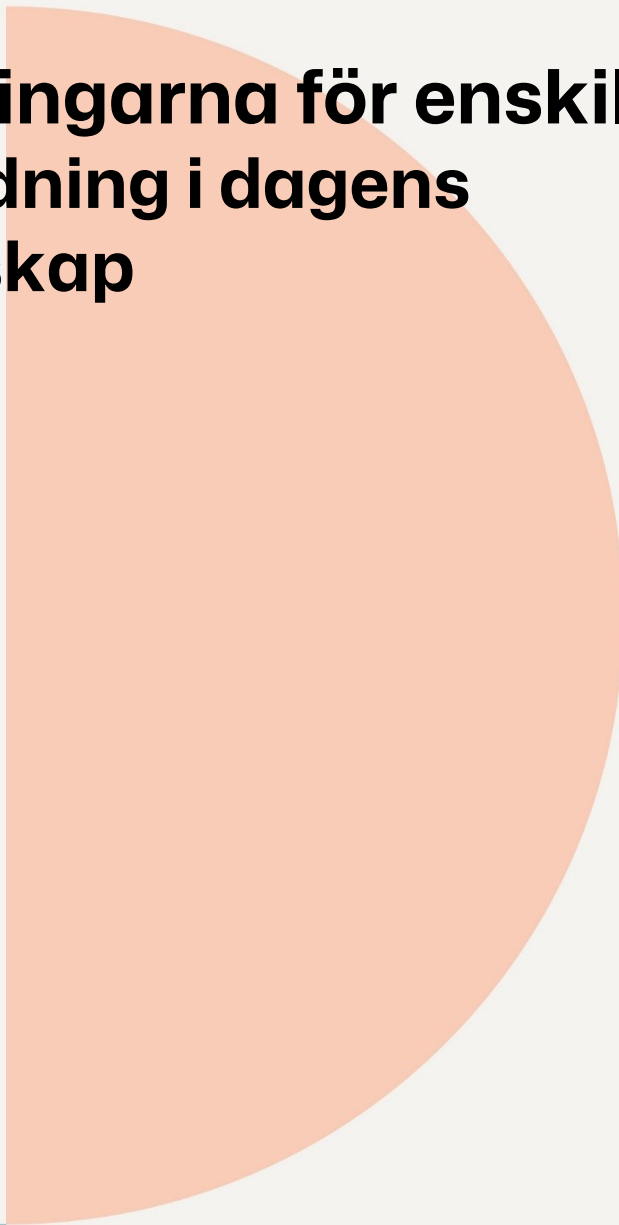
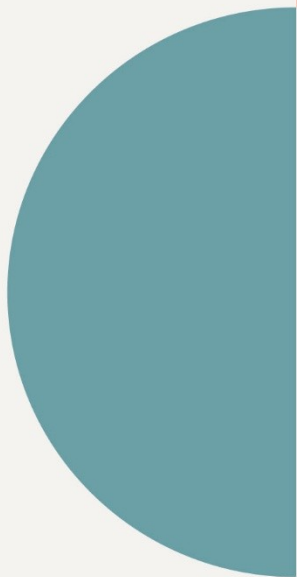


Förutsättningarna för enskildas opinionsbildning i dagens medielandskap



Mediemyndigheten

Innehåll

1	Sammanfattande slutsatser.....	3
1.1	Resultat.....	3
2	Inledning	7
2.1	Syfte och frågeställningar	7
2.2	Centrala begrepp och avgränsningar	8
2.2.1	Läsanvisningar	10
3	Metod och material.....	10
4	Användning av och förtroende för de sociala medieplattformarna i Sverige.....	12
4.1	Olika plattformar i olika befolkningsgrupper	12
4.2	Förtroendet för medier varierar	14
4.3	Digitalt utanförskap.....	14
5	Medielandskapets förändring och betydelse för individens opinionsbildning	15
5.1	Plattformisering	15
5.1.1	Från sociala till parasociala medier i uppmärksamhetsekonomin.....	16
5.2	Digitala medier och demokrati	17
5.2.1	Algorithmic audiencing och algorithmic public opinion.....	19
5.2.2	Politiska influencers och opinionsbildning	20
6	Vilka plattformsfunktioner påverkar den enskildes faktiska möjligheter att opinionsbilda?	22
6.1	Innehållsmoderering	23
6.1.1	Moderering som ett flerdimensionellt system.....	23
6.1.2	Lagligt, olagligt eller oönskat innehåll.....	25
6.1.3	Innehållsmoderering i praktiken	25
6.1.4	Innehållsmoderering i praktiken – i Sverige	26
6.1.5	Automatiserad innehållsmoderering	27
6.2	Algoritmisk kurering och rekommendationssystem.....	30
6.2.1	Dold nedprioritering (shadowbanning).....	31
6.3	Innehållsmodereringens för- och nackdelar för det demokratiska samtalet	34
7	Hur påverkar hot, hat och trakasserier viljan att delta i den digitala samhällsdebatten?.....	39
7.1	Konsekvenser och åtgärder på individ- och strukturell nivå	40
8	Medie- och informationskunnighetens betydelse för opinionsbildning.....	42
8.1	MIK:s betydelse för tolkning och värdering av information.....	42
8.2	MIK som skydd mot desinformation och otillbörlig informationspåverkan.....	43
8.3	MIK, medborgerlig kompetens och demokrati	44
8.4	MIK och politiskt deltagande	46
8.5	Algoritmisk medvetenhet	47
8.6	Begränsningar och komplexitet i sambandet mellan MIK och opinionsbildning	49
9	Utvärdering och forskning om DSA	51
9.1	Transparenskravens genomslag.....	52
9.1.1	Ökad insyn – men begränsad tillförlitlighet.....	52
9.1.2	Fortsatt stora skillnader mellan plattformar.....	52

9.1.3	Automatisering är en viktig del i modereringen, men uppgifterna är svåra att tolka	53
9.1.4	Databrister försvårar jämförelser och analys	53
9.1.5	Begränsade förändringar i moderering under valperioden	54
9.1.6	Begränsad insyn i rekommendationssystemens påverkan.....	54
9.2	DSA:s betydelse för grundläggande rättigheter och opinionsbildning	55
9.2.1	En procedurell regleringsmodell som kan stärka grundläggande rättigheter.....	55
9.2.2	Risk för övermoderering	56
9.2.3	Systemrisk, algoritmer och den praktiska yttrandefriheten	57
9.2.4	Tillsynens betydelse	59
9.2.5	Samlad analys: vad betyder DSA för enskildas opinionsbildning?	60
10	Bilaga. Transparens och reglering: DSA	62
10.1	DSA och skyldigheter för VLOP och VLOSE	63
10.2	DSA och självreglering	64
10.2.1	Frivilliga åtaganden och system för interna klagomål	64
10.2.2	Metas tillsynsråd.....	65
10.3	Transparens om villkor och begränsningar	65
10.4	Borttagning av innehåll och svensk grundlag	66
11	Referenser	68

1 Sammanfattande slutsatser

Regeringen gav Mediemyndigheten i uppdrag den 5 mars 2026 att göra en studie av förutsättningarna för enskildas opinionsbildning i dagens medielandskap (Ku2026/00259). Denna skrivelse utgör svaret på detta uppdrag.

Studien är en kunskapsöversikt baserad på befintliga undersökningar och studier, främst artiklar publicerade i vetenskapliga tidskrifter men även konferenspapers och rapporter från bland annat myndigheter och organ inom EU. Särskilt fokus läggs på enskildas ställning i förhållande till sociala medieplattformar och transparens kring innehållsmoderering, algoritmer och andra funktioner som kan påverka enskildas deltagande i samhällsdebatten.

