LUNDS UNIVERSITET
Ekonomihögskolan

Institutionen för informatik

AI-genererat innehåll på sociala medier

Användaruppfattning av märkningssystem på Instagram och TikTok

Kandidatuppsats 15 hp, kurs SYSK16 i Informatik

Författare: Milica Milosavljevic
Emira Delic
Izzy Ronja Tiger

Handledare: Umberto Fiaccadori

Rättande lärare: Markus Lahtinen
Paul Pierce

AI-genererat innehåll på sociala medier: Användaruppfattning av märkningssystem på Instagram och TikTok

ENGELSK TITEL: AI-generated Content on Social Media: How Labeling Systems are Perceived on Instagram and TikTok

FÖRFATTARE: Emira Delic, Milica Milosavljevic, Izzy Ronja Tiger

UTGIVARE: Institutionen för informatik, Ekonomihögskolan, Lunds universitet

EXAMINATOR: Osama Mansour, Docent

FRAMLAGD: Maj, 2026

DOKUMENTTYP: Kandidatuppsats

ANTAL SIDOR: 38

NYCKELORD: AI-genererat innehåll, AI-märkning, Användaruppfattning, Instagram, TikTok, Transparens, Trust in Information Systems

SAMMANFATTNING:

Denna studie undersöker hur användare uppfattar märkningar för AI-genererat innehåll på Instagram och TikTok, med fokus på transparens och förtroende. Studien bygger på en kvantitativ forskningsdesign där data samlas in genom en enkätundersökning. Den teoretiska referensramen utgår från *Trust in Information Systems* samt forskning om AI-märkningars design, synlighet och kognitiv belastning. Resultaten visar att majoriteten av respondenterna är väl bekanta med AI-genererat innehåll och AI-märkningar, men att många samtidigt upplever märkningarna som inkonsekventa eftersom AI-genererat innehåll ofta förekommer utan märkning. AI-märkningar uppfattades generellt som relativt lätta att förstå, men inte alltid eller helt tillförlitliga. Studien visar även att olika sätt att presentera AI-märkningar påverkar användarnas förhållning till de i både förtroende och känslomässigt. Vidare indikerar resultaten att AI-märkningar till viss del bidrar till ökad transparens och påverkar användarnas förtroende för innehåll, men att de inte tydligt stärker förtroendet för själva plattformarna. Studien belyser därmed vikten av tydliga, konsekventa och synliga AI-märkningar för att stödja användares förståelse och förtroende i digitala informationsmiljöer.

Innehåll

Modeller

Figurer

1 Introduktion

I detta kapitlet introduceras studiens bakgrund och problemområde med fokus på AI-genererat innehåll och AI-märkningar på sociala medier. Kapitlet redogör för varför användares uppfattningar av AI-märkningar är relevanta att undersöka i relation till förtroende och transparens. Därefter presenteras studiens forskningsfråga, syfte, avgränsningar och kunskapsintressenter.

1.1 Bakgrund

Utveckling av artificiell intelligens har under de senaste åren lett till en kraftig ökning av AI-genererat innehåll på sociala medieplattformar. AI-genererat innehåll förekommer idag i stor utsträckning på digitala sociala medier, som exempelvis Instagram och TikTok, och har blivit en central del av dagens informationsmiljö. Sådant innehåll påverkar samhället på många nivåer, men specifikt hur information produceras och konsumeras på sociala medier (Gamage et al., 2025).

Idag används generativ AI för att skapa text, ljud, bilder och videor genom att transformera användares prompts till realistiska digitala outputs. AI-innehåll har blivit en integrerad del av sociala medier och användare exponeras alltmer för AI-genererade deepfakes, bilder, videor och text som efterliknar mänsklig kommunikation och visuella uttryck. Detta innehåll kan i många fall vara svårt att skilja från mänskligt skapat material.

Med denna utveckling uppstår frågor kring transparens, autenticitet och informationspålitlighet på sociala medier. När användare inte kan avgöra om innehållet är skapat av människor eller AI påverkar det hur information uppfattas och sprids. Detta bidrar till ökad risk för missinformation och feltolkning, samtidigt som trovärdigheten hos både innehållet och sociala medieplattformarna kan påverkas negativt.

För att hantera dessa utmaningar har flera plattformar, inklusive Meta Platforms och TikTok, börjat införa märkningssystem för AI-genererat innehåll. Dessa system syftar till att öka transparensen genom att informera användare om omfattningen av AI som används i innehållet, alltså vad som är och inte är AI-genererat (Bickert, 2024). Tidigare forskning visar att märkning av AI-genererat innehåll påverkar hur användare tolkar innehåll på sociala medier, vilket förstärker behovet av mekanismer som tydliggör när innehåll är AI-genererat (Seeger et al., 2026).

Forskning betonar samtidigt vikten av hur sådana system utformas och kommuniceras för att vara effektiva (Jung et al., 2025). Presentation och design av märkningen anses också påverka i vilken utsträckning användare förstår och lägger märke till den. Tidigare studier visar även att olika typer av märkningar skapar olika reaktioner hos användare beroende på terminologi, tydlighet och hur märkningen uppfattas i relation till innehållet (Gamage et al., 2025).

1.2 Problematisering

Trots att AI-märkningar implementerats på flera sociala medieplattformar finns det fortfarande osäkerhet kring hur dessa märkningar fungerar i praktiken ur ett användarperspektiv. Märkningarna är i vissa fall inkonsekventa, med mänskligt innehåll som felaktigt identifieras som AI och AI-innehåll som saknar märkning. Varningsetiketter för AI-innehåll möjliggör transparens, men det är inte bevisat om transparensen faktiskt leder till förtroende för plattformen. Dessa märkningar kan dessutom påverka hur användare engagerar sig med innehållet (Gamage et al., 2025).

Det är i nuläget fortfarande oklart hur dessa märkningar uppfattas av användare och i vilken utsträckning de faktiskt ökar transparens kring AI-genererat innehåll i praktiken. Därför finns det ett behov av att undersöka hur märkningarna uppfattas, särskilt i relation till tydlighet, synlighet och förståelse på specifika plattformar som Instagram och TikTok.

1.3 Val av område

Under de senaste åren har AI-genererat innehåll blivit allt mer utbrett på sociala medier. Som användare har vi själva observerat konsekvenserna av denna utveckling, där personer i vår omgivning, exempelvis familjemedlemmar och andra användare med varierande digital kompetens, i vissa fall har haft svårt att avgöra vad som är autentiskt innehåll. Detta har väckt ett intresse kring hur AI-genererat innehåll presenteras och uppfattas på digitala plattformar.

En central observation är att AI-genererat innehåll inte alltid är tydligt märkt, och att märkningarna är inkonsekventa med innehåll som felaktigt märks som AI och AI-innehåll som saknar märkning. Samtidigt kan befintliga märkningar i vissa fall vara svåra att uppfatta eller förstå. Detta kan bidra till att information misstolkas och sprids vidare. Vi har även observerat att vissa märkningar är placerade på ett diskret sätt vilket gör att de är lätta att missa vid konsumtion av innehåll på sociala medier.

Detta har väckt frågor hos oss om hur synliga dessa märkningssystem är och hur de uppfattas av användare. Studien tar därmed sin utgångspunkt i ett användarperspektiv med fokus på transparens och förtroende i sociala medier.

1.4 Forskningsfråga

Hur uppfattar användare märkningssystem för AI-genererat innehåll på Instagram och TikTok i relation till förtroende och transparens?

I denna studie avser **användaruppfattning** hur användare uppmärksammar, tolkar och värderar AI-märkningar på Instagram och TikTok. Begreppet omfattar användares upplevelser av märkningarnas synlighet, tydlighet, begriplighet och tillförlitlighet samt hur dessa upplevelser relaterar till förtroende för innehållet och upplevd transparens.

1.5 Syfte

Syftet med denna studie är att undersöka hur användare uppfattar AI-märkningar på sociala medieplattformarna Instagram och TikTok, med särskilt fokus på hur märkningarnas tydlighet, synlighet och upplevda konsekventa tillämpning påverkar användarnas uppfattning om transparens och förtroende för digitalt innehåll. Studien avser även att analysera hur dessa uppfattningar kan förstås utifrån ett förtroendeperspektiv, genom det teoretiska perspektivet Trust in Information Systems, för att belysa vilken betydelse AI-märkningar har för användares tillit till AI-genererat innehåll i digitala informationsmiljöer.

Ämnet är relevant att studera eftersom användares förståelse och uppfattning av AI-märkning påverkar hur information på sociala medieplattformar tolkas, värderas och sprids vidare. Ökad kunskap inom området kan därmed bidra till en bättre förståelse för hur transparens kring AI-genererat innehåll kan kommuniceras.

1.6 Avgränsningar

Studien avgränsas till sociala medieplattformarna Instagram och TikTok. Just dessa plattformar valdes då de är bland de som används mest och de är på många sätt väldigt lika varandra. Eftersom de är så lika har valet gjorts att inte jämföra dem, utan bara undersöka frågeställningen utifrån båda och sammanställa resultaten.

Forskningsområdet begränsas till användares uppfattningar av märkning för AI-genererat innehåll på dessa plattformar. Fokus ligger på hur märkningarnas tydlighet och synlighet uppfattas av användare i relation till transparens kring innehållets ursprung och förtroende för både innehållet och plattformarna. Studien genomförs ur ett användarperspektiv och baseras på användares erfarenheter och tolkningar av AI-märkningar i dessa sociala medier.

1.7 Kunskapsintressenter

Det förväntade bidraget till informatik är ökad kunskap och förslag om hur märkningssystem för AI-genererat innehåll kan förstås och uppfattas inom sociala medieplattformar. Det ger insikter kring hur märkning bör presenteras för användare samt organisatoriska och tekniska överväganden kring implementering. Därav bidrar arbetet till förståelsen för hur informationssystem kan stödja transparens och ansvarsfull användning av AI i digitala informationsmiljöer. Resultaten kan vara relevanta både för forskare inom informatik och för praktiker som utvecklar och förvaltar digitala plattformar. Målgruppen inkluderar teknikföretag och organisationer som arbetar med digital transparens och AI-reglering, exempelvis intressenter inom Meta Platforms och TikTok. Användare av dessa plattformar och allmänheten som vill förstå hur AI-genererat innehåll märks och uppmärksammas av användare på digitala plattformar anses också vara en intressentgrupp.

2. Litteraturgenomgång

I detta kapitel presenteras det teoretiska ramverket och den tidigare forskning som ligger till grund för studien. Först presenteras den tidigare forskning om AI-märkningar och märkningars utformning på sociala medieplattformar. Därefter presenteras de teoretiska perspektiv och modeller som används för att förstå användares uppfattningar av AI-märkningar i relation till förtroende och transparens.

2.1 Tidigare forskning

Detta avsnitt behandlar tidigare forskning om AI-märkningar på sociala medier, med fokus på deras funktion som transparensmekanismer, deras utformning samt deras begränsningar och oavsiktliga effekter.

2.1.1 AI-märkningars påverkan på användaruppfattning och engagemang

H. Chen et al. (2025) har undersökt AI-märkningens påverkan på användarengagemang i plattformar för kortformsvideo, det vill säga sociala medieplattformar där användare främst konsumerar korta videoklipp, exempelvis TikTok och Instagram Reels. Resultat av undersökningen indikerar att användarnas negativa attityder mot AI-genererat innehåll kan lindras genom att plattformarna erbjuder mer information om AI och dess förmågor. Samtidigt visar resultaten att redovisning av AI-användning påverkar hur användare uppfattar innehållets autenticitet och trovärdighet, vilket kan påverka hur de engagerar sig med kortformsvideor. H. Chen et al. (2025) tar också hänsyn till hur innehållsskapare kan öka engagemanget hos användare, där de uppmuntras att förbättra sin förmåga att reglera innehållskvalitet och följa plattformens riktlinjer att säkerställa noggrannhet och effektivitet av AI i skapandet av innehållet.

I sin undersökning om AI-labeling presenterar Jung et al. (2025) hur användare reagerar på olika typer av AI-märkningar. Studien fokuserar mest på utformning och kommunikation av märkningssystem i sociala medieplattformar. Resultaten visar att användare har delade åsikter kring AI-märkningar, där vissa uppfattar märkningarna som trenddrivna signaler snarare än informationsbärande transparensmarkörer, medan vissa upplevde automatiskt att innehållet är fejk.

När det gäller vem som skapade märkningen visar användare att de föredrar *self-labeling* (där innehållsskapare själva markerar innehållet som AI-genererat) över plattformarnas märkning. Denna studie stödjer H. Chen et al. (2025) i form av praktiska implikationer, eftersom även den uppmuntrar användare som lägger ut innehåll att själva lägga AI-märkning vid utläggning av innehållet. När AI-innehåll inte markeras av skaparen utan istället flaggas av plattformen belyser detta en bristande transparens, vilket kan minska förtroendet för både skaparen och innehållet.

Jung et al. (2025) föreslår att märkningens meddelande och språk bör vara tydligt och mer specifikt, samt att AI-märkningar bör vara uppdelade i flera kategorier baserat på hur stor andel av innehållet som är AI-genererat. Samtidigt visar studien att användaruppfattning av

innehållets ursprung påverkas av otydlig användning av AI i innehållet, alltså i vilken utsträckning AI användes i skapandet av innehållet.

Vidare undersöker J. Chen et al. (2025) hur detaljnivån i AI-märkningar påverkar användares uppfattningar om AI-genererat innehåll på sociala medier. Resultaten visar att mer detaljerade märkningar ökar upplevd transparens, informativitet och trovärdighet. Innehållets kontext visas också påverka hur trovärdigt AI-genererat innehåll uppfattas, där "*låg-stakes*"-innehåll generellt bedöms som mer autentiskt. Studien föreslår att sociala medieplattformar kan använda maximalt detaljerade märkningar för att öka transparens utan att negativt påverka engagemanget. Märkningsstrategier bör anpassas beroende på innehållets *stakes*, där detaljnivån för märkningen kan vara mer avgörande för innehåll som är "*high-stakes*", exempelvis nyheter eller hälsa. Slutligen kan kreatörer specificera innehållets ursprung och på så sätt öka transparens utan att riskera minskat engagemang (J. Chen et al., 2025).

