel. Character.ai[55] och Chai[56]. Dessa diskuteras inte i den vetenskapliga litteraturen, gissningsvis för att den vetenskapliga publicering inte hunnit till dessa ännu, men studier på dessa förekommer i grå litteratur särskilt när dessa diskuterar potentiell påverkan på unga [57]. Begreppet parasociala relationer [58] blir centralt här, då användare tenderar att utveckla upplevda relationer med AI-karaktärer, inklusive AI-influencers på Instagram och plattformar där användaren skapar AI-förhållanden (till exempel. Reddit's MyBoyfriendIsAI[59]).

3.4 Frågor kring regleringar

Flera lagförslag och regleringar kring generativ AI presenterades och trädde i kraft under den period som omfattas av denna studie. Exempelvis utgör EU:s AI Act (AIA)[60], Storbritanniens Online Safety Act[61] och USA:s Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI (2023)[62] några centrala referenspunkter i litteraturen. Detta avsnitt sammanfattar de mest publicerade rättsvetenskapliga frågorna kring AI-genererat innehåll i sociala medier, såsom de framkommer i det analyserade materialet. Ett återkommande tema i den rättsliga litteraturen är spänningsförhållandet mellan å ena sidan teknikens utvecklingstakt och å andra sidan lagstiftningens möjligheter att reagera i tid.

Copyright, yttrandefrihet och GDPR: vem, "äger" generativt AI-innehåll?

Generativa AI-modeller som producerar video, bild, ljud och text tränas på stora mängder data. En rättslig fråga som återkommer i materialet rör i vilken utsträckning dessa träningsdata, särskilt om de innehåller personlig information, faller under GDPR:s tillämpningsområde [63]. Den rättsvetenskapliga diskussionen är även relevant för områden som rör konst, journalistik och utbildning, där användning av upphovsrättsligt skyddat material för träning av modeller kan utgöra intrång i immateriella rättigheter[64]. Om så är fallet är en pågående diskussion i den rättsvetenskapliga debatten.

Ska generativ AI regleras och i så fall hur?

EU AI Act klassificerar General Purpose AI (GPAI) utifrån risknivå: oacceptabel, hög, begränsad eller minimal[65]. Endast system som anses utgöra oacceptabel risk förbjuds helt. För system med hög risk gäller särskilda krav på dokumentation och transparens, där ansvarsfördelningen är tydligt definierad: det är utvecklarna som bär det primära ansvaret, distributörer ett delat ansvar med utvecklarna och användare bär ett mer begränsat ansvar. Användare ska även informeras om att de

---

[53] Tingxuan Wu m.fl., MSM-BD: Multimodal Social Media Bot Detection Using Heterogeneous Information, arXiv, 2024.
[54] Jaron Mink et al., 'DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks', in Proceedings of the 31st USENIX Security Symposium 2022.
[55] https://character.ai/ [hämtad 2025-12-11]
[56] https://www.chai-research.com/ [hämtad 2025-12-11]
[57] Michael Robb och Supreet Mann, *Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions*. Common Sense Media, 2025.
[58] Jeongeun Park m.fl., 'AI vs. Human-Generated Content and Accounts on Instagram: User Preferences, Evaluations, and Ethical Considerations', *Technology in Society* 79 (December 2024): 102705, https://doi.org/10.1016/j.techsoc.2024.102705. [hämtad 2025-12-11]
[59] https://www.reddit.com/r/MyBoyfriendIsAI/ [hämtad 2025-12-11]
[60] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 [hämtad 2025-12-11]
[61] https://www.legislation.gov.uk/ukpga/2023/50/enacted [hämtad 2025-12-11]
[62] https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence [hämtad 2025-11-10]
[63] Felipe Romero Moreno, 'Generative AI and Deepfakes: a human rights approach to tackling harmful content', *International Review of Law, Computers & Technology*, 38(3), pp. 297–326 (2024).
[64] Beatriz Kira 'When Non-Consensual Intimate Deepfakes Go Viral: The Insufficiency of the UK Online Safety Act'. Computer Law and Security Review 54 (2024); Pope, Audrey. 'NYT v. OpenAI: The Times's About-Face'. Harvard Law Review, 2025; U.S. Copyright Office. Part 1: Digital Replicas. Copyright and Artificial Intelligence. 2024.
[65] EU. (2024). *EU AI Act*. EU AI Act

interagerar med AI-system, exempelvis vid användning av chatbots eller deepfakes. Teknik med låg eller minimal risk omfattas inte av bindande regler.

En särskilt omdiskuterad fråga rör så kallade deepfakes. I dagsläget klassificeras dessa inte nödvändigtvis som hög risk enligt AI-förordningen (AIA) [66]. Samtidigt har USA:s Copyright Office (2024) föreslagit ny lagstiftning som specifikt reglerar så kallade ”digital replicas”, det vill säga AI-genererade representationer av verkliga personer. Syftet med den föreslagna regleringen är att skydda individer vars identitet används utan samtycke, oavsett om identiteten har kommersiellt värde eller inte.

Vems ansvar är det att moderera eller reglera generativt AI-innehåll?

Frågan om ansvar är nära kopplad till riskklassificeringen i AIA. För innehåll som genereras av GPAI-system med låg eller begränsad risk blir ansvarsfördelningen mer oklar. Här uppstår ett juridiskt tolkningsutrymme: ska generativt AI-innehåll betraktas som ett särskilt slags innehåll eller omfattas det av befintliga regler om exempelvis olagligt, vilseledande eller våldsamt innehåll på digitala plattformar såsom sociala medier? Argumentationen skiljer sig mellan jurisdiktioner och plattformar, vilket leder till varierande grader av självsanering, extern reglering och ansvarsutkrävande.

En del av den rättsvetenskapliga litteraturen lyfter behovet av att definiera generativt AI-innehåll som en särskild kategori inom digitala medier [67]. Detta, menar man, skulle möjliggöra mer träffsäkra regelverk och skapa incitament för teknikplattformar att införa tekniska och juridiska skyddsmekanismer. Samtidigt varnas det i flera källor för att överreglering kan hämma innovation, särskilt i sektorer som utbildning, kreativ produktion och medieteknik.

3.5 Risker för barn och unga

Endast ett fåtal vetenskapliga artiklar (6) i det samlade materialet fokuserar explicit på generativt AI-innehåll i sociala medier i relation till barn, utanför användning i utbildningsområdet [68]. En möjlig förklaring kan vara att den sociala medieplattform som oftast används i forskningen, fortfarande är Twitter/X och den inte har en profil mot bland barn och unga. I de artiklar som berör barn handlar det främst om riskbeteenden eller skadliga effekter i form av exempelvis cybermobbning [69].

Fokus på barn och unga är däremot mer framträdande i den grå litteraturen. Här diskuteras både bland annat den data som används för att träna generativa AI-modeller däribland data som kan komma från minderåriga. Där diskuteras också barns exponering för eller interaktion med AI-genererat innehåll [70]. Enskilda fall där barn och unga har begått självmord efter kontakt med AI-chatbots lyfts fram i flera rapporter [71], vilket indikerar behov av vidare forskning kring psykologiska effekter av AI-genererade kommunikationsmönster[72]. Yucong et al. (2025) undersöker i sin experi-

---

[66] Beatriz Kira ‘When Non-Consensual Intimate Deepfakes Go Viral: The Insufficiency of the UK Online Safety Act’. Computer Law and Security Review 54 (2024); Felipe Romero Moreno, ‘Generative AI and Deepfakes: a human rights approach to tackling harmful content’, International Review of Law, Computers & Technology, 38(3), pp. 297–326 (2024).

[67] Abiri Gilad, ‘Generative AI as Digital Media’, *Harvard Journal of Sports and Entertainment Law* 15, no. 2 (2024): 279–332.

[68] Mohamed Chawki, ‘AI Moderation and Legal Frameworks in Child-Centric Social Media: A Case Study of Roblox’; Yucong, Lao, Noora Hirvonen, och Stefan Larsson. ‘AI and Authenticity: Young People’s Practices of Information Credibility Assessment of AI-Generated Video Content’. Journal of Information Science, 15 April 2025; Tama Leaver och Suzanne Srdarov, ‘Generative AI and Children’s Digital Futures’, *Journal of Children and Media* 19 (1): 65–70, 2025; Leung, Janni, Tianze Sun, Daniel Stjepanović, m.fl. “Generative Artificial Intelligence With Youth Codesign to Create Vaping Awareness Advertisements.” JAMA Network Open 8, no. 7 (2025); Milosevic, T., K. Verma, M. Carter, m.fl. ‘Effectiveness of Artificial Intelligence–Based Cyberbullying Interventions From the Youth Perspective’. Social Media and Society 9, no. 1 (2023); Szandra A. Laczi och Valéria Póser, ‘Impact of Deepfake Technology on Children: Risks and Consequences’ in 2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY); Yaman Yu m.fl., ‘Understanding Generative AI Risks for Youth’, arXiv *pre-print*, 2025.