Studiens två övergripande frågeställningar är:

- Hur påverkar plattformarnas ställning, utformning och funktioner enskildas möjligheter till opinionsbildning?
- Vilken betydelse har DSA för reglering och transparens kring plattformarnas utformning och funktioner och därmed för enskildas möjligheter till opinionsbildning?

1.1 Resultat

Förutsättningarna för enskilda att bilda opinion har nog aldrig varit så goda som i dagens medielandskap. Hög tillgång till digital teknik (smartphone) och till sociala medieplattformar med potentiellt internationell räckvidd ger historiskt unika möjligheter för den svenska allmänheten att delta i det offentliga samtalet och driva frågor som man själv finner intressanta och relevanta. Verktygen finns, men som föreliggande rapport visar finns det vissa hinder i praktiken.

- Hur påverkar plattformarnas ställning, utformning och funktioner enskildas möjligheter till opinionsbildning?

Det är några få plattformar, och ännu färre ägare, som har mycket starka positioner i dagens medielandskap och som därmed fungerar som grindvakter för den som vill publicera och sprida material online. De plattformar som används mest i Sverige (Instagram, Facebook, Snapchat, Tiktok, Youtube) är även bland de mest använda globalt och uppfyller villkoren för VLOP eller VLOSE i DSA. Förändringen av medielandskapet i riktning mot några få starka plattformar – *plattformiseringen* – har omformat den offentliga sfärens struktur och förflyttat ansvar och makt över denna, från demokratiskt styrda institutioner till privatägda, kommersiellt drivna plattformsföretag med begränsad insyn. Forskningen pekar på framför allt två problem med dessa plattformar idag:

- **Plattformarnas affärsmodeller** som kortfattat går ut på att fånga och behålla användarnas uppmärksamhet så länge som möjligt för att kunna samla in data om användarna, som används för prediktion, skapande av personliga flöden, anpassade annonser med mera. Detta är premissen för plattformarnas utformning, funktioner och moderering samt för all användning och interaktion på plattformarna.
- **Den algoritmiska styrningen** av innehåll på plattformarna som, å ena sidan, är ett effektivt och nödvändigt verktyg för plattformarna men som, å andra sidan, ger automatiserade system betydande makt samtidigt som det är svårt att beskriva, granska och förstå för användarna och andra oberoende parter.

Forskningen visar att moderering är ett flerdimensionellt system som inbegriper både borttagning av innehåll och olika former av synlighetsstyrning. Moderering handlar inte bara om vad som får sägas eller synas utan om vem som ska få höra eller se vad – inte bara om rätten att få *yttra sig* utan om att kunna *bli hörd*. Det är också ett regulatoriskt kontinuum med flera parter: staten/överstatliga organ, plattformsföretagen, tredjepartsgranskare och användarna. Det är reglering ”av” och ”genom” plattformarna. Moderering görs på grundval av vad som är lagligt/olagligt samt affärsmässigt gynnsamt. Det är särskilt regleringen av innehåll som är ”lagligt men oönskat” som blivit omdiskuterat då det styrs av plattformsföretagens användarvillkor och *guidelines*, vilka primärt baseras på företagets värderingar och kommersiella överväganden. Det är här som tolkningsutrymmet är störst.

I praktiken sker modereringen genom en hybrid av människor och automatiserade system men exakt hur kombinationen ser ut är plattformsspecifikt och föränderligt. Detta innebär dels att modereringslandskapet är svåröverskådligt och att det inte går att dra generella slutsatser om hur moderering går till, dels att rättssäkerheten varierar. Därmed varierar även den enskildes ställning i förhållande till plattformarna och dennes möjlighet att komma till uttryck där, särskilt om det rör sig om innehåll som är ”lagligt men oönskat”.

Innehållsmoderering är alltså generellt svårgenomträngligt och inte så enkelt att få grepp om och när det involverar AI uppstår det som beskrivits som *the black box problem*. Forskare pekar på flera problem med AI-baserad algoritmisk moderering vilka sammantagna försvårar insyn och begriplighet, genererar frågor om hur rättvisa systemen egentligen är samt riskerar att dölja och avpolitisera beslutsprocesser som bland annat kan ha konsekvenser för den enskildes möjligheter att opinionsbilda. Flera forskare har belyst problemen med *shadowbanning*, som ibland nästan framstår som mytisk då företagen förnekar att det existerar samtidigt som användare upplever att de har utsatts för det. Flera studier visar hur *the black box problem* skapar epistemiska konflikter mellan användare och plattformar (*black box gaslighting*, Cotter, 2023), folkteorier om hur systemen fungerar och om hur vissa sociala identiteter premieras respektive marginaliseras: vem och hur man ska vara för att bli en av de algoritmiskt

privilegierade och undvika algoritmisk marginalisering. Denna upplevt oförklarliga marginalisering och *algorithmic precarity* kan, enligt Register et al. (2023), leda till frustration, skam, självzensur och till att användare blir hypervaksamma och emotionellt bundna till en oförutsägbar plattform. Det asymmetriska förhållandet mellan plattformarna och användarna skapar således osäkerhet och anpassning till regler och system som vissa användare upplever som obegripliga och inte kommer att ges full insyn i. Det bör också noteras att marginaliserade grupper verkar uppleva sig råka ut för detta i högre grad än andra vilket reser frågor om dold osynlighetsstyrning också är dold orättvisa och diskriminering.

Samtidigt ger modereringssystemen och teorierna om dessa även upphov till motståndshandlingar bland användare, såsom *algospeak* och *algorithmic resistance*. Det är tydligt att *the black box problem* och bristen på transparens skapar tillitsproblem och en dragkamp mellan användare och plattformar om vad som ska få uttryckas och spridas i det digitala offentliga rummet.

Givet de sociala medieplattformarnas centrala position i dagens medielandskap och därmed för opinionsbildning har deras funktioner, särskilt AI-baserade algoritmer, stor betydelse för enskildas möjligheter till opinionsbildning men det är svårare att säga exakt hur de påverkar. Forskningen tyder på att olika grupper och individer kan ha olika möjligheter att få sitt innehåll upp- eller nedprioriterat, men detta styrs av ett flerdimensionellt system med hög grad av komplexitet och låg grad av insyn (vilket i sig är problematiskt för den enskilde). Både resultaten av modereringen – vad som prioriteras etcetera – och svårigheterna med att förstå och få insyn i systemen kan således vara problematiska för den enskildes möjligheter att delta i samhällsdebatten. Svårigheterna med att få insyn i och granska systemen delas med forskare och andra oberoende aktörer vilket medför att forskningen i nuläget inte heller kan ge några klara, empiriskt baserade, besked.

Samtidigt visar forskning att kunskaperna om hur algoritmer fungerar – den algoritmiska medvetenheten – ofta är ojämn i befolkningen och varierar med ålder, utbildning, internetfärdigheter och användningsvanor. Förmågan att förstå och navigera i algoritmiskt styrda system varierar således mellan grupper. Algoritmisk medvetenhet är på så sätt inte bara en fråga om individuell kunskap utan också om digital jämlikhet.

- Vilken betydelse har DSA för reglering och transparens kring plattformarnas utformning och funktioner och därmed för enskildas möjligheter till opinionsbildning?

Forskningen visar att DSA stärker användarens position i förhållande till plattformarna genom att skydda dennes grundläggande rättigheter samt ge rätt till insyn i och prövning av plattformarnas modereringsbeslut. Samtidigt kräver DSA att mycket stora plattformar analyserar och tar ansvar för systemrisker som kan påverka

samhället och offentligheten på strukturell nivå. DSA syftar således till att både skydda den enskildes rättigheter och det demokratiska samhället som helhet.