2.1.2 AI-märkningars utformning, synlighet och terminologi

AI-märkningar är inte endast informationselement, utan kan även påverka hur användare tolkar och bedömer digitalt innehåll. Forskning av Gamage et al. (2025) visar att användare ofta förlitar sig på snabba visuella heuristiker, såsom ikoner, färger, symboler och placering, för att bedöma tillförlitligheten och äktheten hos onlineinnehåll innan de medvetet bearbetar detaljerad information. Morrow et al. (2022) betonar särskilt att sociala medier är visuellt orienterade miljöer där små designval såsom placering, visuell hierarki och grafiska element påverkar hur användare uppfattar och interagerar med märkt innehåll. Detta är relevant för AI-märkningar eftersom deras funktion inte enbart beror på att de finns i gränssnittet, utan också på hur de visuellt presenteras och uppmärksammas av användare.

Förtroende är dock inte en faktor som enbart bestäms av närvaron av information, utan av hur information uppfattas och tolkas. Enligt Epstein et al. (2023) kan en användares förståelse av AI-märkningar variera avsevärt beroende på terminologi och förkunskaper om AI. Till exempel är etiketter som "AI-genererad" generellt förknippade med innehåll skapat av AI-system, medan termer som "deepfake" är starkare förknippade med vilseledande innehåll, oavsett om AI var inblandat. Detta indikerar att användare inte alltid tolkar etiketter som de är avsedda, vilket direkt påverkar hur förtroende formas.

Epstein et al. (2023) visar även att AI-märkningar kan påverka användares attityder till både innehållet och dess källa. Märkningar som "deepfake" eller "manipulerad" tenderar att framkalla mer negativa reaktioner då de begreppen har mer negativa konnotationer och därmed minskar tron på innehållet, medan mer neutrala märkningar har svagare effekter. Detta bevisar att hur innehållet är märkt inte är den enda faktorn som påverkar användares förtroende för innehållet, utan också av hur märkningen är utformad (Epstein et al., 2023).

Ett relevant perspektiv i detta sammanhang är kognitiv belastningsteori, vilket förklarar hur användare har begränsad mental kapacitet när de bearbetar information online. När etiketter (labels) innehåller överdriven information eller är visuellt komplexa kan användare ignorera dem eller uppleva informationsöverbelastning. Däremot kan enkla och igenkännbara visuella signaler hjälpa användare att bearbeta information mer effektivt. En studie av Gamage et al. (2025) fann att deltagarna föredrog etiketter som var tydliga, koncisa och lätta att förstå vid en blick, medan alltför detaljerade etiketter skapade förvirring och ökad kognitiv belastning.

En annan viktig aspekt är placeringen och synligheten av AI-märkningar. Studien fann att deltagarna föredrog etiketter placerade diskret i hörn eller utanför bilden snarare än att direkt skymma innehållet. Etiketter som var för påträngande påverkade användarupplevelsen negativt, medan etiketter som var för subtila riskerade att ignoreras helt. Detta belyser den balans som plattformar måste uppnå mellan synlighet och användbarhet när de utformar AI-märkningar.

2.1.3 AI-märkningars begränsningar och oavsiktliga effekter

AI-märkningar kan förstås som en funktion som är avsedd för att stödja användares förtroendebedömningar genom att ge ytterligare ledtrådar om hur innehållet skapades (Wittenberg et al., 2023). Trots detta kan de också ha vissa begränsningar och oavsiktliga effekter på förtroende för plattformen och innehållet. Wittenberg et al. (2023) i sin artikel "Labeling AI-generated media online" visar att AI-relaterade avslöjanden kan minska användarnas sannolikhet att tro på och dela vilseledande information, men deras effektivitet beror på hur de är utformade och presenterade. Detta innebär att AI-märkningar inte automatiskt leder till ökad trovärdighet eller förtroende, utan att deras funktion påverkas av hur användare uppfattar märkningarnas tydlighet, placering och relevans.

En central oavsiktlig effekt är implied truth effect. Morrow et al. (2022) beskriver implied truth effect som en möjlig oavsiktlig konsekvens av innehållsmärkningar i sociala medieflöden. Effekten innebär att om vissa inlägg märks som falska eller problematiska kan användare börja tolka omärkt innehåll som mer autentiskt eller trovärdigt just eftersom det saknar märkning. I kontexten av AI-genererat innehåll innebär detta att inkonsekvent märkning kan skapa en falsk trygghet, där användare antar att omärkt innehåll är mänskligt skapat eller mer pålitligt.

En ytterligare oavsiktlig effekt av AI-märkningar är att de kan forma användares övergripande uppfattning om en informationsmiljö. Li et al. (2026) visar att AI-märkningar kan göra användare mer skeptiska mot digitalt innehåll generellt. Detta kan ha både positiva och negativa konsekvenser. Ökad skepticism kan stärka användares kritiska bedömning av innehåll, men kan också bidra till minskat förtroende för online information och sociala medieplattformar.

2.2 Teoretiskt ramverk

Detta avsnitt presenterar de teoretiska perspektiv och modeller som används för att analysera användares uppfattningar av AI-märkningar i relation till förtroende, transparens och informationsbearbetning.

2.2.1 Trust in Information Systems

Trust in Information Systems används i denna studie som ett teoretiskt perspektiv för att förstå hur användare utvecklar förtroende för digitala system och den information som dessa system tillhandahåller. Förtroende kan definieras som viljan hos användare att lita på ett system baserat på deras uppfattningar om reliabilitet, kredibilitet och integritet (Müller et al., 2025). På digitala plattformar, där innehåll skapas, sprids och konsumeras i stora mängder, blir

förtroende centralt för användarens förmåga att bedöma informationens trovärdighet och digitalt innehålls ursprung (Müller et al., 2025).

Trust in Information Systems är relevant för denna studie eftersom AI-märkningar på Instagram och TikTok är informationselement i de två sociala medieplattformarnas gränssnitt. Syftet med märkningarna är att informera användare om innehållets ursprung, till exempel om ett inlägg är AI-genererat eller modifierat med hjälp av AI. För att märkningarna ska kunna fungera som ett stöd till användare och deras bedömningar är det viktigt att användarna uppfattar märkningarna som tillförlitliga. Om en AI-märkning upplevs som otydlig eller inkonsekvent kan det påverka en användares förtroende negativt (Li & Yang, 2024).

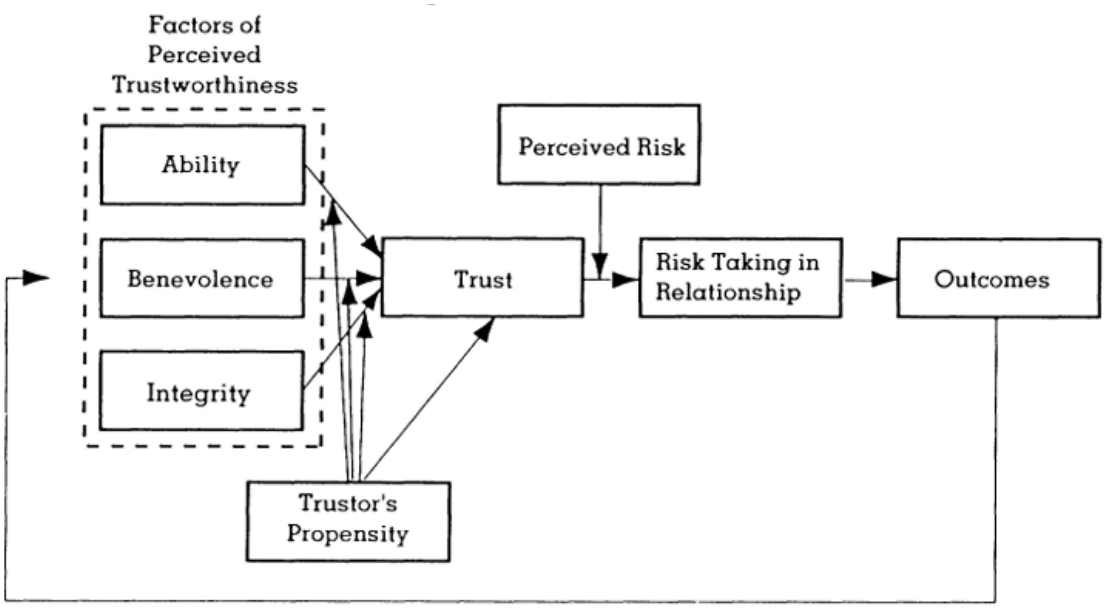
McKnight et al. (2002) visar att förtroende i digitala miljöer består av flera dimensioner som påverkar hur användare förhåller sig till digitala system. I digitala miljöer interagerar användare genom gränssnittet, plattformens funktioner och med den information som presenteras, och inte genom direkt kontakt med aktören bakom systemet. Detta gör förtroende särskilt viktigt på sociala medieplattformar där användare snabbt behöver tolka och värdera information. I relation till AI-märkningar innebär detta att användares förtroende påverkas av hur märkningarna presenteras, hur begripliga de är och om de uppfattas som pålitliga signaler om innehållets ursprung.

Utifrån detta bidrar Trust in Information Systems till att förstå varför AI-märkningars tydlighet, begriplighet och tillförlitlighet är viktiga för användares upplevelse av förtroende och transparens. Trust in Information Systems ger därmed en övergripande grund för att analysera hur användare uppfattar AI-märkningar som informationselement i Instagram och TikTok, och hur dessa märkningar relaterar till förtroende och transparens. För att analysera dessa förtroendedimensioner mer specifikt används Mayers Trust Model, som presenteras i nästa avsnitt.

2.2.2 Mayers Trust Model

Förtroende kan även förstås genom Mayer, Davis & Schoormans (1995) integrerade modell, där trust definieras som en individs vilja att göra sig sårbar gentemot en annan part baserat på positiva förväntningar om dess agerande. Modellen förklarar att förtroende formas genom tre centrala dimensioner: ability, benevolence och integrity (Modell 1). Denna modell har fortsatt ha betydelse inom mer modern forskning om förtroende, och Schoorman, Mayer och Davis (2007) betonar att dessa tre dimensioner fortsätter att vara centrala för att förstå hur förtroende formas.

Ability dimensionen avser uppfattningen om de kompetenser och egenskaper en aktör har för att utföra en specifik uppgift, samt om aktören har resurser, expertis och kapacitet. Benevolence avser uppfattningen att en aktör har goda intentioner gentemot den som visar förtroende. Det innebär att användaren upplever att aktören inte enbart agerar utifrån egenintresse och har inget egocentriskt vinstmotiv. Integrity handlar om till vilken grad en aktör uppfattas agera enligt principer som användaren finner acceptabla (Mayer et al., 2007). Det handlar om konsekvens, ärlighet och efterlevnad av uttalade principer.



Modell 1: Mayers Model of Trust

Ability, benevolence och integrity kan vara oberoende faktorer, men saknad av någon av dessa faktorer kan minska förtroendet. Alla tre dimensioner är därför viktiga för att öka förtroendet hos användarna.

Dessa tre dimensioner, när de tillämpas i kontexten av sociala medieplattformar, kan användas för att förstå hur användare tolkar AI-märkningar. Till exempel kan AI-märkningar signalera plattformens integritet genom att informera användaren om innehållets ursprung. Samtidigt kan detta kopplas till om AI-märknings tillämpning är konsekvent och begriplig, samt om hur användaren uppfattar att plattformen öppet redovisar hur märkningen fungerar. Tydliga och konsekventa märkningssystem kan därmed bidra till att stärka uppfattningen om Instagram och TikToks integritet, medan bristande transparens riskerar att skapa osäkerhet och skepticism.

I samma kontext kan ability förstås som användares uppfattning om plattformens kapacitet att korrekt identifiera, klassificera och märka AI-genererat innehåll. Om märkningarna å andra sidan upplevs som inkonsekventa, felaktiga eller otydliga kan det försvaga uppfattningen om plattformens tekniska förmåga.

Samtidigt kan AI-märkningens funktionalitet och implementering ha en påverkan på hur en användare uppfattar till vilken grad en plattform agerar i deras intresse, vilket kopplas till benevolence faktorn. AI-märkningar kan därför förstås som uttryck för ansvariga för de bägge plattformarnas vilja att bidra till transparens och hjälpa användare att förstå innehållets ursprung. Däremot kan otydliga eller ytliga märkningar minska upplevelsen av benevolence (välvilja).

Modellen är särskilt relevant för studien eftersom den möjliggör en analys av hur användares uppfattningar om AI-märkning kan förstås genom konkreta förtroendedimensioner: ability, benevolence och integrity. Inom kontexten av informationssystem kan användares förtroende dessutom riktas mot flera olika objekt, exempelvis systemet, teknologin, systemleverantören eller informationen som systemet presenterar (Söllner et al., 2016). I denna studie innebär det att användares förtroende både kan avse AI-märkningen som informationselement och

Instagram eller TikTok som plattformar. Mayer-modellen ger på det sätt en struktur för att analysera hur användares uppfattningar om AI-märkningar relaterar till förtroende och transparens.

2.2.3 Kognitiv belastningsteori och informationsöverbelastning

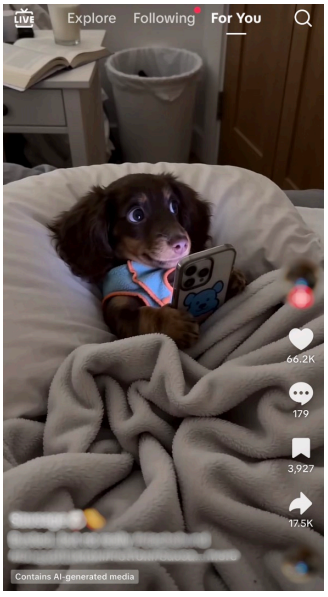
Denna studie beaktar också kognitiv belastningsteori i relation till informationsöverbelastning. Kognitiv belastningsteori utgår från att en hög grad av informationsöverbelastning kan påverka användarens informationsbearbetning negativt (Paas et al., 2003; Marcus & Vasser, 2024). Detta kan öka risken för att information inte uppfattas korrekt, missas helt eller att användaren blir distraherad från övrig information (Eppler & Mengis, 2004).