[69] Milosevic, T., K. Verma, M. Carter, m.fl. ‘Effectiveness of Artificial Intelligence–Based Cyberbullying Interventions From the Youth Perspective’. Social Media and Society 9, no. 1 (2023)

[70] Tama Leaver och Suzanne Srdarov, ‘Generative AI and Children’s Digital Futures’, *Journal of Children and Media* 19 (1): 65–70, 2025

[71] Michael Robb och Supreet Mann, *Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions*. Common Sense Media, 2025.

[72] Yucong, Lao, Noora Hirvonen, and Stefan Larsson. ‘AI and Authenticity: Young People’s Practices of Information Credibility Assessment of AI-Generated Video Content’. Journal of Information Science, 15 April 2025

mentella studie olika sätt på vilka unga människor verifierar AI-genererat video innehåll, och visar att teknik, konsekvenser och associerade människor i videos används som verifikationskriterier. Samtidigt visar studien att unga ganska ofta litar på sin intuition, vilket understryker behöver av mer arbete med mediakunskap (MIK).

Flera plattformar har påbörjat egna policyförändringar. Exempelvis har Character.ai tillkännagivit att de kommer att begränsa tillgången till sina AI-chatbots för användare under 18 år [73], och Instagram (Meta) har lanserat nya AI-säkerhetsfunktioner riktade mot föräldrar till tonårsanvändare [74]. Dessa åtgärder visar att även om de stora aktörerna skurit ner på moderingen av sociala medier under 2025 [75] så har aktörerna i alla fall till viss del uppmärksammat de specifika riskerna med AI-chatbots, vilket understryker behovet av tydligare regulatoriska ramar.

Sammantaget visar studien att medan barn och unga utgör en central målgrupp i policydebatten, är de fortfarande relativt underrepresenterade i den vetenskapliga litteraturen om generativ AI i sociala mediemiljöer. Detta tyder på en potentiell forskningslucka med direkt relevans för aktörer som ansvarar för mediepolicy, barns rättigheter och digitalt skydd.

3.6 Användning av generativ AI i informationspåverkan

Den ökande tillgången på generativ AI har inte förändrat målsättning eller struktur på desinformationskampanjer, men har gjort det betydligt enklare att producera desinformationen snabbare, billigare och i större omfattning [76]. De strategier som används är i hög grad desamma som tidigare: exempelvis ”astroturfing”[77], ”barrage jamming”[78] och ”zero day”[79]-kampanjer. Det som förändrats är produktionskapaciteten, samt svårigheten att identifiera AI-genererat innehåll från falska profiler. De kampanjer som undersöks i litteraturen är ofta knutna till välkända politiska eller geopolitiska teman: valpåverkan [80], pandemikommunikation [81], samt desinformation kopplad till kriget i Ukraina [82].

AI-genererade profiler fungerar ofta som sociala "chatagents" som publicerar inlägg innehållande både text och bild. Studier visar att dessa inlägg i många fall är svåra att identifiera som AI-genererade, särskilt om de är väl utformade och anpassade till plattformens normer. Multimodalt innehåll, i synnerhet videor, har visat sig få störst spridning på sociala medier även när det bevisligen rör sig om manipulerat eller falskt material [83]. Ibland blir gränsen mellan avsiktlig desinformation och oavsiktlig missinformation allt mer otydlig när generativ AI används (till exempel en viral bild under 2023 på påven i vit dunjacka). Detta skapar utmaningar både för reglering, detektion och användarutbildning särskilt när innehållet cirkulerar mellan olika plattformar.

---

[73] Zachary Folk, ‘Character.Ai Banning Children Under 18 From Using Their AI Chatbots’, Forbes, 2025.
[74] Siladitya Ray, ‘Instagram Unveils New AI Safety Features For Parents Of Teen Users’, Forbes, 2025.
[75] UNRIC ‘Metas agerande leder till informationskaos och hatretorik’, *Förenta Nationerna*, 15 January 2025.
[76] Rand Corporation, *Artificial Intelligence, Deepfakes, and Disinformation: A Primer*, 2022.
[77] En strategi där falska profiler eller automatiserade konton skapar intrycket av en genuin folklig opinion, trots att kampanjen är organiserad och styrd uppifrån.
[78] Informationspåverkan genom massiv spridning av motstridiga, överväldigande eller repetitiva budskap för att förvirra mottagaren och försvåra källkritik.
[79] Desinformationsinsatser som utnyttjar nya händelser direkt när de inträffar, innan verifierad information hunnit etableras, för att snabbt forma narrativet.
[80] Alphaeus Dmonte m.fl., ‘Classifying Human-Generated and AI-Generated Election Claims in Social Media’, arXiv:2404.16116, preprint, arXiv, 26 April 2024, https://doi.org/10.48550/arXiv.2404.16116. [hämtad 2025-12-11]
[81] Ashley Kearney, Nihal Poredi, Joseph A. Shelton, Seden Akcinaroglu, et al., "Echoes amplified: a study of AI-generated content and digital echo chambers," Proc. SPIE 13480, Disruptive Technologies in Information Sciences IX, 134800L (2025).
[82] Estibaliz García-Huete m.fl., ‘Evaluating the Role of Generative AI and Color Patterns in the Dissemination of War Imagery and Disinformation on Social Media’, *Frontiers in Artificial Intelligence* 7 (January 2025): 1457247.
[83] Nuria Alina Chandra m.fl., ‘Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024’, arXiv:2503.02857, preprint, arXiv, 27 May 2025.

3.7 Journalistisk trovärdighet

Utveckling och användning av generativ AI i journalistik är ett omfattande och omdiskuterat område både ur ett praktiskt perspektiv, där tekniken kan leda till arbetsförlust för många journalister, och ur ett moraliskt perspektiv, där journalistisk integritet ställs mot marknadens krav på snabbhet, underhållning och personlig attraktionskraft hos den person som förmedlar nyheter. Sex st. artiklar i det funna materialet går in på detta tema [84].

Journalister uttrycker oro för att "sanningen" och fakta kan bli otydliga när AI genererar trovärdiga texter, videor eller bilder som inte har genomgått en stringent verifieringsprocess, i synnerhet när de distribueras på sociala medier [85]. Samtidigt kan tekniken effektivisera många delar av det journalistiska arbetet [86]. Det funna materialet visar dock att transparent användning av generativ AI i journalistiska texter kan påverka läsarens tillit negativt inte bara till den enskilda texten, utan till journalistiken som helhet [87]. En del av debatten rör generativa AI-chatbots (till exempel. ChatGPT) som tränas på upphovsrättsskyddat material och sedan sammanfattar artiklar som görs tillgängliga gratis via dessa tjänster [88]. En annan diskussion handlar om möjligheten att ersätta mänskliga reportrar med AI-genererade "personas", som utformas efter vad som anses göra en rapportör mer trovärdig eller publikvänlig [89]. Även om dessa debatter inte fokuserar specifikt på sociala medier, är de på olika sätt relevanta även för denna studie.

3.8 Detektion och märkning av AI-genererat innehåll

Även om det finns omfattande forskning om detektion och verifiering av generativt AI-innehåll generellt, är det få studier som fokuserar specifikt på detektion av genAI i kontexten av sociala medier. En möjlig förklaring är att många detektionsmodeller utvecklas och testas i kontrollerade miljöer med syntetiska dataset, snarare än i dynamiska digitala ekosystem där innehåll modifieras och sprids snabbt [90].