Forskningen visar dock att det finns en hel del som kan förbättras, vilket till stor del hänger samman med de AI-baserade algoritmernas centrala roll i systemen. Dels för att svårigheterna med att få insyn i och förstå systemen kvarstår, dels för att DSA inte fokuserar på tekniken och hur den implementerar reglerna eller genererar skaleffekter. När AI-baserade algoritmer har en så central roll i opinionsbildningsprocessen kan det framstå som något paradoxalt, men samtidigt svårt att göra något åt givet hur komplex och föränderlig tekniken är.

Sammantaget visar forskningen att DSA har stor potential för användarna men att dess betydelse för enskildas opinionsbildning kommer att bero på hur regleringen implementeras framöver. Några forskare argumenterar för att skärpta skrivningar i DSA skulle kunna stärka den enskildes rättigheter ytterligare, andra menar att det grundläggande problemet är plattformarnas affärsmodeller och så länge de inte ändras kommer inte den enskildes rättigheter och demokratin prioriteras av plattformarna. Det krävs såväl andra plattformar vilka bättre kan främja meningsfull debatt, kunskapsutbyte och användarnas förmåga till eftertanke och reflektion.

Såväl plattformarnas ställning, utformning och funktioner som regleringen av dem är alltså tämligen tvetydig och än så länge saknas större empiriska studier av regleringens effekter, liksom av de automatiserade systemen. Genom DSA ska dock forskare ges tillgång till data vilket förhoppningsvis kan leda till fler studier framöver, inte minst studier som synliggör den nu genomträngliga och osynliga tekniken.

2 Inledning

Regeringen gav Mediemyndigheten i uppdrag den 5 mars 2026 att göra en studie av förutsättningarna för enskildas opinionsbildning i dagens medielandskap (Ku2026/00259). Denna skrivelse utgör svaret på detta uppdrag.

Det fria ordet och en levande samhällsdebatt är grunden för ett fritt och öppet samhälle. Yttrandefriheten och informationsfriheten är därför centrala för vår demokrati. I dagens medielandskap sker en betydande del av samhällsdebatten på sociala medieplattformar med närmast global räckvidd, vilka drivs av stora internationella företag. Det ger enskilda chansen att nå ut brett, dela sina åsikter och skapa uppmärksamhet kring frågor utan att använda traditionella mediekkanaler. Det ger också stora möjligheter att interagera med andra, utbyta åsikter och delta i det demokratiska samtalet. Samtidigt innebär utvecklingen också nya risker och utmaningar, bland annat riskerar hot, trakasserier och hat på sociala medier att påverka vilka som vågar delta i samhällsdebatten. Villkoren och ramarna för den enskildes deltagande sätts av företagen, vilket påverkar användarens faktiska möjligheter att använda tjänsten. Hur företagen tillämpar dessa villkor har kritiserats för bristande transparens bland annat av EU-kommissionen och av användare av tjänsterna. Under de senaste åren har det tillkommit reglering inom EU som gäller innehåll på plattformar, bland annat EU:s förordning om digitala tjänster, vilken trädde i kraft för drygt två år sedan. Mot bakgrund av detta har regeringen gett Mediemyndigheten i uppdrag göra en samlad analys som på olika sätt belyser frågan om den enskildes förutsättningar för att delta i samhällsdebatten.

2.1 Syfte och frågeställningar

Syftet med föreliggande studie är att beskriva och analysera förutsättningarna för enskildas opinionsbildning i dagens medielandskap. Särskilt fokus läggs på enskildas ställning i förhållande till sociala medieplattformar och transparens kring innehållsmoderering, algoritmer och andra funktioner som kan påverka enskildas deltagande i samhällsdebatten.

Studiens två övergripande frågeställningar är:

1. Hur påverkar plattformarnas ställning, utformning och funktioner enskildas möjligheter till opinionsbildning?
2. Vilken betydelse har DSA för reglering och transparens kring plattformarnas utformning och funktioner och därmed för enskildas möjligheter till opinionsbildning?

Det centrala för denna rapport är således plattformarnas sätt att fungera och hur de, genom DSA, regleras. Men den enskildes möjligheter att delta i samhällsdebatten på

sociala medieplattformar påverkas av fler faktorer, däribland egna färdigheter och andra användares beteenden. Studien kommer därför även belysa hur hot, hat och trakasserier på sociala medieplattformar kan påverka den enskildes vilja och förmåga att delta i den allmänna debatten samt vilken betydelse den enskildes medie- och informationskunnighet (MIK), digitala kompetens och algoritmiska medvetenhet har i sammanhanget.

2.2 Centrala begrepp och avgränsningar

Sociala medier är ett begrepp som används flitigt i såväl den allmänna debatten som inom forskningen, men det saknas likväl en enhetlig definition av begreppet. En anledning till det är att de tekniska system och tjänster som åsyftas är komplexa, förekommer i många olika varianter med små eller stora skillnader emellan och de förändras dessutom över tid. Det är alltså svårt att fånga in vad exakt som utmärker sociala medier. De flesta är dock överens om att sociala medier har några karaktäristiska drag, såsom att social interaktion, nätverk och användargenererat innehåll är centralt. I denna rapport används begreppet på samma sätt som i den nyligen presenterade utredningen *En åldersgräns för barns tillgång till sociala medier*, det vill säga ”i vid mening, som det mångfacetterade begrepp det är” (SOU 2026:35 s. 48). Det väsentliga i föreliggande rapport är att det handlar om plattformar där vanliga människor har möjlighet att skapa konton och ladda upp eget innehåll (text, ljud, bild) för att sprida och synliggöra detta för andra användare, som i sin tur kan interagera med detta innehåll genom kommentarer, vidare spridning och rekontextualisering.

EU:s förordning om digitala tjänster, *Digital Services Act* (DSA), använder inte begreppet sociala medier som en egen rättslig kategori. I stället fångas sociala medier normalt upp av begreppet ”onlineplattform”, syftande på en värdtjänst som, på begäran av en tjänstemottagare/användare, lagrar och sprider information till allmänheten. Spridningen ska inte bara vara en mindre, rent underordnad funktion till en annan tjänst och kan ske till en potentiellt obegränsad publik.¹ När DSA beskrivs i rapporten åsyftas de onlineplattformar som omfattas av lagen, i synnerhet de som klassas som mycket stora onlineplattformar (*Very Large Online Platforms*, VLOP eller *Very Large Online Search Engines*, VLOSE). Dessa definieras av att de har minst 45 miljoner aktiva användare i EU och omfattas av striktare regler och särskilda skyldigheter jämfört med plattformar med färre användare.

¹ [Article 3, the Digital Services Act \(DSA\)](#).

De beskrivningar av modereringssystem och automatisering som ges i rapporten ligger på en generell nivå och går inte i detalj in på hur det fungerar på olika plattformar. Ibland nämns dock specifika plattformar som exempel.

I den forskning som refereras i rapporten undersöks och diskuteras ibland specifika plattformar, ibland sociala medieplattformar i allmänhet, och när studierna gäller enskilda plattformar kommer det framgå i rapporten.