Kognitiv belastningsteori utgår på att förstå hur människor kognitivt hanterar inlärning och bearbetning av ny information. Teorin grundar sig i att människor har ett arbetsminne med begränsad kapacitet när det gäller att processa ny och obekant information (Paas et al., 2003; Marcus & Vasser, 2024). Arbetsminnet är begränsat både i hur mycket information som kan hanteras samtidigt och hur länge den stannar kvar samtidigt som långtidsminnet har en närmast obegränsad kapacitet för att lagra bekant kunskap. Kunskap finns lagrad i långtidsminnet i form av scheman, vilka är kognitiva konstruktioner som tillåter att stora mängder information organiseras och bearbetas som en enda enhet. Genom övning kan dessa scheman automatiseras, vilket innebär att de kräver minimalt med resurser från arbetsminnet vid framtida användning. I vardagen förlitar sig människor på både arbetsminnet och på långtidsminnet (Marcus & Vasser, 2024).

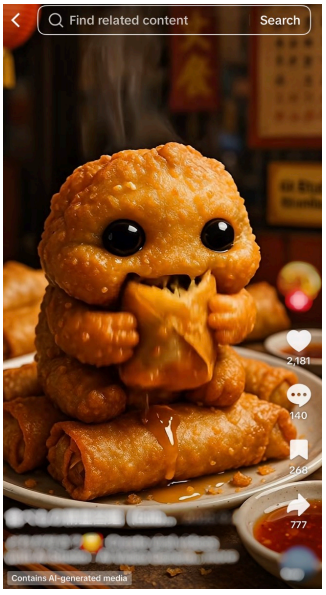
Inom teorin skiljer man mellan tre typer av kognitiv belastning, dessa är inre, överflödig och relevant belastning (Paas et al., 2003). För denna studien blir det relevant att fördjupa sig i överflödig kognitiv belastning (extraneous load). Denna typ av överbelastning handlar om hur informationen presenteras och i vilken utsträckning designen tvingar användaren att lägga mental energi på irrelevanta aktiviteter, såsom att söka efter information eller försöka tyda den (Marcus & Vasser, 2024). Inom kontexten av att användare av sociala medier konsumerar innehåll som är AI-markerade, är det primära målet att genom design minimera den överflödiga belastningen. Märkningen riskerar annars att distrahera från innehållet som leder till att användarupplevelsen försämras och användarna riskerar att feltolka information.

För att fördjupa sig i vad som påverkar den kognitiva belastningen av AI märkningar måste man diskutera transparens. Kopplingen mellan kognitiv belastning och transparens blir särskilt tydlig när man studerar märkningssystem för AI-genererat innehåll på plattformar som Instagram och TikTok. AI-märkningar fungerar som ett informerande element i innehållet som syftar till att öka transparensen kring innehållets ursprung, men deras effektivitet begränsas av hur de påverkar användarens kognitiva resurser. Om en märkning är för visuellt komplex, innehåller överdriven information eller tar upp för mycket plats och drar blicken till sig för mycket, kan det leda till informationsöverbelastning, vilket resulterar i att användare antingen ignorerar märkningen helt eller upplever förvirring kring dess betydelse (Eppler & Mengis, 2004). Istället bör märkningarna utformas minimalistisk och enkelt, men samtidigt tydligt och synligt. Detta är avgörande i informationsrika miljöer där AI-märkningar konkurrerar med bilder, videor och andra interaktiva funktioner om användarens uppmärksamhet (Marcus & Vasser, 2024).

För att illustrera hur AI-märkningar presenteras och utformas på Instagram och TikTok visas nedan exempel från plattformarna:



Figur 1: TikTok-innehåll med AI-märkningen “Contains AI-generated media”



Figur 2: TikTok-innehåll med AI-märkningen “Contains AI-generated media”

Figur 3: Instagram-innehåll med AI-märkningen “AI info”.

Figur 4: Instagram-innehåll med metadata i samma placering som AI-märkningen.

Bilderna ovan visar hur AI-märkningar på TikTok och Instagram integreras på olika sätt. På TikTok placeras märkningen diskret längst ned i innehållet. På Instagram finns märkningen långt upp i innehållet i samma område som annan metadata, i detta fall ljudinformation. Trots att märkningarna är synliga riskerar de att hamna utanför användarens huvudsakliga fokusområde eftersom en användares uppmärksamhet främst riktas mot det centrala visuella innehållet. Utifrån kognitiv belastningsteori blir därför märkningarnas placering, tydlighet och visuella utformning viktiga faktorer för hur effektivt de uppfattas av användare. Därmed kan denna teori användas för att förstå hur AI-märkningar på TikTok och Instagram synliggörs och i vilken utsträckning användare faktiskt uppmärksammar dem i praktiken.

Ytterligare en viktig aspekt för att märkningarna ska uppnå hög transparens är att de måste designas med principen om igenkänning hellre än återkallande (recognition rather than recall). Detta innebär att användaren genom konsekventa visuella signaler snabbt kan igenkänna informationen från tidigare upplevelser och lärdomar, då belastas arbetsminnet inte alls lika mycket (Marcus & Vasser, 2024). Sammanfattningsvis bör märkningarna alltid se likadana ut på en plattform, användaren kan då mycket snabbare identifiera märkningen utan att lägga särskilt mycket energi på den, alltså leder det till låg kognitiv belastning.

En annan viktig aspekt av transparens är konsekvens, i detta fall hur ofta AI-genererat innehåll har märkningar och om korrekt innehåll är märkt. Exempelvis genom inkonsekvent märkning där innehåll saknar märkning men användaren misstänker att innehållet är AI-genererat, ökar användarens kognitiva belastning ytterligare eftersom de tvingas lägga resurser på att själva försöka bedöma autenticiteten.

Sammanfattningsvis påvisar ovan förklaring till att effektiv transparens är en balans där märkningar är tillräckligt synliga för att uppmärksammas, men tillräckligt enkla för att inte bidra till den överflödiga kognitiva belastningen. Dessutom bör märkningar vara så konsekventa och så korrekta som möjligt.

2.3 Sammanfattning och kritisk reflektion över litteraturgenomgången

Utifrån litteraturgenomgången kan AI-märkningar förstås som informationsmarkörer vilka syftar till att skapa transparens kring innehålls ursprung. Tidigare forskning visar att AI-märkningar kan ha en påverkan på hur användare tolkar AI-genererat innehåll i relation till autenticitet, trovärdighet och engagemang. Å andra sidan, visar det även att närvaron av märkningarna inte automatiskt leder till ökat förtroende. Deras effekt beror på hur märkningarna utformas, hur synliga de är, vilken terminologi som används och om de uppfattas som konsekventa.

Den tidigare forskningen presenterar även att AI-märkningar kan få vissa oavsiktliga effekter. Om AI-genererat innehåll är inkonsekvent märkt, kan omärkt innehåll uppfattas som mer autentiskt eller trovärdigt, vilket kan kopplas till implied truth effect. Relevansen av detta för studien då användares förtroende inte enbart påverkas av de märkningar som finns, utan också i de situationer där AI-genererat innehåll saknar märkning. Därför blir både närvaron och frånvaron av AI-märkningar viktiga för att förstå hur transparens och förtroende formas hos användare på Instagram och TikTok.

Det teoretiska ramverket för studien består av Trust in Information Systems, Mayers Trust Model samt kognitiv belastningteori i relation till informationsöverbelastning. Trust in Information Systems är ett persoektiv som används för att förstå hur användare utvecklar förtroende för digitala system samt informationen som presenteras inom systemen. För att fördjupa detta perspektiv, används Mayers Trust Model genom dimensionerna ability, benevolence och integrity. De tre dimensionerna används för att analysera hur användare bedömer plattformarnas förmåga, intentioner och konsekvens i relation till AI-märkningar. Kognitiv belastningsteori och informationsöverbelastning används också som komplettering för förtroendeperspektivet genom att förklara hur märkningarnas placering och synlighet kan påverka om användare faktiskt uppmärksammar och förstår dem.

Å andra sidan har både den tidigare forskningen och det teoretiska ramverket vissa begränsningar. En stor del av forskningen om märkningar behandlar innehållsmärkningar eller generella AI-märkningar, snarare än användare uppfattningar av AI-märkningar på Instagram och TikTok. På grund av detta behöver resultaten från tidigare studier användas med försiktighet i denna kontext. Trust in Information Systems och Mayers Trust Model ger studien en användbar grund för att analysera förtroende, men de förklarar inte fullt ut hur användare uppmärksammar visuella gränssnittselement på sociala medier. Kognitiv belastningsteori används därför då den bidrar med ett sådant perspektiv, men förklarar i sin tur främst informationsbearbetning och uppmärksamhet, snarare än sociala och emotionella aspekter av förtroende.

Mot denna bakgrund möjliggör kombinationen av Trust in Information Systems, Mayers Trust Model och kognitiv belastningsteori både en analys av hur förtroende formas i relation till märkningarna och hur märkningarnas synlighet, tydlighet och utformning påverkar användares möjlighet att uppmärksamma och förstå dem. Den tidigare forskningen har främst fokuserat på AI-märkningars design och användarengagemang, samt generella innehåll märkningar. Däremot finns det mer begränsad forskning om hur användare uppfattar AI-märkningar på etablerade sociala medieplattformar i relation till förtroende och transparens, särskilt i kontexten av Instagram och TikTok. Detta motiverar studiens fokus och visar varför det är relevant att undersöka AI-märkningar ur ett användarperspektiv.

3 Metod och forskningsdesign

I detta kapitel redogörs studiens metodologiska tillvägagångssätt. Först diskuteras studiens forskningsdesign samt motivering till valet av en kvantitativ ansats. Därefter beskrivs datainsamlingen, urvalet och enkätundersökningens utformning. Det redogörs även för hur det insamlade materialet analyserats samt vilka etiska överväganden som har gjorts. Avslutningsvis diskuteras metodvalets styrkor och begränsningar i relation till studiens syfte och forskningsfråga.

3.1 Forskningsdesign

Syftet med denna studie är som tidigare förklarat att få en bild av hur märkningssystem för AI-genererat innehåll uppfattas av användare på Instagram och TikTok och hur det påverkar deras förtroende. För att kunna besvara detta behövs det samlas in en bredare datamängd, därav passar en kvantitativ forskningsdesign på studien mer för denna frågeställning en än en kvalitativ hade gjort. En kvalitativ studie hade kunnat ge en djupare förståelse av enstaka individers uppfattningar, men fokuset i just den här undersökningen ligger snarare på att få en mer översiktlig bild av vad användare upplever för att då kunna identifiera mönster (Lim, 2025).

Då undersökningen är kvantitativ genomfördes datainsamlingen med en enkätundersökning. För att se till att frågeställningen besvaras korrekt och att datan förhåller sig inom de tänkta ramarna, användes modellen som beskrevs i Lim's artikel för att utforma enkäten. Modellen belyser tre övergripande överväganden vid design av kvantitativa undersökningar, dessa är val inom *data source* (källorna datan inhämtas från), *temporality* (tidsperiod som undersökningen sker) och *causality* (vilka faktorer som påverkar data). Genom att förhålla sig till rätt val inom dessa kategorier kan man säkerställa att frågeställningen besvaras. För denna undersökningens frågeställning valdes primary research inom data source då det behövdes data från den nuvarande situationen med AI-märkningar. Föråldrad data är inte relevant till exakt den här undersökningens frågeställning. Inom temporality valdes cross sectional vilket betyder att resultaten mäts vid en viss tidpunkt och inte över en längre period. Detta är mer passande då frågeställningen riktar sig till påverkan av nulägets AI-märkningar och de kan komma att förändras över tid. Inom causality valdes non-experimental vilket betyder att ingen manipulation av variabler kommer att ske för att jämföra konsekvenser och korrelationer. Det är utanför vår frågeställning och skulle innebära en mycket mer omfattande undersökning (Lim, 2025).

Utöver enkäten tar denna studie även hjälp av tidigare forskning om digital mediekonsumtion, algoritmisk transparens och informationens pålitlighet, vilket även ligger till grund för formuleringen av studiens frågeställning.

3.2 Datainsamling

Enkäten består av 18 frågor. De tre första är utformade för att ge en bakgrund om deltagarna, närmare bestämt deras åldersgrupp, hur ofta de använder TikTok respektive Instagram och hur bekanta de är med AI-genererat innehåll.

Sedan är resten av frågorna riktade till deras uppfattningar om AI-genererat innehåll och märkningar. Frågeformuläret fokuserar på följande områden:

- Kännedom om AI-genererat innehåll
- Uppmärksamhet på märkningar
- Trovärdighet och tillit
- Upplevd tydlighet och synlighet av märkningar

Frågorna är utformade på två olika sätt beroende på område och vad för slags svar som söks. Frågorna utformas med hjälp av:

- Binära frågor, i detta fall ja eller nej
- Likertskalor (t.ex. 1–5) för att mäta attityder och uppfattningar

Binära frågor användes för enklare ja och nej svar. Likertskalor valdes då dessa är utmärkta för att kunna få en djupare insikt kring hur respondenterna upplever en aspekt. Det ger upphov till fler nyanser än endast, exempelvis positivt eller negativt. Respondenterna kan därmed svara mer ingående till vilken nivå det är positivt respektive negativt, eller även neutralt.