Forskningen visar dock att sociala medier utgör en särskilt utmanande miljö för detektion av AI-genererat innehåll [91]. Innehåll på dessa plattformar är ofta multimodalt, och formatet är anpassat till respektive plattforms normer till exempel korta videor, redigerade bilder eller komprimerad text vilket försvårar både teknisk och visuell verifiering [92]. Inom det tekniska forskningsfältet utvecklas detektionsmodeller i huvudsak genom att skapa eller använda referensdatabaser innehållande både syntetiskt och verkligt innehåll. Modellerna testas sedan experimentellt [93]. En artikel av review-typ som inte är med i underlaget enligt de fastlagda exklusionskriterierna, men ändå är värd att nämnas är Hannah Lee et al. (2024) [94]. De visar att det finns olika metoder för detektion beroende på innehållstyp, till exempel. bild, audio och video. Där framgår att GAN-genererade bilder kan hittas genom unika artefakter, men de nyare tekniker som använder diffusionsmodeller är svårare. För ljud och tal analyseras ljudsignalens mönster. För video görs oftast bild-för-bild analys, och analys av

---

[84] Amy R. Arguedas och Felix M. Simon, *Automating Democracy: Generative AI, Journalism, and the Future of Democracy.*; Borchardt et al., *Trusted Journalism in the Age of Generative AI*; Hannes Cools and Nicholas Diakopoulos, 'Uses of Generative AI in the Newsroom: Mapping Journalists' Perceptions of Perils and Possibilities'. *Journalism Practice,* 1–19, 2024; Phoebe Matich et al., 'Old Threats, New Name?Generative AI and Visual Journalism', *Journalism Practice, 19*(10), 2402–2421, 2025; Petra Petruccio et al., '"A Part of Our Work Disappeared": AI Automated Publishing in Social Media Journalism', *Journalism and Media*, 2025; TJ Thomson et al., 'Generative AI and Journalism : Content, Journalistic Perceptions, and Audience Experiences'. RMIT University, 2025.

[85] Borchardt m.fl., *Trusted Journalism in the Age of Generative AI*. Working paper, University of Oxford, 2024.

[86] Petra Petruccio m.fl., '"A Part of Our Work Disappeared": AI Automated Publishing in Social Media Journalism', *Journalism and Media*, 2025

[87] Arguedas och Simon, *Automating Democracy: Generative AI, Journalism, and the Future of Democracy.*

[88] Audrey Pope, 'NYT v. OpenAI: The Times's About-Face', *Harvard Law Review*, 2025, https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/. [hämtad 2025-12-11]

[89] Hannes Cools och Nicholas Diakopoulos, 'Uses of Generative AI in the Newsroom: Mapping Journalists' Perceptions of Perils and Possibilities'. *Journalism Practice*, 1–19, 2024

[90] Nuria Alina Chandra m.fl., 'Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024', arXiv:2503.02857, preprint, arXiv, 27 May 2025.

[91] Nuria Alina Chandra m.fl., 'Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024', arXiv:2503.02857, preprint, arXiv, 27 May 2025.

[92] Jonas Ricker m.fl., 'AI-Generated Faces in the Real World' in International Symposium on Research in Attacks, Intrusions and Defenses (RAID), 2024, arXiv.

[93] Reza Babaei m.fl., 'Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis', *Journal of Sensor and Actuator Networks* 14, no. 1 (2025): 17.

[94] Hannah Lee m.fl., 'The Tug-of-War Between Deepfake Generation and Detection', arXiv:2407.06174, preprint, arXiv, 21 August 2024.

fysiologiska ovanligheter (till exempel. hur ofta personen på en video blinkar, handrörelser och interaktion med saker runtomkring), och jämför ljud och bild i videon. Det finns ett pågående arbete att utveckla flera ramverk och modeller för detektion (till exempel. Distil-DIRE [95]). Det är ingen skillnad på att detektera genererat material på sociala medier eller på andra plattformar, förutom att det på sociala medier ofta är multimodalt och kortare snuttar (dvs mindre material att detektera på). Detektionsmetoderna för enskilda modaliteter är desamma.

En svårighet för forskningen är att dessa databaser sällan speglar den faktiska spridningen av innehåll på plattformar som TikTok, Instagram eller Discord [96]. Därför visar de funna studierna att hög prestanda i laboratoriemiljö inte alltid kan översättas till effektivitet i praktiken. Språkliga begränsningar är ett annat problem [97]. Många av de mest använda detektionsmodellerna är utvecklade för engelskspråkigt innehåll och kan därför vara mindre träffsäkra på andra språk, till exempel svenska.

Vad gäller verifieringsmetoder återkommer två huvudstrategier: (1) kvalitativ analys av visuella artefakter till exempel onaturliga skuggor, orealistisk ögonplacering eller felaktiga objekt och (2) maskininlärningsmodeller tränade på stora datamängder. Det finns också modeller som kombinerar dessa två metoder, särskilt för att identifiera missinformation/desinformation [98], men ingen metod är i nuläget helt tillförlitlig. Mänskliga granskare har i vissa fall visat sig ha rätt i ungefär 30 % av bedömningarna, när det gäller deepfakes av hög kvalitet [99], vilket understryker teknikens nödvändighet men också dess begränsningar.

Benchmark-dataset och detektionsplattformar som utvecklas inom olika projekt[100] uppdateras ständigt, men i för långsam takt för att hålla jämna steg med den teknologiska utvecklingen. Endast ett fåtal av dessa databaser bygger på verkligt innehåll från sociala medier (till exempel. TwiBot-22[101]).

En pågående debatt rör hur AI-genererat innehåll bör märkas. Ett förslag är att alla former av generativt innehåll ska vattenmärkas, medan ett alternativt förslag är att endast verifierat "verkligt" innehåll ska märkas exempelvis genom metadata om tid, plats och ursprung [102]. En svårighet som tas inte upp i den aktuella tidsramen men förmodligen kommer hända framöver är att vissa genAI-modeller som redan nu tränats på vattenmärkt data har lärt sig att generera olika typer av vattenmärken.

Artiklar i materialet visar dock att märkning av AI-genererat innehåll kan påverka allmänhetens tillit till digitala medier generellt [103]. Samtidigt visar studier att AI-genererat innehåll, särskilt i form av annonser relaterade till samhällsfrågor såsom vaccination eller rökprevention kan ha positiv påverkan på unga om det är tydligt vem som står bakom innehållet [104].

---

[95] Yewon Lim m.fl., 'DistilDIRE: A Small, Fast, Cheap and Lightweight Diffusion Synthesized Deepfake Detection', arXiv:2406.00856, preprint, arXiv, 2 June 2024.

[96] Nick Hajli m.fl., 'Social Bots and the Spread of Disinformation in Social Media: The Challenges of Artificial Intelligence.', British Journal of Management, 33: 1238-1253, 2022.

[97] Nuria Alina Chandra m.fl., 'Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024', arXiv:2503.02857, preprint, arXiv, 27 May 2025.

[98] Kalina Bontcheva m.fl., *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities* (AI4Trust, AI4Media, vera.ai, Innovate UK, Swiss State Secretariat for Education, Research and Innovation(SERI)), 2024.

[99] Andrew Lewis m.fl., 'Deepfake Detection with and without Content Warnings', *Royal Society Open Science* 10, no. 11 (2023).

[100] Några exempel: AI4TRUST (övervakning, detektion och verifikation plattform, EU Horizon finansierad AI4TRUST Platform | Home) https://pilot.platform.ai4trust.eu/ , *DeepfakeEval* (2024 benchmark dataset utvecklad speciellt för deepfakes relaterade till val; Chandra et al., 'Deepfake-Eval-2024'.), *TrueMedia.org* (detektion och verifikation plattform baserad i Georgetown universitet, McCourt School, TrueMedia.org https://www.truemedia.org/) och *DeepMedia* (plattform för detektion av NSFW deepfakes och oetisk användning av deepfakes, Deep Media) https://deepmedia.ai/about-us [hämtad 2025-12-11]

[101] Tingxuan Wu m.fl., MSM-BD: Multimodal Social Media Bot Detection Using Heterogeneous Information, arXiv, 2024.

[102] Sarah A. Fisher, 'Something AI Should Tell You - The Case for Labelling Synthetic Content', Journal of Applied Philosophy 42, no. 1 (2025): 272–86

[103] Chloe Wittenberg m.fl., 'Labeling AI-Generated Media Online', *PNAS Nexus* 4, no. 6 (2025): pgaf170, https://doi.org/10.1093/pnasnexus/pgaf170. [hämtad 2025-12-11]

[104] Leung, Janni, Tianze Sun, Daniel Stjepanović, m.fl. "Generative Artificial Intelligence With Youth Codesign to Create Vaping Awareness Advertisements." JAMA Network Open 8, no. 7 (2025)

4 Diskussion

För att tydliggöra sambandet mellan studiens forskningsfrågor och de tematiska resultat som presenteras i kapitel 3 kan det noteras att de fyra teman som strukturerar analysen är direkt härledda ur de tre forskningsfrågorna. Frågor om AI-genererat innehålls karaktär, användning och påverkan belyses framför allt i temana om detektion, barns sårbarhet och journalistisk trovärdighet. Metodfrågor behandlas främst inom temat kring tekniker för identifiering och verifiering, medan frågor om rättsliga och etiska ramar återkommer både i diskussionen om rättsliga implikationer och i relation till barns rättigheter och journalistikens villkor. Temana utgör därmed inte fristående analysnivåer, utan en syntetisk struktur genom vilken forskningsfrågorna besvaras och sammanställs.