Opinionsbildning definieras i Nationalencyklopedin som ”processer genom vilka opinion uppstår, vidmakthålls och förändras”. Det finns flera statliga offentliga utredningar som berör opinionsbildning men begreppet definieras sällan tydligt utan beskrivs och används funktionellt i relation till exempelvis demokrati, yttrandefrihet och det offentliga samtalet där uppfattningar, värderingar och ståndpunkter formas och omformas genom information, argumentation och debatt. I några utredningar betonas tydligare att det är verksamhet som syftar till att påverka människors åsikter, attityder eller beteenden (SOU 2007:107). Opinionsbildning är på så sätt starkt kopplat till åsikts- och yttrandefrihet och ger medborgarna möjlighet att påverka politiska beslut indirekt. Inom medie- och kommunikationsvetenskapen betonas kommunikationen, mediernas roll och dagordningsmakt i processen.

Opinionsbildning kan då förstås som en kommunikativ process där frågor, perspektiv och tolkningar görs synliga i offentligheten i syfte att påverka hur mottagare uppfattar vad som är viktigt, rimligt eller legitimt. Det som inom forskning brukar kallas *priming*, *framing* och agendasättande. Givet uppdragets syfte ligger fokus i denna studie på de opinionsbildande processer och verksamheter som sker på sociala medier och beaktar eller jämför inte med andra sätt att driva opinion, exempelvis gällande effektivitet och utfall. Alla kommunikativa processer har flera komponenter och i denna studie är det plattformarnas funktionssätt, med betoning på moderering och algoritmisk kurering, samt reglering som är i fokus.

I föreliggande studie förstås ”den enskilde” som en person som använder sociala medier för att sprida information, uppfattningar och ståndpunkter i någon fråga. Det vill säga ”den enskilde” betraktas först och främst i rollen som avsändare, inte mottagare. Studien går därför inte in på hur den enskildes egna uppfattningar kan *påverkas* på olika sätt, exempelvis av desinformation/missinformation. Rapporten går inte in på opinionsbildning gällande specifika samhällsfrågor och gör inga avgränsningar i det avseendet. Studien gör inte heller några avgränsningar gällande vem den enskilde är, men beskriver hur olika personer kan ha olika förutsättningar för sitt deltagande i den offentliga debatten, kopplat till bland annat tillgång och förmåga att använda digital teknik.

För övrigt kan noteras att problemområdet – enskildas förutsättningar för opinionsbildning i dagens medielandskap – är mycket brett och rör sig i gränslandet mellan teknik, politik, juridik och medieekonomi. Det har därför krävts hårda avgränsningar i litteraturgenomgången och denna rapport kommer inte att kunna ge full rättvisa åt all forskning som har, och hade kunnat, inkluderas.

2.2.1 Läsanvisningar

Föreliggande rapport är en genomgång av (primärt) akademisk forskning som på olika sätt berör problemområdet. Vilken litteratur som omfattas och hur den samlats in redogörs för i avsnitt 3. Forskningen är i de allra flesta fall engelskspråkig och i några avsnitt behålls engelska termer och begrepp om översättningar kan medföra betydelseförskjutningar eller att nyanser går förlorade.

Den första övergripande frågeställningen (fråga 1 ovan) besvaras i avsnitt 4–6. Avsnitt 4–5 beskriver ”plattformarnas ställning” medan avsnitt 6 fokuserar på utformning och funktioner, närmare bestämt moderering och algoritmisk kurerung av innehåll. Den andra frågeställningen (fråga 2 ovan) besvaras i avsnitt 9 och 10. Avsnitt 10 beskriver hur plattformarna regleras genom DSA, borttagning av innehåll och svensk grundlag och ligger som bilaga för den läsare som önskar en sådan beskrivning. Om läsaren redan har god kunskap om DSA kan bilagan hoppas över. I avsnitt 9 redogörs för forskning, utvärdering och diskussion av DSA i förhållande till den enskildes ställning och rättigheter samt opinionsbildning.

Avsnitt 7 och 8 tar kortfattat upp den psykosociala kontexten på plattformarna och användarens individuella färdigheter. Närmare bestämt hur hat, hot och trakasserier kan påverka viljan att delta i den digitala debatten samt betydelsen av medie- och informationskunnighet och algoritmisk medvetenhet för den enskilde.

3 Metod och material

Enligt regeringsuppdraget är Mediemyndighetens uppgift att, utifrån befintliga undersökningar och studier, beskriva och analysera hur medielandskapet påverkar enskildas möjligheter och förutsättningar att opinionsbilda. Det vill säga att göra en kunskapsöversikt, inte genomföra egna empiriska studier. Materialet som denna rapport bygger på är därför främst internationell forskning, publicerad i vetenskapliga tidskrifter med kollegial granskning. En aktuell avhandling från Uppsala universitet (Carretta, 2025) har också varit viktig för uppdraget. Utöver detta har några rapporter som publicerats av svenska myndigheter och organisationer samt EU-kommissionen inkluderats. Mediemyndigheten har även ställt frågor till ett urval plattformsföretag (Meta, Google, Tiktok, X, Reddit och Flashback) angående modereringsåtgärder som berör användare i Sverige, vilket redovisas i avsnitt 6.1.4.

Föreliggande rapport är således först och främst en kunskapsöversikt. Sökstrategin har varit triangulering med Google Scholar, Google Scholar Lab och Copilot. Skillnaden mellan dem är att Google Scholar är en ”klassisk” sökmotor för forskning där man med hjälp av sökord och söksträngar får fram listor på publikationer som kan vara relevanta för frågeställningen. I både Google Scholar Lab och Copilot

promptas en AI vilket innebär en mer utvecklad och flexibel sökning jämfört med en traditionell sökmotor. Vi har primärt använt Copilot, med övriga som komplement och kontroll.

Oftast görs materialinsamling till forskningsöversikter genom sökning i forskningsdatabaser, men när man använder Copilot görs sökningen på webben. Det finns både för- och nackdelar med detta. Fördelen med databaser är att de är avgränsade och att deras innehåll är definierat, vilket innebär att man har kontroll över var man söker och kan redogöra för hur många artiklar man funnit i en viss databas med den aktuella söksträngen. Nackdelen är man ofta får mycket material som måste gås igenom och relevansgranskas och där, i princip alltid, merparten av det insamlade materialet faller bort. Det är en tidskrävande process för att sälla fram relevanta artiklar. Med promptar kan man avgränsa och precisera sina sökningar så väl att det material man får in redan är ”relevansgranskat”. Nackdelen med att en AI gör relevansgranskningen är att vi inte vet vad som har sällats bort och därmed inte kan garantera att alla relevanta och intressanta artiklar har inkluderats. I just detta uppdrag har vi, som tidigare nämnts, behövt göra hårda avgränsningar så det hade ändå inte varit möjligt att inkludera mer material.

Användningen av Copilot i detta uppdrag har samtidigt varit ett utforskande av hur AI kan användas mer i myndighetens arbete. Mediemyndigheten är en av de statliga myndigheter som har i uppgift att utveckla och implementera AI-verktyg i sin verksamhet och försöket har visat att Copilot kan effektivisera materialinsamlingen till kunskapsöversikter. Här kan också noteras att en bidragande faktor till att det fungerat så väl är att alltmer forskning publiceras öppet (*open access*).

Den typ av kunskapsöversikt som presenteras i det följande är tematisk, narrativ. Artiklarna som inkluderas är publicerade inom flera olika discipliner såsom statsvetenskap, rättsvetenskap, politisk kommunikation, människa–datorinteraktion, medie- och kommunikationsvetenskap med flera. I stora delar är det forskning som har normativa utgångspunkter på så sätt att den är pro-demokratisk, pro mänskliga rättigheter och lägger an ett kritiskt perspektiv på de sociala medieplattformarna. Att forskningen har värderande inslag innebär dock inte att dess metoder och tillvägagångssätt brister i objektivitet eller systematik.