Den insamlade datan kommer slutligen att analyseras genom deskriptiv statistik och enkel bivariatanalys. Målet är att skapa figurer och genom dessa kunna få en översiktsbild över deltagarnas förtroende och vilka faktorer som påverkar. Detta kommer att avslöja olika trender och mönster som sedan gör resultaten enklare att analysera. En djupare och mer ingående förklaring finns i stycke 3.5 “Analysmetod”.

3.3 Urval

Urvalet består av ett icke-sannolikhetsurval, närmare bestämt ett bekvämlighetsurval (convenience sampling). Detta innebär att deltagare rekryteras baserat på vilka som är tillgängliga och nås lättast för den som utför studien. I detta fall inkluderar det familj, vänner och studenter (främst vid Lunds universitet). Enkäten distribueras digitalt via diverse sociala medieplattformar till urvalsgruppen, vilket möjliggjorde en bred spridning och effektiv datainsamling (Galanis & Katsiroumpa, 2026).

Målgruppen är aktiva användare av Instagram och/eller TikTok, med fokus på yngre vuxna. Denna grupp är vald då de är särskilt aktiva på dessa plattformar och därmed mer exponerade för AI-genererat innehåll. Enkäten kommer därmed även vara enklare för dem att besvara och datan kommer således dessutom att vara mer pålitlig.

Urvalets storlek bör vara tillräckligt stor för att möjliggöra statistisk analys. Målet är att få åtminstone mellan 40–100 respondenter. Det är viktigt att vara medveten om att denna

urvalsmetod kan begränsa studiens generaliserbarhet, men den är lämplig med hänsyn till studiens tidsram och resurser.

3.4 Enkätundersökning

Enkätfrågorna utformades med utgångspunkt i studiens teoretiska ramverk. De frågor som är relaterade till plattformarnas förmåga att korrekt identifiera AI-genererat innehåll kopplades till dimensionen ability i Mayer et al. (1995) modell för förtroende. Frågorna relaterade till plattformarnas intentioner och om AI-märkningar uppfattades vara i användarnas intresse kopplades till benevolence. Vidare kopplades de frågor om märkningarnas konsekvens, trovärdighet och transparens till integrity. Att utforma frågorna på detta sätt möjliggjorde en analys av hur olika aspekter av förtroende uppfattas i relation till AI-märkningar på Instagram och TikTok.

3.5 Analysmetod

Det insamlade materialet kommer att analyseras med hjälp av kvantitativa analysmetoder för att undersöka hur användare uppfattar och uppmärksammar märkningar av AI-genererat innehåll på Instagram och TikTok. Analysen genomförs i flera steg för att skapa både en översiktlig och jämförande förståelse av resultaten samt för att besvara enkätfrågor av olika slag.

Inledningsvis görs en analys av de tre första frågorna då de var till för att skapa en förståelse av respondenternas bakgrund, alltså en uppfattning om vilka deltagarna är och om deras förutsättningar till svaren. Här samlas information om deras ålder, hur ofta de använder TikTok respektive Instagram och hur bekanta de är med AI-genererat innehåll. Dessa frågor är viktiga då deltagare med olika bakgrund, exempelvis olika åldersgrupp, kommer garanterat att svara på frågorna med betydande skillnader på grund av att de äldre generationerna känt inte använder sociala medier i samma utsträckning. Dessa skillnader måste därför tas i åtanke, både när urvalet till respondenterna väljs, samt vid analysen av svaren. Slutsatserna av analysen blir på så vis mer meningsfulla.

Därefter används deskriptiv statistik för att sammanställa respondenternas svar, men även bivariatanalys utfördes för att jämföra skillnader mellan olika respondentgrupper. Resultaten presenteras genom svars fördelningar, procenttal, medelvärden och figurer för att identifiera övergripande mönster och tendenser i materialet. Detta möjliggör en analys av hur användarna uppfattar AI-genererat innehåll, hur ofta de uppmärksammar AI-märkningar samt hur tydliga och tillförlitliga märkningarna uppfattas vara. För frågor som använder Likertskalor omvandlas svaren till numeriska värden för att möjliggöra statistisk bearbetning. Medelvärden används sedan för att analysera respondenternas generella uppfattningar kring exempelvis tydlighet, trovärdighet, transparens och förståelse av AI-märkningar. Resultat som ligger nära skalans mittpunkt tolkas med försiktighet, eftersom svarsfördelningen också behöver beaktas för att avgöra om uppfattningarna faktiskt är neutrala eller om svaren är polariserade. Resultat med högre eller lägre värden indikerar mer positiva respektive negativa uppfattningar.

Analysen syftar främst till att beskriva och jämföra användarnas uppfattningar snarare än att utveckla avancerade statistiska modeller eller göra prediktioner. Därför genomförs inte mer avancerade analyser såsom regressionsanalys eller logistisk regressionsanalys.

Slutligen kommer de sammanställda resultaten att analyseras med hjälp av Mayers Modell. Denna modell är hjälpsam då den består av tre aspekter som när analyserade tillsammans skapar en övergripande förståelse för förtroende hos respondenterna. Modellen valdes eftersom den möjliggör en strukturerad analys av hur olika dimensioner av förtroende relaterar till användares uppfattningar av AI-märkningar.

3.6 Etik

Studien följer forskningsetiska principer genom att förhålla sig till en del aspekter, varav krav på informationsgivning, samtycke, konfidentialitet och ansvarsfull hantering av data. Respondenterna informeras om studiens syfte innan de deltar och deltagandet är frivilligt. Alla svar behandlas anonymt och ingen personlig information som kan identifiera deltagarna samlas in eller publiceras. Det insamlade materialet används sedan enbart för forskningsändamål inom denna undersökning och hanteras konfidentiellt. Detta görs för att skydda deltagarnas integritet och säkerställa att studien genomförs på ett etiskt korrekt sätt (Onwuegbuzie & Leech, 2025).

Utöver dessa grundläggande forskningsetiska principer beaktas även etiska och reflexiva aspekter kopplade till kvantitativ forskning. Eftersom kvantitativa studier bygger på mätning, statistisk analys och tolkning av data är det viktigt att resultaten presenteras på ett rättvist och transparent sätt. Studien eftersträvar därför att minimera risker för bias och snedvridna tolkningar genom att använda tydliga analytiska procedurer och konsekventa bedömningsgrunder. Detta är särskilt viktigt eftersom statistiska resultat kan påverka hur individer eller grupper förstås och representeras (Onwuegbuzie & Leech, 2025).

Vid analysen tas även hänsyn till att metodologiska och analytiska val kan påverka studiens resultat och slutsatser. Val av variabler, tolkning av data genomförs därför med försiktighet och transparens för att stärka studiens trovärdighet och rättvisa. Studien strävar efter att redovisa resultaten på ett sätt som inte överdriver samband eller osäkerheter och som samtidigt ger en så rättvis bild som möjligt av verkligheten.

3.7 Metoddiskussion

Valet av en kvantitativ forskningsdesign har varit för att kunna besvara studiens syfte på bästa möjliga sätt, då målet var att identifiera generella mönster och trender kring hur AI-märkningar uppfattas av en bredare grupp användare. Genom att använda en enkätundersökning kunde vi samla in strukturerad data som möjliggjorde sammanställningar och jämförelser med hjälp av olika slags diagram. En kvalitativ metod hade dock kunnat ge en djupare förståelse för enskilda individers upplevelser, men det hade inte gett den översiktliga bild och identifiera mönster av användarnas uppfattningar som eftersträvades i denna studie.

En aspekt av kvantitativa studier, speciellt just denna, är studiens urval och generaliserbarhet. Användandet av ett bekvämlighetsurval (convenience sampling) var nödvändigt på grund av

studiens tidsram och resurser. Detta innebär dock att urvalet inte är slumpmässigt, vilket begränsar möjligheten att generalisera resultaten till alla användare av TikTok och Instagram. Majoriteten av respondenterna (över 93 %) tillhörde åldersgruppen 15–27 år, vilket gör att studien främst speglar unga vuxnas perspektiv. Även om tanken från början var att fokusera på denna grupp främst då de är dagliga användare av plattformarna (86,4 %), kan resultaten skilja sig markant från de äldre åldersgrupperna (Galanis & Katsiroumpa, 2026).

För att säkerställa studiens kvalitet tas hänsyn till både validitet och reliabilitet. Validitet handlar om att undersökningen ska mäta det som avses att undersökas, vilket i denna studie innebär att enkätfrågorna utformas i enlighet med studiens syfte. Varje fråga för sig själv kommer inte ensamt kunna svara på frågeställningen, men genom att noga ompröva enkätens alla frågor och justera ordvalet, går det att säkerhetsställa att frågorna gemensamt täcker hela frågeställningen. Reliabilitet handlar om undersökningens tillförlitlighet och att resultaten ska kunna bli liknande vid en upprepning av studien. För att stärka reliabiliteten används tydligt formulerade frågor och standardiserade svarsalternativ, exempelvis Likertskalor. Dock finns det även brister med reliabiliteten i denna undersökning på grund av antalet respondenter. Ett större urval hade stärkt studiens reliabilitet. Denna undersökning eftersträvade mellan 40 och 100 deltagare, vilket kan anses vara en mindre respondentgrupp, speciellt då antalet respondenter för denna undersökning blev närmare 40. En mindre grupp respondenter innebär en högre risk för slumpmässiga variationer i datan, vilket kan påverka resultatets stabilitet vid en eventuell upprepning av studien (Fendler, 2016).

Slutligen bör det noteras att studien data är användarnas självrapporterade uppfattningar. Det finns en risk att respondenterna överskattar sin förmåga att upptäcka AI-genererat innehåll eller märkningar. Det är i princip omöjligt att verifiera om det innehåll de någonsin sett faktiskt var AI-genererat eller inte. Ett experimentellt upplägg där användare faktiskt får interagera med olika märkningar i realtid hade kunnat ge ytterligare precision kring märkningarnas faktiska synlighet men det är en större studie utöver denna studiens resurser.

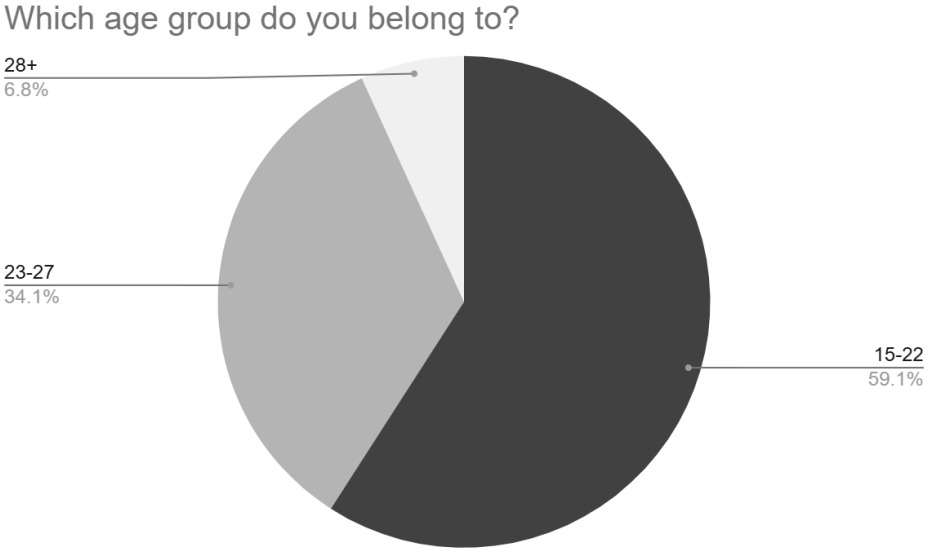
4 Empiri

I detta avsnitt kommer resultaten från enkätundersökningen att presenteras. Resultaten kommer att redovisas genom deskriptiv statistik i form av svarsfördelningar, procenttal, medelvärden och figurer. Utöver den deskriptiva statistiken genomförs en enkel bivariatanalys i form av jämförelser mellan två åldersgrupper, 15–22 år och 23–27 år.

Enkätundersökningen besvarades av totalt 44 respondenter. Samtliga 44 respondenter besvarade enkätens huvudsakliga frågor. I frågan om hur omärkt AI-genererat innehåll har påverkat respondenternas förtroende deltog 42 respondenter, eftersom denna fråga riktades specifikt till dem som uppgav att de tidigare sett AI-genererat innehåll utan märkning. För de frågor som bygger på likertskalor kommer resultaten att tolkas både utifrån medelvärden och svarsfördelning, då ett medelvärde nära skalans mittpunkt inte nödvändigtvis betyder att respondenterna är neutrala.

4.1 Respondenternas bakgrund

För att skapa en överblick över deltagarna presenteras respondenternas bakgrund utifrån ålder och användningsfrekvens av Instagram och TikTok.



Figur 5: Fördelning av respondenternas åldersgrupper.

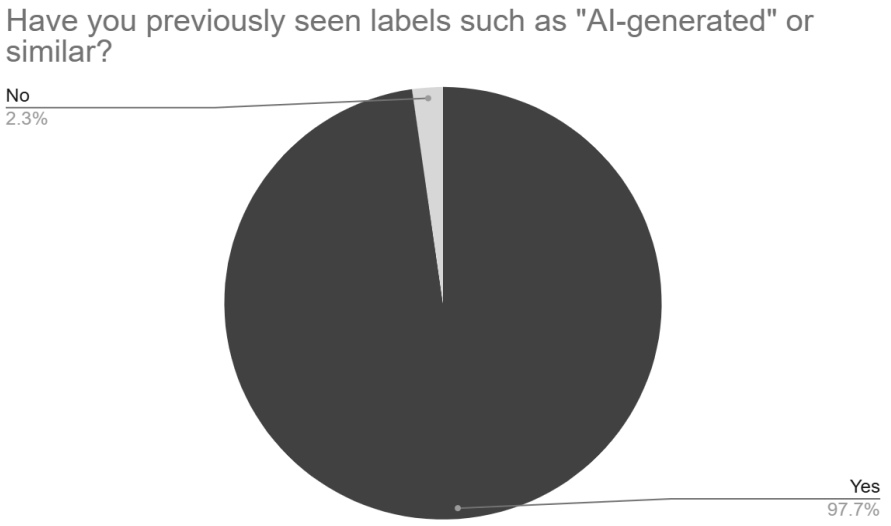
Majoriteten av respondenterna befann sig i åldersgruppen mellan 15-22 (59,1%) och 23-27 (34,1%).