Denna systematiska litteraturöversikt visar att forskningen om generativ AI i sociala medier befinner sig i en expansiv men fragmenterad fas. Det råder ingen tvekan om att generativ AI har förändrat förutsättningarna för digital kommunikation, såväl vad gäller produktion som distribution och konsumtion av innehåll. Samtidigt präglas det vetenskapliga fältet av en stark teknikorientering, där frågor om social påverkan, reglering och användarbeteenden ofta behandlas perifert eller i separata studier.

Fyra tematiska axlar – detektion, barns sårbarhet, journalistikens trovärdighet och regulatoriska frågor – har visat sig vara särskilt återkommande i materialet, men behandlas sällan integrerat i en sammanhållen analys. Detta splittrade kunskapslandskap speglar inte enbart disciplinära gränser, utan även en brist på gemensamma begrepp, definitioner och normativa utgångspunkter för vad AI-genererat innehåll är och bör vara.

En annan tydlig tendens är att forskningen fokuserar oproportionerligt mycket på plattformar som Twitter/X och på engelskspråkigt innehåll. Det innebär att andra relevanta plattformar – såsom TikTok, Snapchat och Roblox samt särskilt viktiga användargrupper som barn och unga ges begränsat utrymme i det vetenskapliga synfältet.

Sammantaget pekar studiens resultat på behovet av en mer sammanhållen och policyinformerad forskning, som inte bara analyserar teknikens funktioner utan också dess samhälleliga konsekvenser.

4.1 Potentiella luckor i forskningen

Flera forskningsluckor framkommer i materialet. För det första är AI-genererat innehåll på sociala medier fortfarande ett nytt område. Det innebär att de flesta empiriska studier som finns är av mindre omfattning, ofta med fallstudier eller små datamängder. I många fall bygger studier på experimentella modeller eller syntetiska data, vilket försvårar överförbarhet till verkliga mediekontexter. Då uppstår en lucka i den ekologiska validiteten.

För det andra finns det få empiriska studier om barns och ungas interaktion med AI-genererat innehåll på sociala medier. De finns några få rapporter från civilsamhället, från tech-företag och från policyaktörer som lyfter denna aspekt[105], men den publicerade vetenskapliga litteraturen på det området är fortfarande begränsad. Detta gäller både användarmönster och psykiska, sociala eller mediepedagogiska konsekvenser.

För det tredje saknas longitudinella studier. Det finns nästan inga studier som följer användare eller innehåll över längre tid. Detta är problematiskt eftersom mycket av AI-genererat innehåll förändras eller tas bort över tid, särskilt på plattformar med kortlivade format (till exempel stories eller livestreams).

För det fjärde är studier med flerspråkigt fokus eller icke-anglosaxiska kontexter underrepresenterade. Merparten av forskningen utgår från engelskspråkiga texter, och fokus ligger på USA,

[105] Tama Leaver och Suzanne Srdarov, ‘Generative AI and Children’s Digital Futures’, Journal of Children and Media 19 (1): 65–70, 2025

Storbritannien och i viss mån EU. Det finns mycket begränsad forskning om svenska, nordiska eller svenska minoritetsspråkliga användares erfarenheter.

En femte forskningslucka gäller ansvarsfrågan. Trots att många artiklar diskuterar etiska och juridiska aspekter, är få av dessa empiriskt grundade. Det saknas till exempel studier om hur plattformar, redaktioner eller andra aktörer i praktiken hanterar ansvar för AI-genererat innehåll. Det saknas också empiriska fallstudier om hur transparenskrav och märkning fungerar i praktiken, både juridiskt och tekniskt.

En sjätte och sista lucka rör användares förmåga att identifiera AI-genererat innehåll. De få studier som finns visar att människor ofta misslyckas med att skilja mellan verkliga och syntetiska bilder, ljud och text särskilt när innehållet är anpassat till plattformens estetik eller ton. Det finns behov av fler studier om mediekompetens och utbildning, särskilt riktade mot unga användare.

4.2 Relevans för policyskapare

Flera av de teman som i denna studie identifierats i litteraturen såsom till exempel. märkning av AI-genererat innehåll, desinformation, barns exponering för medier och journalistikens trovärdighet ligger nära policyskapares nuvarande och framtida tillsynsbehov. Studien visar exempelvis att det saknas etablerade normer för hur AI-genererat innehåll bör märkas, särskilt i journalistiska eller opinionsbildande kontexter. Det råder även osäkerhet kring hur ansvar bör fördelas mellan utvecklare, plattformsoperatörer och användare. Detta är en fråga som delvis adresseras i EU:s AI Act, men där den funna litteraturen visar att tolkningen av reglernas innebörd fortfarande är oklar.

Studien belyser också frågor om kulturell och språklig mångfald. Om generativa AI-system främst tränas på engelska och optimeras för internationella mediekulturer, finns risk att minoritetsspråk, nationella särdrag och svenska publicistiska normer trängs undan. Detta kan få konsekvenser för synligheten och representationen av svenskproducerat innehåll.

De tre forskningsfrågorna har besvarats genom den tematiska struktur som utvecklades i analysen. Karakteriseringen av AI-genererat innehåll behandlas i samtliga teman, särskilt genom analyser av innehållstyper, plattformar och användningsområden. Metodfrågan adresseras främst i diskussionen om tekniker för identifiering och verifiering, medan rättsliga och etiska perspektiv behandlas i temat om regulatoriska implikationer samt i relation till barns rättigheter och journalistik. Temana utgör därmed en samlad syntes av hur frågorna hanterats i det analyserade materialet.

Slutligen indikerar resultaten att nuvarande regelverk såsom Digital Services Act (DSA) och Audiovisuella medietjänstdirektivet (AV-direktivet) kan behöva ses över i ljuset av den växande användningen av generativ AI på sociala medier. För Mediemyndigheten innebär detta ett behov av att delta aktivt i samtalet om vilka regler som krävs och hur de ska tillämpas. Mediemyndigheten har enligt myndighetsförordningen ansvar att följa utvecklingen om myndighetens ansvarsområde på EU-nivå för att hålla regeringen underrättad om utvecklingen där. Att Sverige deltar aktivt i den europeiska debatten är en förutsättning för att reglerna om generativ AI ska hamna på ”rätt nivå” i förhållande till svenska intressen. Studien erbjuder därmed ett kunskapsunderlag som kan stödja myndighetens arbete med policyutveckling, tillsyn och strategisk samverkan.

4.3 Metodreflektion

Studiens styrka ligger i att den tillämpat en metodologiskt stringent, systematisk litteraturöversikt enligt till exempel. PRISMA och Tranfield (2003) m.fl. Den kombinerar kvantitativa översiktsverktyg med tematisk analys och kodning av både tekniskt och policyinriktat material. Samtidigt finns metodologiska begränsningar. Den snabba publiceringstakten på området gör att ny litteratur ständigt tillkommer, och det finns risk för snabb föråldring i vissa delar. Valet att inkludera mestadels engelskspråkiga vetenskapliga artiklar (samt ett selektivt urval av grå litteratur) innebär även att viss nationell eller icke-akademisk kunskap kanske kan ha exkluderats.

En annan utmaning var variationen i begreppsanvändning. Vad som betraktas som ”generativ AI”, "deepfake", "syntetiskt innehåll" eller "AI-agent" skiljer sig åt mellan studier. Detta gjorde kodning och syntes komplex, särskilt i gränsområden där tekniska och samhälleliga perspektiv överlappar. Den tematiska kodningen utvecklades abduktivt, vilket möjliggjorde att nya teman kunde växa fram under analysens gång men också ställde krav på flexibilitet och tolkningsprecision. Studiens avgränsningar inkluderade inte forskning som fokuserar på spridningsmönster på sociala medier eller forskning som har med psykologiska effekter av gen-AI på social medier, vilka är viktiga aspekter för vidare forskning.