4 Användning av och förtroende för de sociala medieplattformarna i Sverige

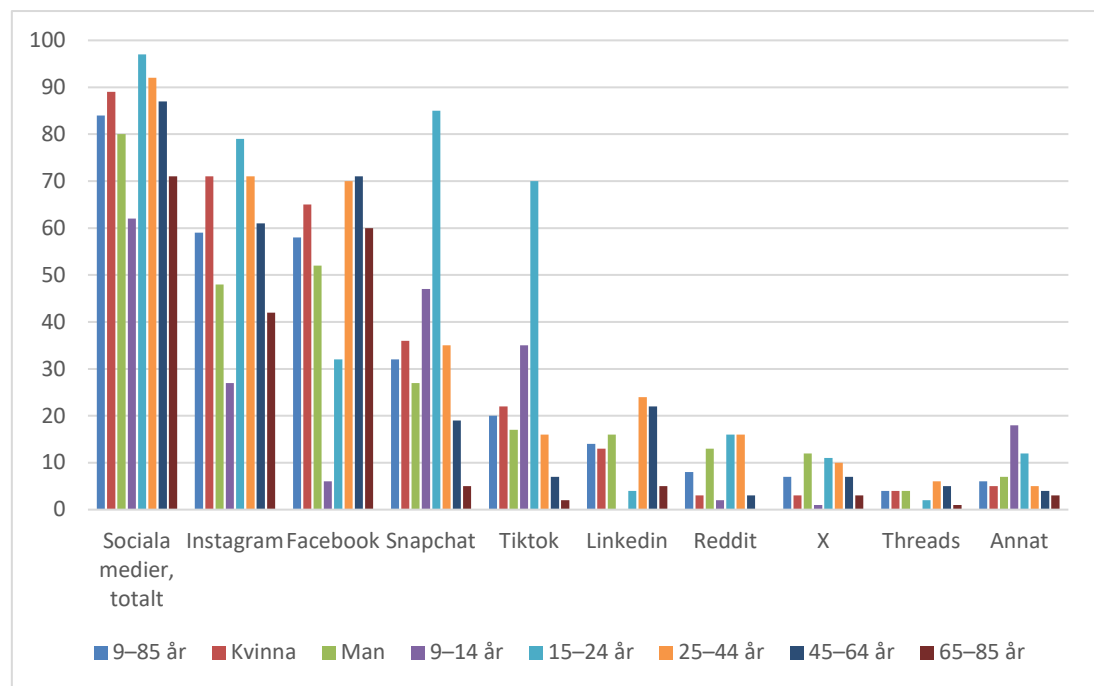
Mediekonsumtionen i Sverige har förflyttats successivt från de traditionella medieplattformarna som tv, tidningar och radio till de digitala plattformarna för streaming, sociala medier, poddar och andra helt digitala medietjänster.

Användningen av sociala medier har ökat över tid och idag använder större delen av befolkningen, 84 procent, sociala medier varje dag. Det är dock stora skillnader mellan olika åldersgrupper. Högst andel dagliga användare finns i åldersgruppen 15 till 24 år, 97 procent, och lägst andel i åldersgruppen 65 till 85 år. Men även i den gruppen använder så många som 7 av 10 sociala medier varje dag (Nordicom, 2026).

4.1 Olika plattformar i olika befolkningsgrupper

Instagram har flest användare en genomsnittlig dag i alla åldersgrupper, 59 procent, strax därefter kommer Facebook med 58 procent. Både Snapchat och Tiktok har färre användare, 32 procent respektive 20 procent en genomsnittlig dag. Hos de yngre är användningen mycket högre på Snapchat, Instagram och Tiktok. Mellan 70 och 85 procent i åldern 15 till 24 år använder Snapchat, Instagram och Tiktok en genomsnittlig dag, men endast 32 procent använder Facebook. Hos den äldsta åldersgruppen i *Mediebarometern 2025*, 65 till 85 år, är det tvärtom Facebook som är störst, 60 procent, och endast 5 procent använder Snapchat en genomsnittlig dag. Lägst daglig användning har Reddit, X och Threads (Nordicom, 2026).

Diagram. Sociala medier, daglig räckvidd, 9–85 år, 2025 (procent). Källa: *Mediebarometern 2025*, Nordicom, Göteborgs universitet.



Att användningen av sociala medier skiljer sig åt mellan de olika åldersgrupperna har till viss del att göra med när plattformen lanserades. Den plattform som lanserades i Sverige redan 2006, Facebook, har numera den äldsta målgruppen. De äldre stannade kvar på Facebook när de yngre generationerna lämnade för nya plattformar med mer visuellt uttryck och korta videos som Tiktok. De yngre rör sig dessutom mellan flera snabbväxande plattformar, medan äldre koncentrerar sin närvaro till ett fåtal. Samtidigt använder många flera plattformar parallellt (Nordicom, 2025).

Det finns också tydliga könsskillnader i den dagliga användningen av de största sociala medierna i Sverige. Sammantaget använder fler kvinnor än män sociala medier en genomsnittlig dag 2025 (89 procent jämfört med 80 procent). Men skillnaderna handlar inte bara om hur många som använder sociala medier, utan också om vilka plattformar som används. Kvinnor använder i högre grad visuella och relationsinriktade plattformar som Instagram, medan män oftare använder text- och diskussionsbaserade plattformar. Detta blir särskilt tydligt för X och Reddit, där användningen är betydligt mer mansdominerad: män använder dessa plattformar ungefär fyra gånger så ofta som kvinnor, även om den totala andelen användare fortfarande är relativt låg (Nordicom, 2026).

En stor del av den svenska befolkningen använder Youtube dagligen (52 procent), men de främsta användarna är unga män. I hela populationen använder 64 procent av männen Youtube dagligen, jämfört med 40 procent av kvinnorna (Nordicom, 2026).

4.2 Förtroendet för medier varierar

I en internationell jämförelse är förtroendet för traditionella medier generellt högt i Sverige, vilket framgår av Reuters Institutes *Digital News Report* (Newman et al., 2025). Förtroendet för traditionella nyhetsmedier, särskilt för public service, är genomgående högre än förtroendet för sociala medier. Medieakademins *Förtroendebarmeter* från 2026 visar att svenskarnas förtroende för hur sociala medieföretag sköter sitt arbete är mycket lågt. Mediemyndighetens analyser bekräftar samma mönster: även om användningen av sociala medier som nyhetskälla är utbredd, särskilt bland yngre, är tilliten låg. Bland unga uppger endast cirka 11–15 procent att de har förtroende för nyheter från sociala medier eller influencers. Samtidigt har samma grupper högt förtroende för traditionella nyhetskällor såsom SVT, SR och etablerade nyhetswebbplatser (Mediemyndigheten, 2024a). Detta innebär att användning och förtroende inte sammanfaller. Sociala medier fungerar i stor utsträckning som distributionskanaler för nyheter, men inte som tillförlitliga källor.

4.3 Digitalt utanförskap

Digitaliseringen har medfört stora fördelar för samhället och befolkningen men det finns också en risk för att personer utan digital kompetens eller tillgång till digital utrustning hamnar utanför. Det gäller framför allt de allra äldsta, personer som lever i hushåll med låga inkomster och personer med vissa funktionsnedsättningar. Även inom gruppen nyanlända finns motsvarande utmaningar, inte minst när det gäller användningen av digitala offentliga tjänster.