De flesta respondenter (86,4%) svarade att de använder Instagram och TikTok dagligen.

4.2 Kännedom om AI-genererat innehåll och märkningar

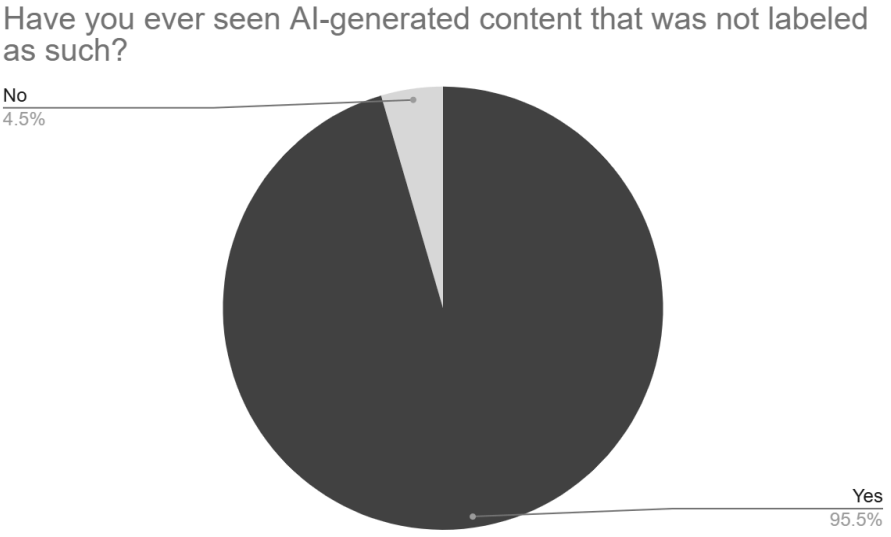
För att undersöka respondenternas erfarenhet av AI-genererat innehåll och märkningar ställdes frågor om kännedom och exponering. Över 68% av respondenterna svarade att de är väldigt bekanta med AI-genererat innehåll, och 29,5% svarade att de är relativt bekanta.

Sedan ställdes frågan om man tidigare har sett AI-märkningar, där majoriteten (97,7%) av respondenterna uppgav att de tidigare sett märkningar såsom “AI-generated” på Instagram och TikTok. (Figur 6)



Figur 6: Respondenternas kännedom om AI-genererat innehåll.

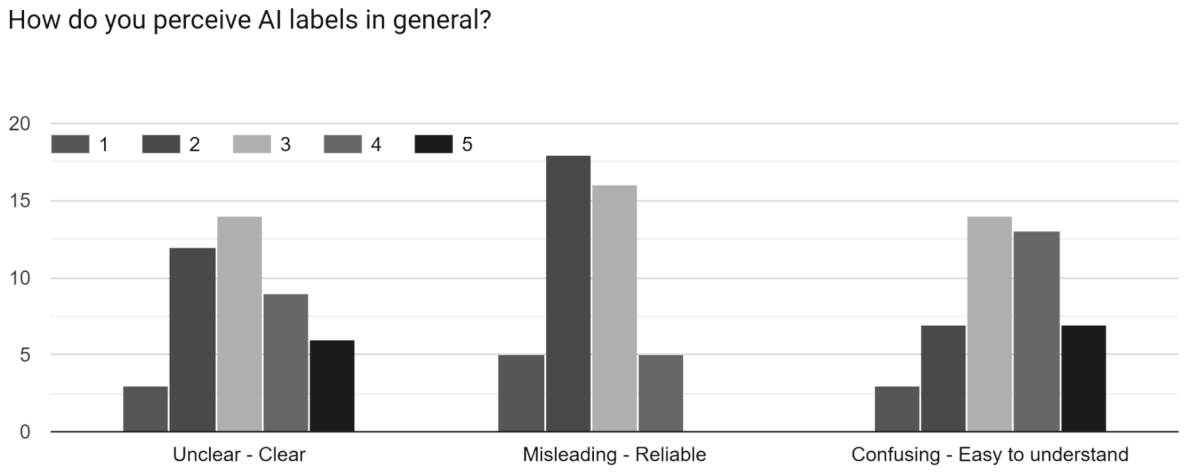
Samtidigt svarade 95,5% av respondenter att de även har sett AI-genererat innehåll som inte har varit märkt som sådant. (Figur 7)



Figur 7: Respondenternas erfarenhet av omärkt AI-genererat innehåll.

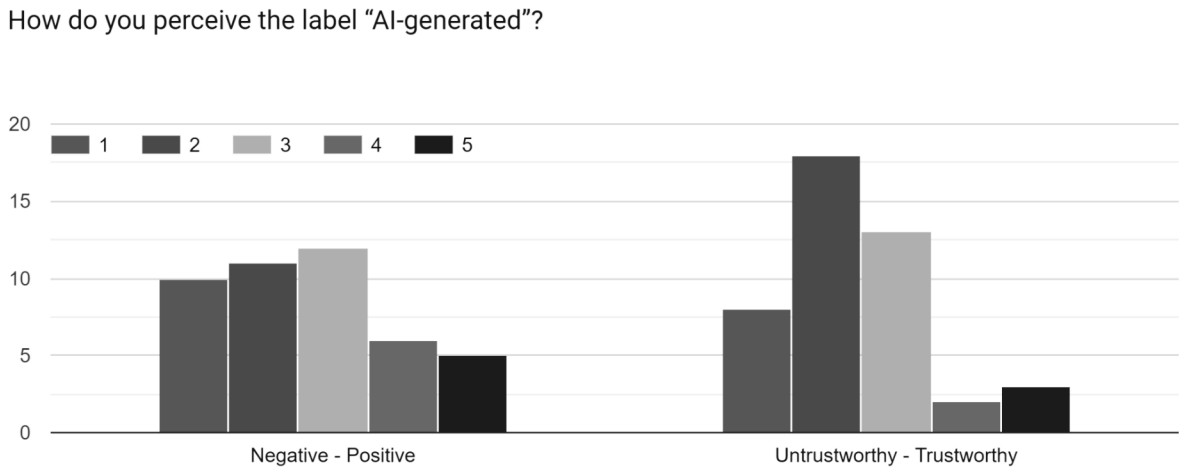
4.3 Uppfattning av AI-märkning

Enkäten inkluderade även frågor relaterade till hur AI-märkningar uppfattas på Instagram och TikTok, där fokus låg på märkningarnas tydlighet och förståelse.



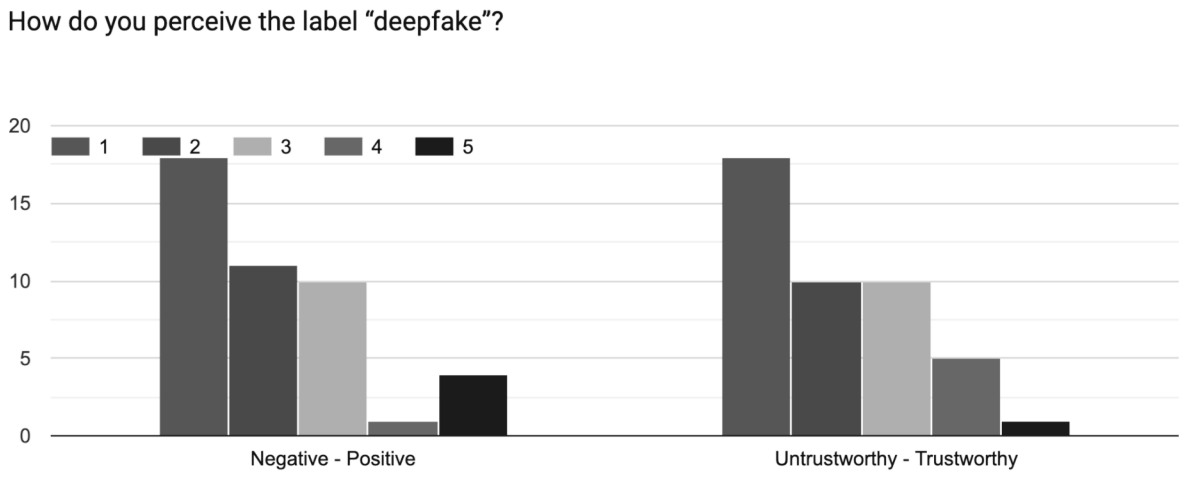
Figur 8: Respondenternas uppfattning av AI-märkningars tydlighet, tillförlitlighet och begriplighet.

Figur 8 visar hur respondenterna uppfattar AI-märkningar. Medelvärdet av första svarsalternativet på skalan "Unclear-Clear" är ≈ 3.07. Figuren visar också hur respondenterna uppfattar AI-märkningars tillförlitlighet. Medelvärdet här ligger på ≈ 2.47, på skalan "Misleading-Reliable". Sista svarsalternativet är på skalan "Confusing-Easy to Understand", och Figur 8 visar hur respondenterna uppfattar förståelsen av AI-märkningar, där medelvärdet är ≈ 3.31.



Figur 9: Respondenternas uppfattning av begreppet "AI-generated".

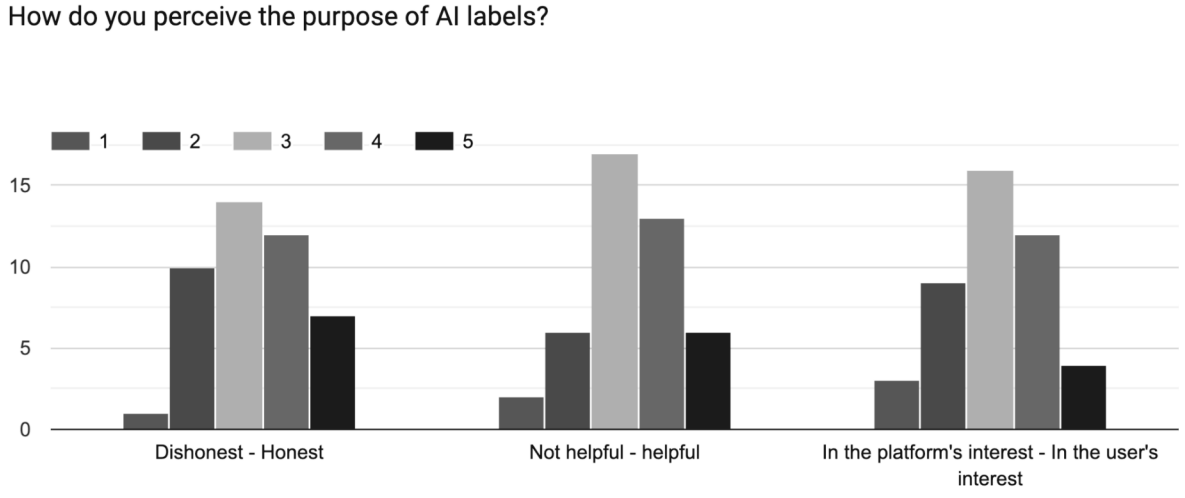
Figur 9 visar hur respondenterna uppfattar märkningen som använder begreppet "AI-generated". Medelvärdet för svarsalternativet på skalan "Negative-Positive" är ≈ 3.29.



Figur 10: Respondenternas uppfattning av begreppet "deepfake".

Figur 10 visar hur respondenterna uppfattar märkning som innehåller begreppet "Deepfake". Här är medelvärdet för svarsalternativet på skalan "Negative-Positive" ≈ 2.13. Samtidigt visar figuren svaren på skalan "Untrustworthy- Trustworthy", där medelvärdet är ≈ 2.11.

Sedan ställdes frågan om respondenternas uppfattning av AI-märkningars syfte.

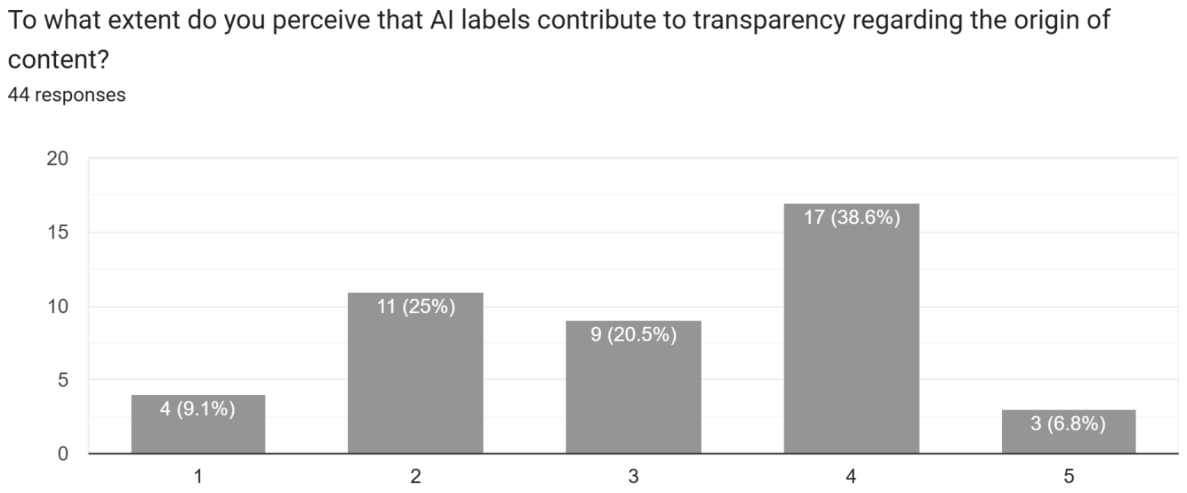


Figur 11: Respondenternas uppfattning av AI-märkningars ärlighet, hjälpsamhet och användarintresse.