4.4 Förslag för framtida forskning

Studien identifierar flera områden där vidare forskning är särskilt angelägen:

- Barn och unga: Empiriska studier om hur barn interagerar med AI-genererat innehåll på sociala medier särskilt utanför skolkontexten, är en kritisk kunskapslucka.
- Plattformsjämförelser: Forskning som jämför hur olika social medieplattformar hanterar AI-genererat innehåll, inklusive märkning, ansvar och användarinteraktioner.
- Svensk och nordisk kontext: Det behövs nationellt förankrade studier, särskilt om policy, juridik och språkliga implikationer av AI-innehåll i svenska mediemiljöer.
- Detektionsmetoder i praktiken: Studier av hur tekniska detektionsmodeller faktiskt fungerar i vardaglig plattformsmiljö.
- Reglering och rättsliga frågor: forskning om implementering och effekter av EU AI Act, DSA och andra regelverk inklusive hur svenska aktörer tolkar och tillämpar dem.
- Mottagarperspektiv: Hur olika målgrupper särskilt unga uppfattar och reagerar på AI-genererat innehåll, med fokus på tillit, identifikation och påverkan.
- En framtida systematisk översikt skulle med fördel kunna integrera fler språk, inkludera icke-akademiska kunskapskällor och följa upp förändringar i takt med att AI-policylandskapet utvecklas.

5 Referenser

Abi-Rafeh, Jad, Leila Cattelan, Hong Hao Xu, Brian Bassiri-Tehrani, Roy Kazan, och Foad Nahai. 'Artificial Intelligence–Generated Social Media Content Creation and Management Strategies for Plastic Surgeons'. *Aesthetic Surgery Journal* 44, no. 7 (2024): 769–78. https://doi.org/10.1093/asj/sjae036. [hämtad 2025-12-11]

Alexander, Sergio. 'Deepfake Cyberbullying: The Psychological Toll on Students and Institutional Challenges of AI-Driven Harassment'. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas* 98, no. 2 (2025): 36–50. https://doi.org/10.1080/00098655.2025.2488777. [hämtad 2025-12-11]

Andrejevic, Mark, och Zala Volčič. 'Automated Parasociality: From Personalization to Personification'. *Television and New Media* 26, no. 4 (2025): 421–37. Scopus. https://doi.org/10.1177/15274764241300436. [hämtad 2025-12-11]

Arguedas, Amy Ross, och Felix M. Simon. *Automating Democracy: Generative AI, Journalism, and the Future of Democracy*. Balliol Interdisciplinary Institute, University of Oxford., 2023.

Ashton, Daniel, och Karen Patel. '"People Don't Buy Art, They Buy Artists": Robot Artists – Work, Identity, and Expertise'. *Convergence* 30, no. 2 (2024): 790–806. Scopus. https://doi.org/10.1177/13548565231220310. [hämtad 2025-12-11]

Ayers, John W., Adam Poliak, Mark Dredze, m.fl. 'Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum'. *JAMA Internal Medicine* 183, no. 6 (2023): 589. https://doi.org/10.1001/jamainternmed.2023.1838. [hämtad 2025-12-11]

Babaei, Reza, Samuel Cheng, Rui Duan, och Shangqing Zhao. 'Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis'. *Journal of Sensor and Actuator Networks* 14, no. 1 (2025): 17. https://doi.org/10.3390/jsan14010017. [hämtad 2025-12-11]

Baele, Stephane J., Elahe Naserian, och Gabriel Katz. 'Is AI-Generated Extremism Credible? Experimental Evidence from an Expert Survey'. *Terrorism and Political Violence* 37 (8): 1060–76. https://doi.org/10.1080/09546553.2024.2380089. [hämtad 2025-12-11]

Bayer, Judit. 'Legal Implications of Using Generative AI in the Media'. *Information & Communications Technology Law* 33, no. 3 (2024): 310–29. https://doi.org/10.1080/13600834.2024.2352694. [hämtad 2025-12-11]

Bontcheva, Kalina, Symeon Papadopoulous, Filareti Tsalakanidou, m.fl. *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities*. AI4Trust, AI4Media, vera.ai, Innovate UK, Swiss State Secretariat for Education, Research amd Innovation(SERI), 2024.

Borchardt, Alexandra, Felix M. Simon, O Zachrison, m.fl. *Trusted Journalism in the Age of Generative AI*. EBU News Report 2024. https://ora.ox.ac.uk/objects/uuid:8c874e2e-34de-4813-ba23-84e6300af110/files/s9593tw89x. [hämtad 2025-12-11]

Bramer, Wichor M., Melissa L. Rethlefsen, Jos Kleijnen, och Oscar H. Franco. 'Optimal Database Combinations for Literature Searches in Systematic Reviews: A Prospective Exploratory Study'. *Systematic Reviews* 6, no. 1 (2017): 245. https://doi.org/10.1186/s13643-017-0644-y. [hämtad 2025-12-11]

Brüns, Jasper David, och Martin Meißner. 'Do You Create Your Content Yourself? Using Generative Artificial Intelligence for Social Media Content Creation Diminishes Perceived Brand Authenticity'. *Journal of Retailing and Consumer Services* 79 (July 2024): 103790. https://doi.org/10.1016/j.jretconser.2024.103790. [hämtad 2025-12-11]

Chandra, Nuria Alina, Ryan Murtfeldt, Lin Qiu, m.fl. 'Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024'. arXiv:2503.02857. Preprint, arXiv, 27 May 2025. https://doi.org/10.48550/arXiv.2503.02857. [hämtad 2025-12-11]

Chawki, Mohamed. 'AI Moderation och Legal Frameworks in Child-Centric Social Media: A Case Study of Roblox'. *Laws* 14, no. 3 (2025). Scopus. https://doi.org/10.3390/laws14030029. [hämtad 2025-12-11]

Chomanski, Bartlomiej, och Lode Lauwaert. 'Automated Propaganda: Labeling AI -Generated Political Content Should Not Be Required by Law'. *Journal of Applied Philosophy* 42, no. 3 (2025): 994–1015. https://doi.org/10.1111/japp.70002. [hämtad 2025-12-11]

Cools, Hannes, och Nicholas Diakopoulos. 'Uses of Generative AI in the Newsroom: Mapping Journalists' Perceptions of Perils and Possibilities'. *Journalism Practice*, 26 August 2024, 1–19. https://doi.org/10.1080/17512786.2024.2394558. [hämtad 2025-12-11]

Dmonte, Alphaeus, Marcos Zampieri, Kevin Lybarger, Massimiliano Albanese, och Genya Coulter. 'Classifying Human-Generated and AI-Generated Election Claims in Social Media'. arXiv:2404.16116. Preprint, arXiv, 26 April 2024. https://doi.org/10.48550/arXiv.2404.16116. [hämtad 2025-12-11]

EU. 'EU AI Act'. EU AI Act, 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. [hämtad 2025-12-11]

Eurostat. *Individuals - Internet Activities*. 2025. https://ec.europa.eu/eurostat/databrowser/bookmark/409256e9-cf9c-4a3b-9f11-d4b037ba1e07?lang=en&createdAt=2025-10-23T10:50:22Z. [hämtad 2025-12-11]

Fisher, Sarah A. 'Something AI Should Tell You - The Case for Labelling Synthetic Content'. *Journal of Applied Philosophy* 42, no. 1 (2025): 272–86. WOS:001293019000001. https://doi.org/10.1111/japp.12758. [hämtad 2025-12-11]

Fisher, Sarah A., Jeffrey W. Howard, och Beatriz Kira. 'Moderating Synthetic Content: The Challenge of Generative AI'. *Philosophy & Technology* 37, no. 4 (2024): 133. https://doi.org/10.1007/s13347-024-00818-9. [hämtad 2025-12-11]

Folk, Zachary. 'Character.Ai Banning Children Under 18 From Using Their AI Chatbots'. Forbes, 2025. [hämtad 2025-12-11] https://www.forbes.com/sites/zacharyfolk/2025/10/29/characterai-will-ban-children-from-speaking-with-chatbots-after-facing-regulatory-pressure-and-lawsuits/.

García-Huete, Estibaliz, Sara Ignacio-Cerrato, David Pacios, m.fl. 'Evaluating the Role of Generative AI and Color Patterns in the Dissemination of War Imagery and Disinformation on Social Media'. *Frontiers in Artificial Intelligence* 7 (January 2025): 1457247. https://doi.org/10.3389/frai.2024.1457247. [hämtad 2025-12-11]

Gilad, Abiri. 'Generative AI as Digital Media'. *Harvard Journal of Sports and Entertainment Law* 15, no. 2 (2024): 279–332.