År 2025 var det 3 procent av befolkningen över 15 år som inte använde internet. Ålder är den faktor som har störst påverkan på internetanvändandet eftersom det i princip bara är personer i pensionsåldern som inte använder internet, i synnerhet de äldsta pensionärerna som är 76 år eller äldre (19 procent icke-användare). De som är icke-användare har ofta inte använt internet i sitt tidigare yrkesliv, eller har på grund av fysiska hinder slutat att använda internet, enligt *Svenskarna och internet 2025*.

För att människor ska vara digitalt inkluderade är tillgång till digitala tjänster, tjänsternas användbarhet och individuella förutsättningar avgörande. Kompetens samt vilja och motivation att använda digitala tjänster är en viktig del av de individuella förutsättningarna. Sammantaget finns det olika anledningar till digitalt utanförskap i Sverige idag då det kan vara självvalt och på grund av bristande intresse, men också på grund av bristande kunskap, fysiska skäl eller ekonomiska skäl. Även om den absoluta merparten av den svenska befolkningen har tillgång till digital teknik och sociala medier för att ta del av, och vara en del av, det digitala samtalet är det viktigt att påminna om denna grupp.

5 Medielandskapets förändring och betydelse för individens opinionsbildning

De sociala medieplattformarna har alltså sammantaget stor räckvidd i den svenska befolkningen och digitaliseringen har förändrat – och förändrar – medielandskapet i snabb takt. Från ett landskap dominerat av massmedier till plattformsstyrda informationsflöden, från redaktionella medier till sociala medier i rollen som grindvakter, vilket har medfört både möjligheter och risker. I det följande avsnittet beskrivs denna process kortfattat, med särskilt fokus på forskning om sociala medier och demokrati i avsikt att ringa in plattformarnas ställning.

5.1 Plattformisering

Perioden mellan 1990 och cirka 2004 brukar beskrivas som den statiska webbens tid, eller web 1.0. Detta innebar att publiken bara kunde läsa och passivt konsumera innehållet, till skillnad från perioden efter, web 2.0, där internet blivit mer dynamiskt och socialt i och med att användarna fått möjlighet att aktivt skapa innehåll, agera och kommunicera med andra användare. Distribution av information har historiskt sett varit dyrt men i och med webben sjönk kostnaderna markant varför betydligt fler kunde publicera sig. Detta skifte var avgörande för relationen mellan publiken och medierna, och för den enskildes möjlighet att bli aktör och uttrycka sig själv i offentligheten. Nya begrepp som ”användargenererat innehåll” och ”prosumers” (Bruns, 2008) etablerades för att beskriva den nya aktiva, deltagande användaren i ”konvergenskulturen” (Jenkins, 2008). Att ”dela” var ett nyckelord (John, 2024). Till en början skedde detta genom digitala nätverk, såsom olika forum och bloggar, men den decentraliserade bloggsfären blev efterhand utkonkurrerad av de framväxande sociala medieplattformarna (Gerrard, 2022; boyd², 2026). En plattform utgör en teknisk infrastruktur som ägs och kontrolleras av en aktör, ofta ett företag (SOU 2020:56, s 63). Dessa företag hade tidigt som affärsidé att knyta samman människor, det vill säga optimera *connectivity* och användarengagemang, globalt. Genom bland annat uppköp av konkurrenter har marknaden gått mot oligopol med några få ”digitala giganter” (van Dijck, 2013; Myers West, 2022). Det görs försök att starta och driva utmanare men i dagsläget är det likväl några få plattformar som har mycket starka positioner. Globalt är de största sociala medieplattformarna Facebook, Instagram, Whatsapp, Youtube och Tiktok, följt av stora regionala eller funktionsspecifika plattformar som Wechat/Weixin, Telegram, Messenger, Snapchat,

² Professor danah boyd stavar sitt namn utan versaler.

Reddit, Douyin, X, Pinterest, Threads och LinkedIn (Meltwater & We Are Social, 2025). De fem största plattformarna, vilka används i hög grad i Norden, ägs av tre företag: Facebook, Instagram och Whatsapp ägs av Meta, Youtube ägs av Google (Alphabet Inc.) och Tiktok ägs av Bytedance. Samtidigt som den enskilde har fått ökade möjligheter att använda internet för att göra sin röst hörd har antalet plattformar där man kan nå en större publik således begränsats till några få, vilka genom sina starka positioner penetrerar de flesta samhällsliga sektorer och delar av livet. Några få stora företag reorganiserar kulturella praktiker och har blivit ”power brokers” (Myers West, 2020 s. 7ff). Om man tidigare behövde passera massmediernas redaktionella grindvakter för att bli publicerad är det nu plattformarna som sätter villkoren – en process som inom forskningen brukar kallas *platformisation* (Poell, Nieborg & van Dijck, 2019).

5.1.1 Från sociala till parasociala medier i uppmärksamhetsekonomin

I takt med att de sociala medieplattformarna vuxit har de dock blivit allt mindre sociala: generellt verkar användare ägna sig mindre åt att interagera och posta eget innehåll och mer åt passiv konsumtion av innehåll (John, 2024; Jungselius & Weilenmann, 2025; boyd, 2026). Innehållet har inte längre samma personliga prägel som det hade för omkring 20 år sedan på forum som Myspace utan har successivt blivit mer genomtänkt, strategiskt konstruerat och professionellt. Det som blir viralt är inte sällan skapat strategiskt i avsikt att bli just viralt, bland annat av den tämligen nya kategorin av professionella kreatörer och influencers vars existens möjliggjorts av de stora plattformarna (Riedl, Lukito & Woolley, 2023). De sociala medieplattformarna har, med danah boyds ord, blivit parasociala, syftande på de ensidiga relationer som karakteriserat förhållandet mellan publik och mediepersonligheter sedan massmediernas begynnelse (boyd, 2026).

Förklaringen till denna förändring av plattformarnas karaktär är deras affärsmodeller. Innehåll som drar in och håller kvar många användare under lång tid är mer lönsamt då publikens uppmärksamhet kan säljas till annonsörer. Sociala medieföretag är delar av en ny digital industri som med hjälp av avancerad teknik extraherar data ur all interaktion som sker online. De digitala spår som en person lämnar kan sammanlagt användas för att skapa digitala profiler som används för att förutsäga och dra slutsatser om personers framtida beteende och preferenser. Användardata är alltså en resurs och för att få så mycket data som möjligt vill plattformarna dra in och behålla sin publik, kort sagt få så mycket uppmärksamhet som möjligt. En affärsmodell som Shosana Zuboff (2021) benämnt ”övervakningskapitalismen”, också kallad uppmärksamhetsekonomin. Detta är premisen för plattformarnas utformning, funktioner och moderering. Det är också premisen för användarna och all interaktion på plattformarna. Man betalar inte för sin användning med pengar utan

med data. Plattformarna har alltså möjliggjort för fler att publicera sig och komma till tals offentligt, med det är inte gratis (se t.ex. Lanier, 2011; Zuboff, 2021).

I början av internets framväxt fanns en något utopisk och idealiserad bild av internet, inte minst idéer om hur tekniken skulle kunna bidra till en demokratisering av den offentliga sfären. I de mest utopiska visionerna innebar den digitala tekniken ”the end of history”, ett kvalitativt nytt skede i mänsklighetens historia, ”the Information Age”, där gamla faktorer som tidigare begränsat oss, den ekonomiska tillväxten, demokratiseringen osv. inte längre gällde (Mosco, 2004). Under 2010-talet användes den digitala tekniken i folkliga uppror och demonstrationer mot auktoritära, korrupta ledare – inte minst under den så kallade arabiska våren. Protesterna har också kallats Twitter- och Facebook-revolutionen eftersom de digitala tjänsterna användes för både organisering och kontakt mellan oppositionella grupper och för spridning av information till omvärlden. Det blev på så sätt tydligt vilken enorm politisk och demokratisk potential som fanns i den digitala tekniken.