Figur 11 visar hur respondenterna uppfattar AI-märkningar i relation till ärlighet, hjälpsamhet och intresse. Medelvärdet för svarsalternativet på skalan "Dishonest-Honest" är ≈ 3.31, och ≈ 3.34 på skalan "Not helpful-Helpful". Sista delen av frågan handlar om respondenternas uppfattning av intresse, där medelvärdet för svarsalternativet på skalan "In the platform's best interest-In the user's interest" ligger runt 3.11.

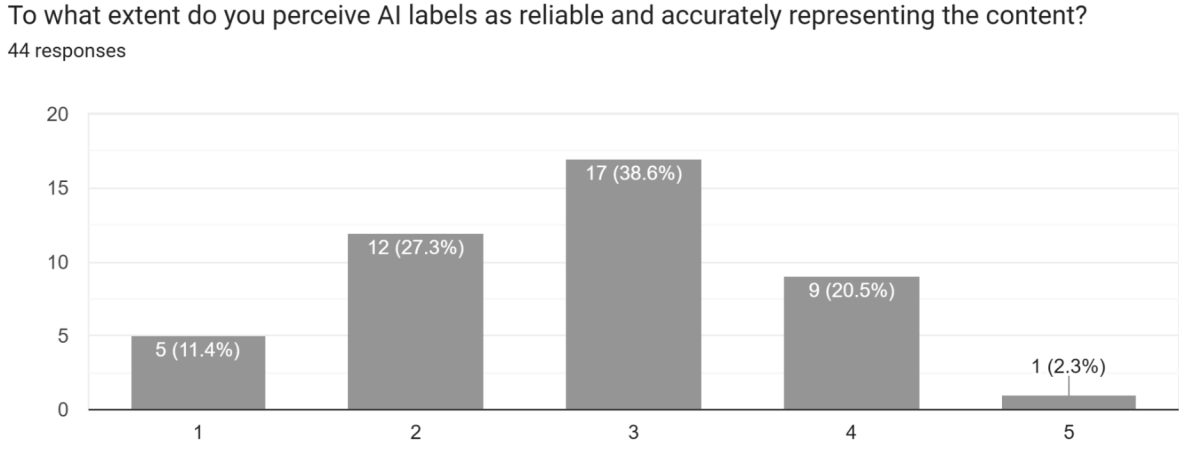
4.4 Förtroende och transparens

Nästa del av undersökningen fokuserade på hur AI-märkning påverkar användarnas uppfattning av transparens och förtroende på Instagram och TikTok.



Figur 12: Respondenternas uppfattning av AI-märkningars bidrag till transparens.

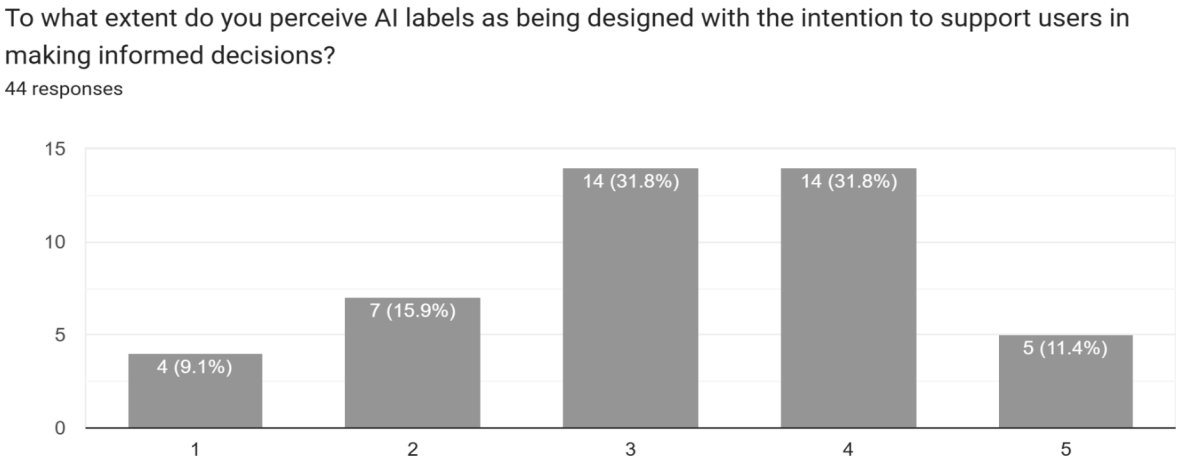
Figur 12 visar hur respondenterna uppfattar AI-märkningars bidrag till transparens angående innehållets ursprung. Medelvärdet för svarsalternativet är ≈ 3.09, vilket ligger nära mittpunkten på skalan. Trots detta, visar svarsfördelningen att respondenterna inte var helt neutrala då den största andelen svar ligger på den positiva sidan av skalan. Nästa fråga handlar om hur respondenterna uppfattar AI-märkningars påverkan på förtroendet förplattformarna, och medelvärdet är exakt likadant.



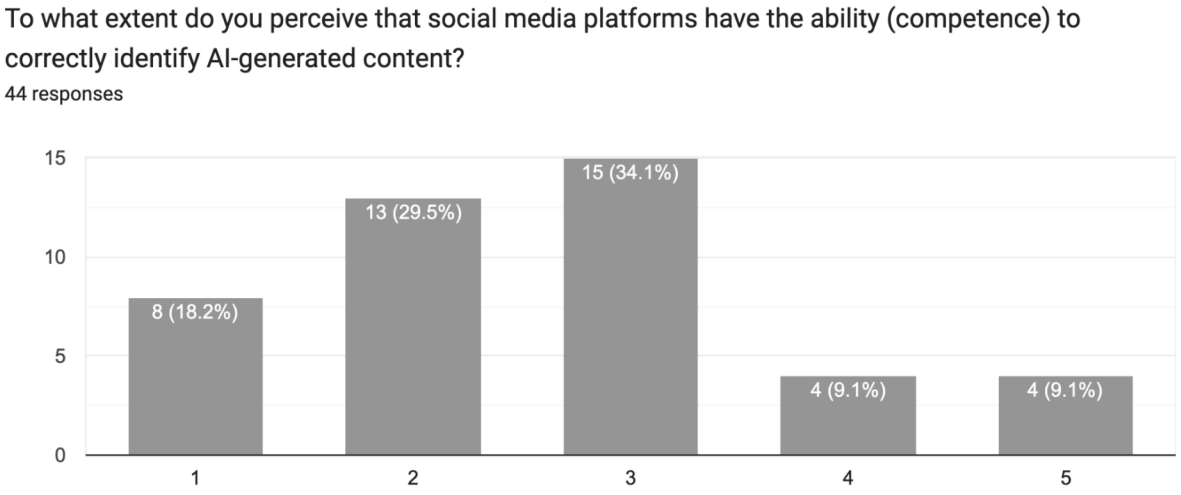
Figur 13: Respondenternas uppfattning av AI-märkningars tillförlitlighet och korrekthet.

Sedan frågades respondenterna om hur de uppfattar AI-märkningars förmåga att korrekt representera innehållet (Figur 13). Medelvärdet för svarsalternativet är = 2.75, som ligger lite under mittpunkten.

Nästa fråga handlade om respondenternas uppfattning av plattformarnas intentioner med AI-märkningar (Figur 14), där medelvärdet är ≈ 3.20.

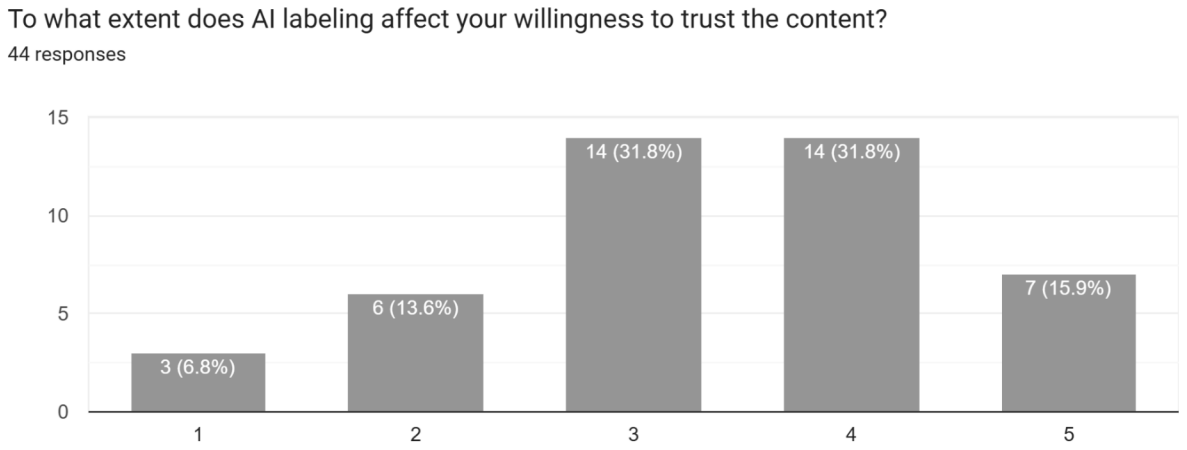


Figur 14: Respondenternas uppfattning av AI-märkningars stöd för informerade beslut.



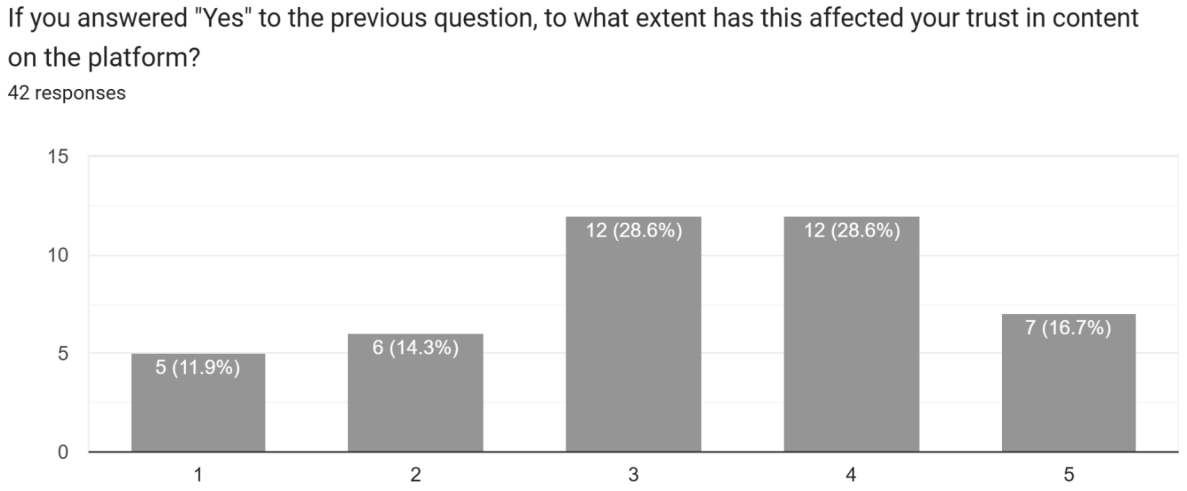
Figur 15: Respondenternas uppfattning av plattformars förmåga att identifiera AI-genererat innehåll.

Figur 15 visar hur respondenterna uppfattar plattformarnas förmåga att korrekt identifiera AI-genererat innehåll, med medelvärdet på ≈ 2.61.



Figur 16: AI-märkningars påverkan på respondenternas vilja att lita på innehåll.

Respondenterna svarade att AI-märkningar i viss utsträckning positivt påverkar deras vilja att lita på innehållet, eftersom medelvärdet är ≈ 3.36. (Figur 16)



Figur 17: Respondenternas förtroende för plattformsinnehåll efter exponering för omärkt AI-innehåll.

Nästa fråga handlade om liknande, fast den fokuserade på vilja att dela innehållet med andra. Medelvärdet här ligger under mittpunkten (≈ 2.77). Sedan fick respondenter som tidigare uppgett att de har sett AI-genererat innehåll utan märkning svara på i vilken utsträckning detta påverkat deras förtroende för innehåll på plattformen (Figur 17), och här visas medelvärdet av svaren vara ≈ 3.24.

4.5 Jämförelse mellan respondentgrupper

För att identifiera eventuella skillnader mellan respondentgrupper genomfördes en enkel jämförelse mellan åldersgrupperna 15-22 år och 23-27 år. De respondenter som tillhörde 28 år och äldre inkluderades inte i jämförelsen eftersom den gruppen bestod av ett begränsat antal

deltagare. Detta skulle inte ge en rättvis bild av resultaten då det hade varit svårt att dra jämförbara slutsatser utifrån en så liten respondentgrupp.

Resultaten visar att åldersgruppen 15-22 generellt uppfattade AI-märkningar som något lättare att förstå jämfört med respondenter i åldersgruppen 23-27. Medelvärdet för förståelse av AI-märkningar låg på ≈ 3.38 för gruppen 15–22 och ≈ 3.13 för gruppen 23–27.

Ytterligare en mindre skillnad mellan respondentgrupperna identifierades i svaren kring hur AI-märkningar påverkar respondenternas vilja att lita på innehåll. Respondenter i åldersgruppen 15–22 uppgav i något större utsträckning att AI-märkningar påverkar deras förtroende för innehåll jämfört med gruppen 23–27. Skillnaderna mellan grupperna var dock relativt små.

Resultaten visar även att båda åldersgrupperna uppfattade begreppet “deepfake” betydligt mer negativt än begreppet “AI-generated”. Medelvärdet för begreppet “deepfake” på skalan “Negative-Positive” låg på ≈ 2.31 för gruppen 15–22 och ≈ 1.80 för gruppen 23–27. På skalan “Untrustworthy-Trustworthy” låg medelvärdet på ≈ 2.27 för gruppen 15–22 och ≈ 1.87 för gruppen 23–27. Dessa resultaten visar därmed att den äldre åldersgruppen uppfattade begreppet mer negativt och mindre trovärdigt jämfört med den yngre gruppen.