GOV.UK. 'Online Safety Act: Explainer'. Guidance Online Safety Act: Explainer, 2025. https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer. [hämtad 2025-12-11]

Granåsen, Magdalena, Oskarsson Per-Anders, Maria Olsen, och Niklas Hallberg. *Tvärsektoriell Krishantering: Värdering Av Förmåga Och Modellering Av System En Systematisk Litteraturöversikt*. R FOI-R--5022—SE. FOI, Stockholm: 2021.

GreyNet. 'Grey Literature -'. Greynet International, 2025. https://www.greynet.org/home/aboutgreynet.html. [hämtad 2025-12-11]

Hajli, Nick, Usman Saeed, Mina Tajvidi, och Farid Shirazi. 'Social Bots and the Spread of Disinformation in Social Media: The Challenges of Artificial Intelligence'. *British Journal of Management* 33, no. 3 (2022): 1238–53. https://doi.org/10.1111/1467-8551.12554. [hämtad 2025-12-11]

The White House. 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence'. *The White House*, 30 October 2023. https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. [hämtad 2025-12-11]

Ienca, Marcello. 'On Artificial Intelligence and Manipulation'. *Topoi* 42, no. 3(2023): 833–42. https://doi.org/10.1007/s11245-023-09940-3. [hämtad 2025-12-11]

Insikt Group. *Targets, Objectives, and Emerging Tactics of Political Deepfakes*. Threat analysis. 2024. https://www.recordedfuture.com/research/targets-objectives-emerging-tactics-political-deepfakes. [hämtad 2025-12-11]

Internetstiftelsen. *Kapitel 3. Sociala Medier*. Svenskarna Och Internet. 2025. https://svenskarnaochinternet.se/rapporter/svenskarna-och-internet-2025/sociala-medier/#vilka-sociala-medier-ar-popularast-i-olika-aldrar. [hämtad 2025-12-11]

Kearney, Ashley, Nihal Poredi, Joseph A. Shelton Seden Akcinaroglu, Ekrem Karakoc, Thi Tran, Yu Chen. 'Echoes Amplified: A Study of AI-Generated Content and Digital Echo Chambers'. Proceedings Volume 13480, Disruptive Technologies in Information Sciences IX; 134800L (2025) https://doi.org/10.1117/12.3053447 [hämtad 2025-12-11]

Kira, Beatriz. 'When Non-Consensual Intimate Deepfakes Go Viral: The Insufficiency of the UK Online Safety Act'. *Computer Law and Security Review* 54 (2024). Scopus. https://doi.org/10.1016/j.clsr.2024.106024.

Laczi, Szandra A. och Valéria Póser, 'Impact of Deepfake Technology on Children: Risks and Consequences' in 2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY). 2024, 215–20. https://doi.org/10.1109/SISY62279.2024.10737593.

Lao, Yucong, Noora Hirvonen, och Stefan Larsson. 'AI and Authenticity: Young People's Practices of Information Credibility Assessment of AI-Generated Video Content'. *Journal of Information Science*, 15 April 2025, 01655515251330605. https://doi.org/10.1177/01655515251330605. [hämtad 2025-12-11]

Leaver, Tama, och Suzanne Srdarov. 'Generative AI and Children's Digital Futures: New Research Challenges'. *Journal of Children and Media* 19, no. 1 (2025): 65–70. https://doi.org/10.1080/17482798.2024.2438679. [hämtad 2025-12-11]

Lee, Hannah, Changyeon Lee, Kevin Farhat, m.fl. 'The Tug-of-War Between Deepfake Generation and Detection'. arXiv:2407.06174. Preprint, arXiv, 21 August 2024. https://doi.org/10.48550/arXiv.2407.06174. [hämtad 2025-12-11]

Leung, Janni, Tianze Sun, Daniel Stjepanović, m.fl. 'Generative Artificial Intelligence With Youth Codesign to Create Vaping Awareness Advertisements'. *JAMA Network Open* 8, no. 7 (2025): e2514040. https://doi.org/10.1001/jamanetworkopen.2025.14040. [hämtad 2025-12-11]

Lewis, Andrew, Patrick Vu, Raymond M. Duch, och Areeq Chowdhury. 'Deepfake Detection with and without Content Warnings'. *Royal Society Open Science* 10, no. 11 (2023): 231214. https://doi.org/10.1098/rsos.231214. [hämtad 2025-12-11]

Lim, Yewon, Changyeon Lee, Aerin Kim, och Oren Etzioni. 'DistilDIRE: A Small, Fast, Cheap and Lightweight Diffusion Synthesized Deepfake Detection'. arXiv:2406.00856. Preprint, arXiv, 2 June 2024. https://doi.org/10.48550/arXiv.2406.00856. [hämtad 2025-12-11]

Lv, Linxiang, Yongheng Liang, Siyun Chen, Gus G. Liu, och Jiancai Liao. 'Good Deeds Deserve Good Outcomes: Leveraging Generative Artificial Intelligence to Reduce Tourists' Avoidance of Ethical Brands Embracing Stigmatized Groups'. *Annals of Tourism Research* 110 (2025). https://doi.org/10.1016/j.annals.2024.103889. [hämtad 2025-12-11]

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, m.fl. *AI Index Report 2025*. AI Index Steering Committee. Institute for Human-Centered AI, Stanford University, 2025. https://hai.stanford.edu/ai-index/2025-ai-index-report. [hämtad 2025-12-11]

Matich, Phoebe, T. J. Thomson, och Ryan J. Thomas. 'Old Threats, New Name? Generative AI and Visual Journalism'. *Journalism Practice*, 11 January 2025, 1–20. https://doi.org/10.1080/17512786.2025.2451677.

Miake-Lye, Isomi M., Selene Mak, Meron M. Begashaw, och Paul G. Shekelle. 'Using Google to Search for Evidence: How Much Is Enough? One Center's Experience'. *Systematic Reviews* 14, no. 1 (2025): 92. https://doi.org/10.1186/s13643-025-02836-w. [hämtad 2025-12-11]

Milosevic, Tijana, Kanishk Verma, Michael Carter, Samantha Vigil, Derek Laffan, Brian Davis, och James O'Higgins Norman. 'Effectiveness of Artificial Intelligence–Based Cyberbullying Interventions from Youth Perspective'. *Social Media and Society* 9, no. 1 (2023). https://doi.org/10.1177/20563051221147325. [hämtad 2025-12-11]

Minici, Marco, Federico Cinus, Luca Luceri, och Emilio Ferrara. 'Uncovering Coordinated Cross-Platform Information Operations: Threatening the Integrity of the 2024 U.S. Presidential Election'. *First Monday* 29, no. 11 (2024). https://doi.org/10.5210/fm.v29i11.13831. [hämtad 2025-12-11]

Mink, Jaron, Licheng Luo, Natã M. Barbosa, Olivia Figueira, Yang Wang, and Gang Wang. 'DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks'. *Proceedings of the 31st USENIX Security Symposium, Security 2022*, 2022, 1669–86. [hämtad 2025-12-11]

Nilsson, Per-Erik, Jarlsbo Mathilde, Stefen Werther, m.fl. *Våldsbejakande extremism och digitala medier: En forskningsöversikt* R R--5500--SE. FOI, Stockholm: 2024. https://in.foi.se/rapporter/rapport?reportNumber=FOI-R--5500--SE.

O'Meara, Jennifer, och Cáit Murphy. 'Aberrant AI Creations: Co-Creating Surrealist Body Horror Using the DALL-E Mini Text-to-Image Generator'. *Convergence* 29, no. 4 (2023): 1070–96. https://doi.org/10.1177/13548565231185865. [hämtad 2025-12-11]

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, m.fl. *The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews*. Research Methods & Reporting. British Medical Journal Publishing Group, 29 March 2021. [hämtad 2025-12-11] l https://doi.org/10.1136/bmj.n71.