Denna optimism har sedan gradvis minskat. För det första, för att sociala medier lika gärna kan användas för statlig övervakning och kontroll, inte minst av auktoritära stater som anammat den digitala tekniken för att trycka tillbaka oppositionella rörelser och tysta avvikande röster (Morozov, 2011; Dumbrava, 2021). I skrivande stund stänger den ryska staten tidvis ner internet, försvårar användning av VPN och tjänster som Whatsapp och Telegram för att styra över den digitala kommunikationen till den statligt kontrollerade appen Max (se även Fornusek, 2026).

För det andra för att ”plattformiseringen” har omformat den offentliga sfärens struktur och förflyttat ansvar och makt över denna, från demokratiskt styrda institutioner till privatägda, kommersiellt drivna plattformsföretag utan insyn. Demokratin utspelas nu, som Balčytienė et al. (2026) uttrycker det, i skuggan av dessa.

5.2 Digitala medier och demokrati

Sociala medieplattformar fungerar idag som villkorande infrastrukturer för opinionsbildning: de erbjuder nya möjligheter till uttryck, men de kontrollerar också synlighet, räckvidd och villkor för deltagande. Forskningen visar att digitala medier ofta sänker deltagandetrösklar och kan stärka politisk delaktighet, men också att effekterna varierar beroende på typ av plattform, politisk kontext och vilka demokrativariabler man studerar (Lorenz-Spreen et al., 2023). Digitaliseringen har i vissa avseenden stärkt sociala rörelser, t.ex. #metoo-upproppen i flera länder, och har i Sverige möjliggjort upprop av enfrågerörelser såsom Arkitektupproret, Bensinupproret, Vårdupproret och Undersköterskeupproret (SOU 2020:56 s. 84). För civilsamhället är sociala medier viktiga medel för kommunikation, opinionsbildning och mobilisering (Chong, 2024).

Flera svenska utredningar har undersökt digitaliseringens påverkan på demokratin och de visar att opinionsbildningen i viss mån har demokratiserats, av ovan nämnda skäl. Enligt utredningen *Det demokratiska samtalet i en digital tid* finns det dock ingenting som tyder på att internet har lett till ökad politisk jämlikhet i Sverige. Det tenderar att vara samma personer som intresserar sig för och engagerar sig politiskt i exempelvis föreningar som också använder sociala medier i sitt politiska engagemang (SOU 2020:56 s. 83). En internationell meta-analys av 38 longitudinella studier visade på liknande sätt att det framför allt är de redan politiskt motiverade och engagerade personerna som använder sociala medier, vilket tyder på att sociala medier snarare förstärker existerande skillnader mellan grupper än mobiliserar nya personer att engagera sig (Oser & Boulianne, 2020). En metaanalys av forskning som undersökt sociala mediers betydelse för politiskt engagemang utanför plattformarna fann svaga till måttliga positiva samband, vilket antyder att den positiva kopplingen mellan användning av sociala medier och reellt politiskt deltagande endast uppstår under vissa omständigheter (Jung, Dai & Alberracin, 2023).

Lorenz-Spreen med kollegor (2023) har gjort en omfattande metaanalys av forskning om sambanden mellan digitala medier och olika aspekter av demokratin. De försöker alltså ta ett brett grepp om digitaliseringens betydelse för demokratin och inkluderar olika slags undersökningar. Metaanalysen bekräftar att utfallen varierar mellan olika sammanhang men visar dessutom att utfallen kan vara både positiva och negativa – ökad delaktighet på vissa områden, men också risker för polarisering, misstro och manipulativ påverkan.

Sambanden i studierna som ingår i metaanalysen är ofta korrelationer men här finns också studier som kan peka på kausalitet. Metaanalysen omfattar totalt 496 artiklar publicerade inom flera olika discipliner och med bred geografisk spridning av det empiriska materialet. Resultaten visar, för det första, att de flesta artiklar finner samband mellan digitala medier och olika demokratiska utfallsmått men att effekten varierar mellan de olika utfallen och den samhälleliga kontexten. Det är få studier som rapporterar noll-effekt. Det tydligaste för demokratin positiva utfallet är ökat politiskt deltagande och mobilisering, från enklare insatser som att gilla och dela politiskt innehåll på nätet till att delta i fysiska demonstrationer. Flera artiklar har också funnit positiva samband mellan användning av digitala medier och ökad politisk kunskap samt ökat intresse för politik bland ungdomar. Resultaten var dock ganska blandade på detta område, vilket även gäller studier av enskildas användning av digitala medier för att uttrycka sina politiska åsikter. Det finns både studier som visar på positiva utfall och på negativa. De negativa utfallen förklaras ofta med teorin om tystnadsspiralen: att individens vilja att uttrycka sina politiska åsikter beror på om individen uppfattar att det finns stöd för åsikten.³

³ [NE - tystnadsspiralen](#).

För flera andra demokratiska utfallsmått var resultaten huvudsakligen negativa. Det finns samband mellan digital medieanvändning och minskat förtroende för institutioner, ökad exponering för desinformation och ökad politisk polarisering. Olika variabler har alltså olika utfall för det demokratiska systemet, men utfallen varierar också med kontexten.

De mest positiva utfallen fanns i länder som genomgår en demokratiseringsprocess i Sydamerika, Afrika och södra Asien. Resultaten är mer blandade i utvecklade demokratier och flest skadliga utfall för demokratin fanns i Europa, USA och Ryssland. I demokratiska länder var både användning av sociala medier och annan internetanvändning förknippad med ökad polarisering. Positiva utfall modererades ofta av politiskt intresse och nyhetskonsumtion medan skadliga utfall modererades av politisk orientering. Utfallen varierar alltså även på individnivå och mellan olika subgrupper. Det är därmed svårt att dra generella slutsatser om *vilka* effekter digitala medier har på demokratin, men det är otvetydigt *att* de spelar roll.

5.2.1 Algorithmic audiencing och algorithmic public opinion

Tittar man mer specifikt på sociala medier i relation till demokrati kan konstateras att det är ett komplext forskningsområde som blivit än mer komplicerat och svårgenomträngligt i och med implementeringen av alltmer avancerad AI och algoritmer på plattformarna. Under det senaste dryga decenniet har forskningsfält som *critical AI studies* (t.ex. Lindgren, 2024) och *critical algorithm studies* (t.ex. Gandini et al., 2022; 2025) vuxit fram, vilka fokuserar på de samhälleliga effekterna av AI och algoritmer, inte minst för opinionsbildning och demokrati. De gemensamma utgångspunkterna är ett kritiskt perspektiv på dels plattformarnas affärsmodeller där användarnas uppmärksamhet och data är den främsta valutan, dels den algoritmiska styrningen av plattformarnas innehåll. Givet plattformarnas centrala ställning i samhället styrs i princip den offentliga debatten av ogenomsådliga, insynsskyddade algoritmiska system. Man kan studera vad som, så att säga, matas in i algoritmen (input) och vad som kommer ut (output) men inte vad som sker däremellan. Algoritmerna är utformade i enlighet med affärsmodellens mål men formas också genom användarnas interaktioner i en feedbackloop mellan systemet och användarna. Både opinionsbildningsprocessen och dess produkt – ”opinionen” – skapas algoritmiskt, men vi vet inte närmare hur det går till (Gandini et al., 2022; 2025).