5. Diskussion

I detta kapitel diskuteras och analyseras studiens empiriska resultat i relation till forskningsfrågan, det teoretiska ramverket och tidigare forskning. Fokus ligger på hur användare uppfattar AI-märkningar på Instagram och TikTok i relation till förtroende och transparens. Det presenteras även de praktiska implikationerna och begränsningarna av studien samt förslag på framtida forskning.

Resultaten från enkätundersökningen visar att majoriteten av respondenterna är väl bekanta med AI-genererat innehåll och AI-märkningar på Instagram och TikTok. Samtidigt svarade nästan samtliga respondenter att de även hade observerat AI-innehåll utan märkning, vilket väcker frågor om märkningarnas konsekvens och tillförlitlighet. Respondenterna uppfattade generellt AI-märkningar som relativt förståeliga, men inte nödvändigtvis tydliga eller helt tillförlitliga. Resultaten visar även att olika begrepp associeras på olika sätt, där det finns en tydlig skillnad i uppfattningar av "AI-generated" och "deepfake". Vidare visar resultaten att AI-märkningar till viss del bidrar till transparens och påverkar användarnas förtroende för innehåll, men att de inte tydligt stärker förtroendet för själva plattformarna. Respondenterna uttryckte även viss skepticism kring plattformarnas förmåga att korrekt identifiera AI-genererat innehåll. Dessutom uppfattades avsaknaden av märkningar i viss utsträckning påverka användarnas förtroende för innehållet som presenteras. Sammantaget tyder detta på att märkningars synlighet, tydlighet, tillförlitlighet och konsekvens har betydelse för hur innehåll uppfattas och tolkas av användare.

5.1 AI-märkningar som transparensmekanism

Ett centralt resultat i studien är att respondenterna generellt uppfattar AI-märkningar som ett positivt initiativ för att skapa transparens kring innehållets ursprung. Detta ligger i linje med tidigare forskning som beskriver AI-märkningar som transparensmekanismer vars syfte är att ge användare information om hur innehåll har producerats och i vilken utsträckning artificiell intelligens har varit involverad (J. Chen et al., 2025). Jung et al. (2025) visar samtidigt att användares reaktioner på AI-märkningar påverkas av hur AI-användningen kommuniceras, till exempel om innehåll markeras som AI-genererat eller AI-modifierat. Detta är relevant för studiens resultat eftersom det visar att AI-märkningar inte enbart fungerar som neutrala informationsmarkörer, utan också påverkar hur användare tolkar innehållets ursprung och graden av transparens. AI-märkningar kan därmed förstås som ett försök från plattformarna att minska osäkerheten kring innehållets autenticitet och ge användare bättre förutsättningar att tolka information.

Samtidigt visar resultaten att märkningarna inte automatiskt leder till ett ökat förtroende för plattformarna. Respondenternas svar låg ofta nära mittpunkten på skalan, vilket indikerar att märkningarna visserligen uppfattas som användbara men inte tillräckligt starka för att förändra användarnas övergripande förtroende. Detta är intressant eftersom det visar att transparens och förtroende inte är samma sak. En plattform kan vara transparent genom att informera användarna om att AI har använts, utan att användarna för den skull upplever att plattformen är mer trovärdig.

Utifrån Trust in Information Systems kan detta förstås som att användare inte endast bedömer förekomsten av information utan också kvaliteten på informationen och systemets förmåga att förmedla den på ett konsekvent sätt (Li & Yang, 2024). Transparens fungerar därför som en grundläggande förutsättning för förtroende, men den är inte tillräcklig i sig själv. För att förtroende ska utvecklas krävs även att användare uppfattar informationen som korrekt, pålitlig och konsekvent över tid.

Studien visar därmed att AI-märkningarnas funktion sträcker sig bortom att endast informera användare. Märkningarna fungerar samtidigt som signaler om hur plattformarna hanterar AI-genererat innehåll och hur stort ansvar de tar för informationsmiljön. När användare upplever att märkningarna fungerar väl kan detta bidra till upplevelsen av transparens. När märkningarna däremot upplevs som inkonsekventa eller otillräckliga riskerar de istället att skapa osäkerhet kring plattformarnas trovärdighet (Li & Yang, 2024) .

5.2 Terminologi som en förutsättning för transparens

Resultaten visar tydligt att användarnas uppfattningar påverkas av vilken terminologi som används i AI-märkningar. Begreppet ”deepfake” associerades i högre grad med manipulation, desinformation och opålitlighet, medan ”AI-generated” uppfattades som mer neutralt och mindre negativt laddat. Detta tyder på att användare inte endast reagerar på informationen i märkningen utan även på de associationer som olika begrepp väcker.

Detta kan förstås i relation till Epstein et al. (2023) som visar att olika utformade AI-märkningar kan framkalla olika nivåer av skepticism hos användare. Terminologi fungerar på så sätt som en tolkningsram genom vilken användare bedömer innehåll. Begrepp som ”deepfake” är ofta kopplade till manipulation och vilseledande innehåll, medan “AI-generated” framstår som ett mer tekniskt och deskriptivt begrepp (Epstein et al., 2023). När användare möter dessa begrepp kan tidigare erfarenheter och föreställningar aktiveras, vilket påverkar hur innehållet uppfattas. Detta kan förklara varför “deepfake” i studien associerades med mer negativa och mindre tillförlitliga egenskaper, medan “AI-generated” uppfattades som mer neutralt. Detta visar samtidigt att terminologi inte enbart påverkar innehållets trovärdighet utan också märkningarnas förmåga att skapa transparens. Om användare tolkar olika begrepp på olika sätt riskerar märkningarnas funktion att bli oklar, vilket kan försvåra användarnas förståelse av vad märkningen faktiskt kommunicerar. I relation till Trust in Information Systems innebär detta att informationselement i ett digitalt system behöver uppfattas som tydliga, begripliga och tillförlitliga för att kunna stödja användares förtroendebedömningar.

Terminologin för AI-märkningar kan också förstås i relation till kognitiv belastning. Om en AI-märkning innehåller begrepp som är otydliga, tvetydiga eller starkt laddade kan detta leda till att användare behöver lägga mer mental energi på att tolka vad märkningen innebär (Marcus & Vasser, 2024). I sociala medieflöden där användare snabbt skannar innehåll kan detta minska märkningens effektivitet som transparensverktyg, särskilt om märkningen inte är tillräckligt tydlig eller lätt att förstå (Gamage et al., 2025). Studien indikerar därmed att terminologin är en central del av märkningarnas design. För att märkningarna ska fungera som effektiva transparensmekanismer behöver användarna förstå dem på ett relativt enhetligt sätt. Annars riskerar märkningarna att skapa nya tolkningsproblem istället för att minska osäkerheten kring innehållets ursprung.

5.3 AI-märkningar i relation till användares förtroende

En stor andel av undersökningen handlar om hur AI-märkningar uppfattas i relation till förtroende och transparens. Resultaten visar att märkningarnas påverkan på användares "trust" för innehåll och plattformar är relativt neutralt till svagt positivt, men samtidigt inkonsekvent beroende på vilken aspekt undersöks. Flera medelvärden ligger nära mittpunkten, vilket tyder på att AI-märkningar inte har en starkt polariserande effekt på användarnas förtroende, utan snarare en mer situationsberoende påverkan.

Samtidigt indikerar resultaten att AI-märkningar i viss utsträckning bidrar till upplevelsen av transparens, eftersom respondenterna svarade varken starkt positivt eller negativt, utan mer neutralt. Utifrån ett analysperspektiv kan detta tolkas som att transparens inte nödvändigtvis översätts direkt till ökat förtroende, samtidigt som förtroende inte enbart formas av närvaron av transparensmekanismer, utan också av hur de uppfattas i relation till tidigare erfarenheter på plattformen.

Ett av studiens tydligaste resultat är att nästan samtliga respondenter hade sett AI-genererat innehåll som saknade märkning, samtidigt som de flesta också hade sett AI-märkningar på plattformarna. Detta kopplades vidare till användarnas förtroende för innehåll på plattformarna, där resultaten visade att användarnas förtroende inte endast påverkas av närvaron av AI-märkningar, utan även situationer där märkning saknas trots att innehållet uppfattas som AI-genererat. Ur ett analysperspektiv kan detta skapa osäkerhet kring vilka typer av innehåll som faktiskt är autentiska och vilka som är AI-genererade. Detta påverkar användarnas upplevelse av förtroende och transparens. Resultatet kan även förstås i relation till Li et al. (2026), som visar att AI-märkningar kan bidra till en mer generell skepticism mot digitalt innehåll. När användare både exponeras för märkta och misstänkt omärkta AI-inlägg kan osäkerheten förstärkas, vilket kan leda till att de blir mer försiktiga i sina bedömningar av innehåll på sociala medier.

Detta kan vidare kopplas till "implied truth effect", där Wittenberg et al. (2023) förklarar att avsaknaden av märkning leder till att användare automatiskt antar att omärkt innehåll är autentiskt eller mer trovärdigt. AI-genererat innehåll som förekommer utan märkning kan därför skapa en mer generell skepticism mot innehållet som presenteras på Instagram och TikTok.

Utifrån resultaten kan användares förtroende förstås som starkt kopplat till hur konsekvent AI-märkningar implementeras i praktiken. Märkningarna i viss utsträckning uppfattas som hjälpsamma och transparensskapande, men deras effekt verkar ändå begränsad när användare upplever att märkningarna inte alltid är korrekta eller konsekventa. Detta indikerar att AI-märkning inte fungerar som en isolerad funktion, utan som en del av större informationsmiljö där både märkt och omärkt innehåll påverkar uppfattningen av trovärdighet.

Därför visar analysen att förtroendet hos användare formas genom samspelet mellan konsekvens, transparens och tidigare erfarenhet på plattformarna. Samtidigt verkar AI-märkningar bidra till ökad medvetenhet kring innehållets autenticitet, men att inkonsekvent implementering skapar osäkerhet kring vad användare faktiskt kan lita på i Instagram och TikTok.

5.4 AI-märkningars synlighet i informationsrika miljöer

Utöver terminologi och förtroende påverkas AI-märkningars effektivitet även av hur märkningarna visuellt integreras i plattformarnas gränssnitt. Detta kan förstås genom kognitiv belastningsteori, där användares begränsade kognitiva resurser påverkar hur information bearbetas i digitala miljöer (Paas et al., 2003; Marcus & Vasser, 2024). På Instagram och TikTok exponeras användare för stora mängder visuellt innehåll samtidigt, vilket kan förstås i relation till informationsöverbelastning. När många informationsobjekt konkurrerar om användarens uppmärksamhet ökar risken för att enskilda element, såsom AI-märkningar, förbises eller inte fullt bearbetas (Eppler & Mengis, 2004). Därför blir AI-märkningars placering och visuella utformning central för uppmärksammande av användare.

Utifrån frågan om hur AI-märkningar uppfattades generellt, tyder resultaten på att respondenterna uppfattar märkningarna som relativt förståeliga, men inte nödvändigtvis tydliga. Detta kan analyseras utifrån Marcus och Vasser (2024), som menar att användare i digitala miljöer ofta snabbt skannar information istället för att aktivt bearbeta allt i gränssnittet i detalj. AI-märkningar som placeras diskret i gränssnittet riskerar därför att hamna utanför användarens huvudsakliga fokusområde och därför kan gå osedda, detta kan även kopplas till kognitiv belastningsteori. Ofta uppmärksammar användare inte märkningarna aktivt eftersom deras uppmärksamhet främst riktas mot det centrala innehållet. AI-märkningar konkurrerar med med bilder, videor, kommentarer, musik och andra visuella element om användarnas uppmärksamhet.Tidigare forskning av Gamage et al. (2025) visar å andra sidan att alltför påträngande märkningar kan påverka användarupplevelsen negativt.

Denna analys visar att AI-märkningars effektivitet inte endast påverkas av närvaron av AI-märkningar, utan även av hur de integreras visuellt i plattformarnas gränssnitt. För att märkningarna ska uppfattas och bidra till ökad transparens behöver de vara tillräckligt synliga för att uppmärksammas, samtidigt som de inte får vara för påträngande så de inte upplevs som störande eller överbelastande i användarupplevelsen.

5.5 Plattformarnas ability, benevolence och integrity

Resultaten kan vidare förstås genom Trust in Information Systems, eftersom AI-märkningar fungerar som informationselement i Instagram och TikToks gränssnitt. Användarnas förtroende påverkas därmed inte enbart av innehållet i sig, utan även av hur plattformarna presenterar, förklarar och tillämpar information om AI-genererat innehåll. För att fördjupa denna förståelse används Mayer, Davis och Schoormans (1995) tre förtroendedimensioner: ability, benevolence och integrity. Dessa dimensioner bidrar till en djupare förståelse av varför AI-märkningar inte automatiskt leder till förtroende trots att de uppfattas som transparensskapande.

Ability handlar om användarnas uppfattning av plattformarnas kompetens. Resultaten visar att många respondenter uttrycker viss skepticism kring plattformarnas förmåga att korrekt identifiera AI-genererat innehåll. När användare upplever att AI-innehåll ibland saknar märkning uppstår tvivel kring plattformarnas tekniska kapacitet. Detta påverkar inte bara förtroendet för märkningarna utan även förtroendet för plattformarnas övergripande informationshantering.

Benevolence handlar om uppfattningen att plattformarna agerar i användarnas intresse. Resultaten visar att användarna i viss utsträckning uppfattar AI-märkningar som hjälpsamma och avsedda att stödja informerade beslut. Samtidigt tyder svaren på en osäkerhet kring plattformarnas motiv. Vissa användare kan uppfatta märkningarna som ett uttryck för ansvarstagande, medan andra kan se dem som ett sätt för plattformarna att uppfylla regulatoriska krav eller skydda sitt eget rykte.