Park, Jeongeun, Changhoon Oh, och Ha Young Kim. 'AI vs. Human-Generated Content and Accounts on Instagram: User Preferences, Evaluations, and Ethical Considerations'. *Technology in Society* 79 (December 2024): 102705. https://doi.org/10.1016/j.techsoc.2024.102705. [hämtad 2025-12-11]

Petruccio, Petra, Tai Neilson, och Christian Stöcker. '"A Part of Our Work Disappeared": AI Automated Publishing in Social Media Journalism'. *Journalism and Media* 6, no. 1 (2025). Scopus. https://doi.org/10.3390/journalmedia6010030. [hämtad 2025-12-11]

Pope, Audrey. 'NYT v. OpenAI: The Times's About-Face'. *Harvard Law Review*, 2025. https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/. [hämtad 2025-12-11]

Puczyńska, Julia, och Youcef Djenouri. 'AI in Disinformation Detection'. *Applied Cybersecurity & Internet Governance* 3, no. 2 (2024): 211–32. https://doi.org/10.60097/ACIG/200200. [hämtad 2025-12-11]

Rand Corporation. *Artificial Intelligence, Deepfakes, and Disinformation: A Primer*. RAND Corporation, 2022. https://www.rand.org/pubs/perspectives/PEA1043-1.html. [hämtad 2025-12-11]

Ray, Siladitya. 'Instagram Unveils New AI Safety Features For Parents Of Teen Users'. Forbes, 2025. https://www.forbes.com/sites/siladityaray/2025/10/17/instagram-announces-new-parental-controls-for-teen-accounts-accessing-its-ai-chatbots/. [hämtad 2025-12-11]

Ricker, Jonas, Dennis Assenmacher, Thorsten Holz, Asja Fischer, och Erwin Quiring. 'AI-Generated Faces in the Real World: A Large-Scale Case Study of Twitter Profile Images'. *The 27th International Symposium on Research in Attacks, Intrusions and Defenses*, ACM, 30 September 2024, 513–30. https://doi.org/10.1145/3678890.3678922. [hämtad 2025-12-11]

Robb, Michael B., och Supreet Mann. *Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions*. Common Sense Media, 2025. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf. [hämtad 2025-12-11]

Romero Moreno, Felipe. 'Generative AI and Deepfakes: A Human Rights Approach to Tackling Harmful Content'. *International Review of Law, Computers & Technology* 38, no. 3 (2024): 297–326. https://doi.org/10.1080/13600869.2024.2324540. [hämtad 2025-12-11]

Rosati, Eleonora. 'Infringing AI: Liability for AI-Generated Outputs under International, EU, and UK Copyright Law'. *European Journal of Risk Regulation* 16, no. 2 (2025): 603–27. https://doi.org/10.1017/err.2024.72. [hämtad 2025-12-11]

Scheirer, Walter J. 'AI Misinformation Detectors Can't Save Us from Tyranny—at Least Not Yet'. *Bulletin of the Atomic Scientists* 80, no. 5 (2024): 308–13. https://doi.org/10.1080/00963402.2024.2388467. [hämtad 2025-12-11]

Security Conference. 'Accord - Munich Security Conference'. A Tech Accord to Combat Deceptive Use of AI in 2024 Elections, 2024. https://securityconference.org/en/aielectionsaccord/accord/. [hämtad 2025-12-11]

Security Hero. *2023 State of Deepfakes*. 2023. https://www.securityhero.io/state-of-deepfakes/.[hämtad 2025-12-11]

SFU Library. 'Grey Literature: What It Is & How to Find It'. 2025. https://www.lib.sfu.ca/help/research-assistance/format-type/grey-literature. [hämtad 2025-12-11]

Smith, Laura G. E., Richard Owen, Alicia Cork, och Olivia Brown. 'How and Why Psychologists Should Respond to the Harms Associated with Generative AI'. *Communications Psychology* 2, no. 1 (2024): 60. https://doi.org/10.1038/s44271-024-00110-8. [hämtad 2025-12-11]

Spada, Marissa. 'History "for the Gram": AI Beauty and the Vintage, Analogue, and "Throwback" Celebrity'. *Celebrity Studies*, ahead of print, Routledge, 2025. Scopus. https://doi.org/10.1080/19392397.2025.2518089. [hämtad 2025-12-11]

Sumsub. *Identity Fraud Report 2024-2025*. 2025. https://sumsub.com/blog/guides-reports/identity-fraud-report-2024-2025/. [hämtad 2025-12-11]

Svenonius, Ola. Varning – desinformation! – Allmänhetens syn på psykologiskt försvar. FOI-R--5264--SE. Stockholm: FOI, 2022. https://www.foi.se/rapporter/rapportsammanfattning.html?reportNo=FOI-R--5264--SE. [hämtad 2025-12-11]

Tahmasebi, Sahar, Eric Müller-Budack, och Ralph Ewerth. 'Multimodal Misinformation Detection Using Large Vision-Language Models'. Version 1. Preprint, arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.14321.

Tang, Jiaru, och Xiyao Liu. '"NO TO AI GENERATED IMAGES": Fan Art Creators Contesting AI Integration on Social Media Platforms'. *Media International Australia*, 17 June 2025, 1329878X251347005. https://doi.org/10.1177/1329878X251347005. [hämtad 2025-12-11]

Tarsney, Christian. 'Deception and Manipulation in Generative AI'. *Philosophical Studies* 182, no. 7 (2025): 1865–87. https://doi.org/10.1007/s11098-024-02259-8. [hämtad 2025-12-11]

TJ Thomson, Ryan J. Thomas, Michelle Riedlinger, Phoebe Matich, 'Generative AI and Journalism : Content, Journalistic Perceptions, and Audience Experiences'. RMIT University, 2025. https://doi.org/10.6084/M9.FIGSHARE.28068008. [hämtad 2025-12-11]

Tranfield, David, David Denyer, och Palminder Smart. 'Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review'. *British Journal of Management* 14, no. 3 (2003): 207–22. https://doi.org/10.1111/1467-8551.00375. [hämtad 2025-12-11]

UNRIC. 'Metas agerande leder till informationskaos och hatretorik'. *Förenta Nationerna*, 15 January 2025. https://unric.org/sv/metas-agerande-underblaser-informationskaos-och-hatretorik/. [hämtad 2025-12-11]

U.S. Copyright Office. *Part 1: Digital Replicas*. Copyright and Artificial Intelligence. 2024.

Wei, Xuan, Zhu Zhang, Mingyue Zhang, Weiyun Chen, och Daniel Dajun Zeng. 'Combining Crowd and Machine Intelligence to Detect False News on Social Media'. *MIS Quarterly* 46, no. 2 (2022): 977–1008. https://doi.org/10.25300/MISQ/2022/16526. [hämtad 2025-12-11]

Wittenberg, Chloe, Ziv Epstein, Gabrielle Péloquin-Skulski, Adam J Berinsky, och David G Rand. 'Labeling AI-Generated Media Online'. *PNAS Nexus* 4, no. 6 (2025): pgaf170. https://doi.org/10.1093/pnasnexus/pgaf170.

Wu, Chuanhui, Yifan Wang, Yuchen Zhang, Houcai Wang, och Yufei Pang. 'Confront Hate with AI: How AI-Generated Counter Speech Helps against Hate Speech on Social Media?' *Telematics and Informatics* 101 (2025). https://doi.org/10.1016/j.tele.2025.102304. [hämtad 2025-12-11]

Wu, Tingxuan, Zhaorui Ma, Yanjun Cui, Ziyi Zhou, och Eric Wang. *MSM-BD: Multimodal Social Media Bot Detection Using Heterogeneous Information*. arXiv, 2024. https://doi.org/10.48550/arXiv.2501.00204. [hämtad 2025-12-11]

Yu, Yaman, Yiren Liu, Jacky Zhang, Yun Huang, och Yang Wang. 'Understanding Generative AI Risks for Youth: A Taxonomy Based on Empirical Data'. Version 2. Preprint, arXiv, 2025. https://doi.org/10.48550/ARXIV.2502.16383. [hämtad 2025-12-11]

6 Appendix 1 Söksträngar

Sökningen skedde genom tre steg enligt våra forskningsfrågor och endast inom tidsintervallet juni 2022-juni 2025:

Steg 1. Generativ AI innehåll i allmänhet på sociala medier

1.1. Samla in alla artiklar som nämner AI-genererat innehåll på sociala medier

Steg 2. Generativ AI innehåll för desinformation på sociala medier

2.1 Samla in alla artiklar som nämner AI-genererat desinformation på sociala medier

Steg 3. Äkthetsverifiering av AI genererat innehåll på sociala medier

Ytterligare genomförs sökning för specifika teman: barn/unga, ansvar och journalistisk trovärdighet.

Steg 1. AI-genererat innehåll på sociala medier

- **Google Scholar**: "generative artificial intelligence" or "generative AI" and content and "social media"
- **Scopus**: (title-abs-key (generative artificial intelligence) or title-abs-key (ai) and title-abs-key (content) and title-abs-key (social media platform)) and (limit-to (subjarea , "soci")) and (limit-to (doctype , "ar") or limit-to (doctype , "cp")) and (limit-to (language , "english"))
- **Web of Science**: AB =generative artificial intelligence or AB= generative AI and AB= content and AB= social media platform
- **IEEE Xplore**: ("All Metadata":generative artificial intelligence) or ("All Metadata":generative AI) and ("All Metadata":content) and ("All Metadata":social media platform)
    - *Manuell filtrering*: ConferencesJournalsIEEEIEEE Access, topics “soci”, OECD only Affiliation
- **HeinOnline**: "generative artificial intelligence" and content and "social media"
    - *Manuell filtrering*: Section type: articles, OECD location
- **ArXiv**: order: -announced_date_first; size: 200; date_range: from 2022-06-30 to 2025-06-30; classification: Computer Science (cs), Economics (econ), Quantitative Finance (q-fin), Statistics (stat); include_cross_list: True; terms: and all=social media platform; and all=generative artificial intelligence content
    - AND all=twitter/x/facebook/snapchat/tiktok (specific sökning)

Steg 2. AI-genererat desinformation på sociala medier

- **Google Scholar**: "generative artificial intelligence" or "generative AI" and content and "social media" and disinformation
- **Scopus**: (title-abs-key (generative artificial intelligence) or title-abs-key (ai) and title-abs-key (content) and title-abs-key (social media platform) and title-abs-key (disinformation)) and (limit-to (subjarea , "soci")) and (limit-to (doctype , "ar") or limit-to (doctype , "cp")) and (limit-to (language , "english"))
- **Web of Science**: AB =generative artificial intelligence or AB= generative AI and AB= content AND AB= social media platform AND AB=disinformation
- **IEEE Xplore**: (("All Metadata":generative artificial intelligence) or ("All Metadata":generative AI) and ("All Metadata":content) and ("All Metadata":social media platform) AND ("All Metadata":disinformation))

- *Manuell filtrering*: Conferences/JournalsIEEEIEEE Access, topics “soci”, OECD only Affiliation

- **HeinOnline**: "generative artificial intelligence" and content and "social media" and disinformtion
  - *Manuell filtrering*: Section type: articles, OECD location
- **ArXiv**: order: -announced_date_first; size: 200; date_range: from 2022-06-30 to 2025-06-30; classification: Computer Science (cs), Economics (econ), Quantitative Finance (q-fin), Statistics (stat); include_cross_list: True; terms: and all=social media platform; and all=generative artificial intelligence content; and all=disinformation

Steg 3. Äkthetsverifiering av AI genererat innehåll på sociala medier

- **Google Scholar**: "generative artificial intelligence" or "generative AI" and content and "social media" and verification or veracity
- **Scopus**: title-abs-key (generative artificial intelligence) or title-abs-key (ai) and title-abs-key (content) and title-abs-key (social media platform) and title-abs-key (verification) or title-abs-key (veracity) and (limit-to (doctype , "ar") or limit-to (doctype , "cp")) and (limit-to (language , "english"))
- **Web of Science**: AB =generative artificial intelligence OR AB= generative AI and AB= content and AB= social media platform and AB=verification or AB=veracity
- **IEEE Xplore**: (("All Metadata":generative artificial intelligence) or ("All Metadata":generative AI) and ("All Metadata":content) and ("All Metadata":social media platform) and ("All Metadata":verification) or ("All Metadata":veracity)
  - *Manuell filtrering*: ConferencesJournalsIEEEIEEE Access2022 – 2025, topics “soci”, OECD only Affiliation
- **HeinOnline**: "generative artificial intelligence" and content AND "social media" and verification
  - *Manuell filtrering*: Section type: articles, OECD location
- **ArXiv**: order: -announced_date_first; size: 200; date_range: from 2022-06-30 to 2025-06-30; classification: Computer Science (cs), Economics (econ), Quantitative Finance (q-fin), Statistics (stat); include_cross_list: True; terms: and all=social media platform; and all=generative artificial intelligence content; and all=veracity; and all=verification

De särskilda teman som var av intresse var: **(1)** ansvar (responsibility/liability/accountability), **(2)** barn/unga (children/youth) och **(3)** journalistisk trovärdighet (jounalism) För det genomförs ytterligare sökningar på Scopus, Web of Science, Google Scholar (50 första träffar sorterade på relevans) och HeinOnline.

(1) Ansvar
   a. **Google Scholar**: "generative artificial intelligence" or "generative AI" and content and "social media" and responsibility OR liability or accountability
   b. **Scopus**: ((title-abs-key (generative artificial intelligence) or title-abs-key (ai) and title-abs-key (content) and title-abs-key (social media platform) and title-abs-key (responsibility) or title-abs-key (liability) or title-abs-key (accountability)) and (limit-to (doctype , "ar") or limit-to (doctype , "cp")) and (limit-to (language , "English"))
   c. **Web of Science**: AB =generative artificial intelligence or AB= generative AI and AB= content and AB= social media platform and AB=responsibility or AB=liability
   d. **HeinOnline**: "generative artificial intelligence" and content and "social media" and (responsibility OR liability) OR accountability
      i. *Manuell filtrering*: Section type: article, OECD location

(2) Barn/unga

a. **Google Scholar**: "generative artificial intelligence" or "generative AI" AND content and "social media" and children or youth
b. **Scopus**: ((title-abs-key (generative artificial intelligence) or title-abs-key (ai) and title-abs-key (content) and title-abs-key (social media platform) and title-abs-key (children) or title-abs-key (youth)) and (limit-to (doctype , "ar") or limit-to (doctype , "cp")) and (limit-to (language , "english"))
c. **Web of Science**: AB =generative artificial intelligence OR AB= generative AI and AB= content and AB= social media platform and AB=children or AB=youth
d. **HeinOnline**: "generative artificial intelligence" and content and "social media" and (children or youth)
    i. *Manuell filtrering*: Section type: article, OECD location
e. **ArXiv**: order: -announced_date_first; size: 200; date_range: from 2022-06-30 to 2025-06-30; classification: Computer Science (cs), Economics (econ), Quantitative Finance (q-fin), Statistics (stat); include_cross_list: True; terms: and all=social media platform; and all=generative artificial intelligence content; and all=children

(3) Journalistisk trovärdighet

a. **Google Scholar**: "generative artificial intelligence" or "generative AI" and content and "social media" and journalism
b. **Scopus**: ((title-abs-key (generative artificial intelligence) or title-abs-key (ai) and title-abs-key (content) and title-abs-key (social media platform) and title-abs-key (journalism))and (limit-to (doctype , "ar") or limit-to (doctype , "cp")) and (limit-to (language , "english"))
c. **Web of Science**: AB =generative artificial intelligence or AB= generative AI and AB= content and AB= social media platform and AB=journalism
d. **ArXiv**: order: -announced_date_first; size: 200; date_range: from 2022-06-30 to 2025-06-30; classification: Computer Science (cs), Economics (econ), Quantitative Finance (q-fin), Statistics (stat); include_cross_list: True; terms: and all=social media platform; and all=generative artificial intelligence content; and all=journalism

7 Appendix 2 Kodningsschema

Study ID:
Title: Title of paper / abstract / report that data are extracted from
Country in which the study conducted

1. United States
2. UK
3. Canada
4. Australia
5. Sweden
6. Other

Notes

Methods:

_____

Aim of study:

_____
Study design:

1. experiment/lab study
2. Survey/quantitative
3. Case study
4. Systematic review
5. Qualitative research: interviews, (n)etnography
6. Prevalence study
7. Case report
8. Text and opinion
9. Data set presentation and testing
10. Other

Study object/subject
Äkthetsverifiering, detektion och moderering

1. platform moderation
2. verification
3. benchmark data
4. labeling
5. community moderation
6. detektion

Platform

1. Facebook
2. Snapchat
3. TikTok
4. Instagram
5. Twitter/X
6. Other

(Total number of) participants:

AI content (abduktiv kodlista, växte under läsning)

1. influencers
2. chat agents
3. deepfake
4. meme
5. pornography
6. cyberbullying
7. political campaigns
8. disinformation
9. misinformation
10. elections
11. covid
12. Rus-Ukr war
13. fraud/deephishing
14. dating
15. education
16. social issues awareness
17. AI generated vs AI manipulated carification
18. medicine
19. multimodal content
20. astroturfing
21. barrage jamming
22. zero day
23. voice
24. video
25. image
26. text
27. marketing
28. art
29. identity/personification
30. extremism