Integrity handlar om uppfattningen av plattformarnas ärlighet och konsekvens. Det är inom denna dimension som studiens resultat framstår som starkast. När användare upplever att AI-innehåll ibland märks och ibland inte märks riskerar plattformarnas integritet att ifrågasättas. Även om märkningarna i sig uppfattas som positiva kan deras värde minska om användarna inte upplever att de tillämpas konsekvent.

Tillsammans visar dessa tre dimensioner att användarnas förtroende formas av fler faktorer än själva märkningens existens. AI-märkningar fungerar som signaler om plattformarnas kompetens, intentioner och konsekvens, vilket innebär att användarnas bedömning av märkningarna samtidigt blir en bedömning av plattformarna som informationssystem.

5.6 Praktiska implikationer

Studiens resultat visar flera praktiska implikationer för sociala medieplattformarna Instagram och TikTok, kring hur i AI-märkningar bör utformas och implementeras för att uppfattas som mer tydliga, trovärdiga och användbara av användare.

Resultaten visar att respondenterna generellt uppfattar AI-märkningar som relativt förståeliga, men samtidigt uttrycker skepticism kring märkningarnas tillförlitlighet och plattformarnas förmåga att korrekt identifiera AI-genererat innehåll. Detta tyder på att märkningar i sig inte är tillräckliga för att skapa tillit, utan att användare även behöver uppleva att plattformarna är mer konsekventa och korrekta i hur AI-genererat innehåll identifieras och märks. För att stärka användarnas förtroende behöver plattformar se till att bli bättre på att identifiera AI-genererat innehåll för att då kunna märka det.

En praktisk implikation är vikten av tydlig och konsekvent terminologi. Studien visar att begreppet “AI-generated” uppfattades betydligt mer neutralt och positivt jämfört med “deepfake”, som associerades med negativa och opålitliga egenskaper. Detta innebär att val av språk och formulering i AI-märkningar kan påverka hur användare tolkar innehållet och därmed även deras förtroende för plattformen. Plattformar bör därför använda tydliga och neutrala begrepp som minskar risken för feltolkning och onödig oro hos användarna.

En annan praktisk implikation är synlighet och placering. Resultaten visar att märkningarnas synlighet och placering är centrala faktorer för hur de uppmärksammas. Eftersom användare ofta snabbt skannar innehåll riskerar diskreta märkningar att förbises, samtidigt som alltför påträngande märkningar kan försämra användarupplevelsen. Plattformar behöver därför hitta en balans mellan synlighet och användbarhet genom att utforma märkningar som är tillräckligt tydliga för att uppmärksammas utan att störa konsumtionen av innehållet.

Slutligen kan resultaten vara relevanta även för organisationer som arbetar med AI-reglering och digital transparens. Studien visar att användarperspektivet är centralt vid utvecklingen av AI-relaterade märkningar och liknande. Därför bör framtida riktlinjer och regleringar inte

endast fokusera på att märkningar implementeras, utan även på hur de utformas och kommuniceras då det påverkar hur användarna förhåller sig till innehållet.

5.7 Begränsningar

Vid tolkning av resultaten finns ett antal begränsningar som bör beaktas. Den första begränsningen är att de respondenter som besvarade enkäten bygger på ett bekvämlighetsurval. Detta är då respondenterna rekryterades främst genom personliga kontakter och studenter vid Lunds Universitet. Detta påverkar vem resultaten kan generaliseras till. Urvalet är därför inte representativt för alla användare av Instagram och TikTok, vilket begränsar möjligheten att generalisera resultaten till en bredare population.

Kopplat till urval är den andra begränsningen, den lilla respondentgruppen på 44 personer. Resultaten ger i sig indikationer på hur användare uppfattar AI-märkningar, men ett större urval hade kunnat bidra till mer tillförlitliga och generaliserbara resultat. Dessutom tillhör majoriteten av respondenterna yngre åldersgrupper som också använder sociala medier dagligen vilket har en direkt påverkan på resultaten och perspektiven som framkom i studien.

Denna studie baseras endast på självrapporterande uppfattningar utifrån en enkätundersökning. Resultaten bygger därför på hur respondenter själva upplever AI-märkningar, inte hur de faktiskt agerar eller uppmärksammar märkningarna i praktiken. Detta innebär att det finns möjlighet till att respondenternas svar skiljer sig från deras faktiska beteenden vid användning av Instagram och TikTok.

Ytterligare en begränsning är att studien fokuserar enbart på Instagram och TikTok. Resultaten kan därför inte överföras till andra digitala plattformar eller sociala medier där AI-märkningar troligtvis är utformade och implementerade på andra sätt.

5.8 Framtida forskning

Framtida forskning inom detta området skulle kunna vara att undersöka hur AI-märkningar uppfattas i större och mer varierade urval. Detta hade skapat en bredare förståelse för användares uppfattningar kring AI-märkningar. Eftersom denna studie främst inkluderade yngre användare som använder sociala medier dagligen, hade det varit relevant att även undersöka hur andra åldersgrupper och användargrupper uppfattar AI-märkningar på digitala plattformar.

Även skulle framtida studier inkludera fler sociala medieplattformar än endast Instagram och TikTok. Då AI-märkningar implementeras och utformas på olika sätt beroende på vilken plattform de förekommer på, med varierande design, funktioner och policyer, innebär det att användares uppfattningar hade skilts sig åt. Det hade därför varit aktuellt och relevant att jämföra AI-märkningar på exempelvis X, Youtube eller Facebook.

Ytterligare en möjlig riktning i framtiden hade varit att undersöka användares faktiskt beteenden till AI-märkningar, snarare än bara självrapporterande uppfattningar. Det hade kunnats utföras eye-tracking-studier för att analysera i vilken utsträckning användare faktiskt lägger märke till AI-märkningar på de olika digitala plattformarna.

Det hade även varit relevant att undersöka hur olika typer av terminologi, design och placeringar påverkar användares förtroende och förståelse av AI-genererat innehåll. Eftersom resultaten i denna studie visar att olika begrepp väcker olika associationer hos användare, finns det ett behov av ytterligare forskning kring hur AI-märkningar kan utformas för att uppfattas som tydliga, trovärdiga och användarvänliga.

6. Slutsats

Syftet med denna studie var att undersöka hur AI-märkningar uppfattas på Instagram och TikTok i relation till transparens och förtroende. Studien visar att AI-märkningar uppfattas som delvis transparensskapande, men att deras förtroendeskapande effekt är begränsad. Respondenterna var generellt väl medvetna om förekomsten av både AI-märkningar och AI-genererat innehåll, och uppfattar dem som relativt förståeliga. Samtidigt framkom en tydlig osäkerhet kring märkningarnas tillförlitlighet, konsekvens och plattformarnas förmåga att korrekt identifiera AI-genererat innehåll.

En central slutsats är att förekomsten av AI-märkningar inte automatiskt leder till ökat förtroende för varken innehåll eller plattformarna. Detta kan förstås utifrån att majoriteten av respondenterna tidigare sett AI-genererat innehåll utan märkning, vilket påverkar hur märkningarnas trovärdighet och konsekvens uppfattades. Användares förtroende formas därför inte enbart av att märkningar finns, utan även av hur konsekvent och korrekt märkningarna implementeras i praktiken.

Studien visar även att terminologi spelar en central roll i hur AI-märkningar tolkas. Begreppet “deepfake” associerades tydligare med negativa egenskaper jämfört med “AI-generated”, som uppfattades mer neutralt. Detta visar att språkliga val påverkar hur användare bedömer innehållets trovärdighet. Även märkningarnas synlighet och placering är viktiga, eftersom märkningar som inte uppmärksammas inte heller kan bidra till ökad transparens eller stödja användares bedömning av innehållets ursprung.

Sammanfattningsvis visar studien att AI-märkningar kan fungera som ett verktyg för ökad transparens och medvetenhet, men att deras effektivitet beror på användarnas uppfattning av tydlighet, konsekvens, synlighet och tillförlitlighet. AI-märkningar bör därför inte förstås som enbart tekniska mekanismer, utan som en del av en större digital informationsmiljö där användares förtroende, tidigare erfarenheter och tolkning av plattformarnas ansvar spelar en central roll. Studien bidrar därmed till informatik genom ökad förståelse för hur transparensmekanismer i digitala miljöer påverkar användares förtroende och tolkning av innehåll på sociala medieplattformar.

7. Referenser

Bickert, M. (2024). *Our approach to labeling AI-generated content and manipulated media*. Meta Newsroom. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/

Chen, H., Wang, P. & Hao, S. (2025). AI in the spotlight: The impact of artificial intelligence disclosure on user engagement in short-form videos. *Computers in Human Behavior*, 163, Article 108448. https://doi.org/10.1016/j.chb.2024.108448

Chen, J., Wang, T.-Y., Williams, M., Jordan, N. A., Shao, M., Zhang, L. & Fussell, S. R. (2025). Examining the impact of label detail and content stakes on user perceptions of AI-generated images on social media. In: *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW Companion '25)*. Association for Computing Machinery, New York, NY, USA, pp. 270–275. https://doi.org/10.1145/3715070.3749237

Eppler, M.J. & Mengis, J. (2004). 'The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines', *The Information Society*, 20(5), pp. 325–344. https://doi.org/10.1080/01972240490507974

Epstein, Z., Fang, C.M., Arechar, A.A. & Rand, D.G. (2023). *What label should be applied to content produced by generative AI?* PsyArXiv. https://osf.io/preprints/psyarxiv/v4mfz_v1

Fendler, L. (2016). 'Ethical implications of validity-vs.-reliability trade-offs in educational research'. *Ethics and Education*. https://www-tandfonline-com.ludwig.lub.lu.se/doi/pdf/10.1080/17449642.2016.1179837

Galanis, P. & Katsiroumpa, A. (2026). *Fundamental principles on qualitative research*. *Archives of Hellenic Medicine*, 43(2), pp. 277–282. https://www.mednet.gr/archives/2026-2/pdf/277.pdf

Gamage, D., Sewwandi, D., Zhang, M. & Bandara, A.K. (2025). 'Labeling synthetic content: User perceptions of label designs for AI-generated content on social media', *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–29. https://doi.org/10.1145/3706598.3713171

Jung, Y., Hua, P., Bao, J.A. & Sundar, S.S. (2025). 'AI-generated or AI-modified? User reactions to labeling AI use in social media posts', *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, Yokohama, Japan, 26 April–1 May. New York: ACM. https://doi.org/10.1145/3706599.3720264

Li, F. & Yang, Y. (2024). Impact of Artificial Intelligence-Generated Content Labels On Perceived Accuracy, Message Credibility, and Sharing Intentions for Misinformation: Web-Based, Randomized, Controlled Experiment. JMIR Formative Research, 8, e60024. https://doi.org/10.2196/60024

Li, F., Liu, Y., Yang, Y. & Yu, G. (2026). 'How users perceive and react to labeled AI-generated content on social media: A longitudinal study of cognitive and emotional

effects', *International Journal of Human–Computer Interaction*. https://doi.org/10.1080/10447318.2026.2618553

Lim, M.W. (2025). *What is quantitative research? An overview and guidelines*. *Asia Pacific Journal of Human Resources*. https://journals-sagepub-com.ludwig.lub.lu.se/doi/10.1177/14413582241264622

Marcus, N. & Vassar, A. (2024). 'UX Design Guided by Cognitive Load Theory', in *Futureshock*. Routledge. https://doi.org/10.1201/9781032702797-6

Mayer, R.C., Davis, J.H. & Schoorman, F.D. (1995). 'An integrative model of organizational trust', *Academy of Management Review*, 20(3), pp. 709–734. https://doi.org/10.2307/258792

McKnight, D.H., Choudhury, V. & Kacmar, C. (2002). 'Developing and validating trust measures for e-commerce: An integrative typology', *Information Systems Research*, 13(3), pp. 334–359. https://doi.org/10.1287/isre.13.3.334.81

Morrow, G., Swire-Thompson, B., Polny, J.M., Kopec, M. & Wihbey, J.P. (2022). 'The emerging science of content labeling: Contextualizing social media content moderation', *Journal of the Association for Information Science and Technology*, 73(10), pp. 1365–1386. https://doi.org/10.1002/asi.24637

Müller, L.S., Nohe, C., Reiners, S., Becker, J. & Hertel, G. (2025). 'Building trust in workplace information systems: A four-company study', *Behaviour & Information Technology*. https://doi.org/10.1080/0144929X.2025.2518236

Onwuegbuzie, A.J. & Leech, N.L. (2025). 'Understanding research frameworks: A conceptual synthesis and typology across quantitative, qualitative, and mixed methods research', *International Journal of Multiple Research Approaches*, 17(2), pp. 52–92. https://doi.org/10.29034/ijmra.v17n2a1

Paas, F., Renkl, A. & Sweller, J. (2003). 'Cognitive load theory and instructional design: Recent developments', *Educational Psychologist*, 38(1), pp. 1–4. https://doi.org/10.1207/S15326985EP3801_1

Schoorman, F.D., Mayer, R.C. & Davis, J.H. (2007). 'An integrative model of organizational trust: Past, present, and future', *Academy of Management Review*, 32(2), pp. 344–354. https://doi.org/10.5465/amr.2007.24348410

Seeger, F., Wessel, M. & Lehrer, C. (2026). 'AI content labeling and user engagement on social media: The role of AI level, content type, and disclosure timing', *Electronic Markets*, 36(1), p. 31.

Söllner, M., Hoffmann, A. & Leimeister, J.M. (2016). 'Why different trust relationships matter for information systems users', *European Journal of Information Systems*, 25(3), pp. 274–287. https://doi.org/10.1057/ejis.2015.17

Wittenberg, C., Epstein, Z., Berinsky, A.J. & Rand, D.G. (2023). *Labeling AI-Generated Content: Promises, Perils, and Future Directions*. MIT Generative AI Initiative. https://mit-genai.pubpub.org/pub/hu71se89/release/1