Algoritmer styr vilket innehåll som ska lyftas fram respektive nedprioriteras för vem, vilket ger personaliserade plattformsflöden. Ur ett yttrandefrihetsperspektiv reser detta flera frågor, både för den som yttrar sig digitalt och för publiken. Dels frågor om vad som kan yttras, dels frågor om vad som kommer att höras/synas och av vem eller vilka. I tidigare digitala nätverk på internet kunde användare kommunicera med och vända sig till en specifik grupp, gemenskap eller vänner, men idag kan inte den

som yttrar sig på öppna konton själv avgöra vem som kommer att nå eftersom det är algoritmer som avgör detta. Algoritmer ”kurerar” innehåll, men även publik (Riemer & Peter, 2021).

Den enskilde användaren blir således algoritmiskt positionerad *som* publik och *för* publik samt exponerad för innehåll som denne i viss mån kan påverka genom sin egen interaktion på plattformarna men som till syvende och sist styrs av plattformarnas ekonomiska logik. I avsnitt 6 beskrivs några centrala system och praktiker som plattformarna använder sig av i dessa processer, men först några ord om betydelsen av influencers i dagens medielandskap.

5.2.2 Politiska influencers och opinionsbildning

Sociala medier har, som tidigare nämnts, möjliggjort framväxten av en ny kategori mediepersonligheter med potentiellt stort inflytande över det offentliga samtalet och opinionsbildningen. Influencers (eller kreatörer som de också benämns idag) har tidigare främst förknippats med livsstilsrelaterat, personligt innehåll, underhållning och marknadsföring, men under senare år har det skett en förskjutning där gränserna mellan det kommersiella, det personliga och det politiska blivit alltmer flytande. Vissa samtida influencers blandar kommersiella budskap och sponsrade samarbeten med uttryck för samhällsengagemang varmed politiska budskap och ideologiska ställningstaganden integreras i det vardagliga innehållet (Riedl, Lukito & Woolley, 2023; von Sikorski et al., 2025). Det innebär att opinionsbildning sker i nya former och sammanhang som skiljer sig från den traditionella politiska debatten.

Influencers genomslag bygger framför allt på upplevd autenticitet, närhet och förtroende (Goodwin et al., 2023; Riedl, Lukito & Woolley, 2023). Den upplevda närheten mellan influencer och följare – alltså den parasociala relationen – kan göra politiska budskap mer relevanta och personliga. Detta innebär att influencers kan göra politiska frågor mer tillgängliga och fungera som pedagogiska förmedlare av grundläggande politisk kunskap, särskilt i format som är anpassade till unga målgrupper. I sådana fall kan influencers komplettera traditionella informationskällor. De kan till exempel erbjuda lättillgängliga förklaringar och perspektiv, utan att ersätta journalistik eller myndighetsinformation. Detta är särskilt relevant för grupper som inte aktivt söker politisk information i traditionella nyhetsmedier (Sehl & Schützeneder, 2023; Harff & Schmuck, 2024; Harff & Schmuck, 2025; Wasike, 2023).

Genom att föra in politiska frågor i nya sammanhang kan influencers också fungera som mellanhänder mellan politiska aktörer, samhällsfrågor och användares vardagliga medieflöden. På så sätt kan de påverka vilka former av deltagande som upplevs som möjliga, relevanta och meningsfulla på sociala medier (Riedl, Lukito & Woolley, 2023; von Sikorski et al., 2025; Goodwin et al., 2023).

Politisk opinionsbildning på sociala medier sker ofta genom handlingar som är knutna till plattformarnas funktioner för interaktion. Det kan till exempel handla om att gilla, kommentera, dela, svara på inlägg, delta i en livechatt eller bidra till att ett visst budskap får större synlighet – det vill säga spela på det man tror trigger algoritmerna i rätt riktning. För enskilda kan detta skapa nya möjligheter att delta i kollektiv opinionsbildning. Politiska influencers kan mobilisera sina följare att förstärka ett svar eller en kritik genom sådana handlingar och därigenom påverka vad som blir synligt i det digitala samtalet. Men konkurrensen om uppmärksamheten är hård och andra aktörer och publikgrupper kan mobilisera motreaktioner och försöka få genomslag för alternativa budskap (Merz & von Sikorski, 2025; Harris, Foxman & Partin, 2023; Riedl, Lukito & Woolley, 2023). En enskild användare som vill bilda opinion kan således välja att ansluta sig till redan existerande kollektiv och kreatörer som driver sakfrågor, uppfattningar och värderingar som överensstämmer med de egna. Eller försöka att själv skapa synlighet för sina frågor och perspektiv, vilket kan vara en utmaning givet konkurrensen från redan etablerade, mer eller mindre professionella influencers. Influencers kan på så sätt vara både en tillgång och ett hinder för den enskilde användaren.

Sammanfattningsvis är plattformarnas ställning central för nutida opinionsbildning, vilket innebär att de AI-baserade algoritmer som styr innehållet i enlighet med företagets affärsmässiga mål kan ha avgörande betydelse för såväl processen som resultatet – opinionsbildning har i mångt och mycket blivit *algoritmiska* ”processer genom vilka opinion uppstår, vidmakthålls och förändras”⁴ och det är detta som den enskilde har att förhålla sig till.

⁴ [NE - opinionsbildning](#).

6 Vilka plattformsfunktioner påverkar den enskildes faktiska möjligheter att opinionsbilda?

Tekniken är, som så många gånger påpekats, varken ond eller god i sig själv men den är heller inte neutral eftersom den sätter gränser för vad som kan göras, underlättar vissa handlingar och försvårar eller omöjliggör andra. Teknikens funktionssätt sätter ramarna för det praktiskt möjliga. Gränssnitt och interaktionsdesign är centralt för relationen mellan användare och plattform och därmed för den enskildes opinionsbildning (Lewis, 2014), men i denna studie ligger inte fokus där utan på de funktioner som avgör vilket innehåll som blir synligt på plattformarna:

- innehållsmoderering (borttagning, kontosanktioner)
- algoritmisk kurering/rekommendationssystem (vad som synliggörs, döljs eller förstärks).

Varje dag laddas enorma mängder material upp på plattformarna och detta måste hanteras och organiseras på något sätt. Delar av det uppladdade materialet är dessutom av en karaktär som inte är önskvärd i det offentliga rummet: filmade övergrepp, misshandel, mord, terrorbrott, desinformation och så vidare. Detta innehåll kan leda till skada i flera bemärkelser: det kan skada samhället och den offentliga debatten om exempelvis propaganda får florera fritt, innehåll kan skada olika grupper, exempelvis barn, och det kan skada plattformsföretagens relation till både annonsörer och publiken (Gerrard, 2022). Att inte moderera innehållet är därför inte ett alternativ – frågan är hur modereringen bör göras. Forskningen visar att moderering av olika slag, inklusive rekommendationssystem, kan påverka enskildas tillit, upplevelse av rättvisa och viljan att uttrycka sig politiskt så hur moderering och kurering sker har stor betydelse. Moderering innebär både borttagning av innehåll och synlighetsstyrning genom bland annat rekommendationssystem. I detta avsnitt ges först en översiktlig beskrivning av moderering, i bemärkelsen av att ta bort oönskat eller skadligt innehåll ("pull"), och därefter den algoritmiska kureringen av innehåll genom rekommendationssystem ("push", Carretta, 2025). Båda systemen är alltså olika former av moderering där den förstnämnda är "hård" medan den andra brukar beskrivas som en mjukare form av moderering (Gorwa et al., 2020).

6.1 Innehållsmoderering

Innehållsmoderering definieras i DSA (artikel 3 t) som de aktiviteter, automatiserade eller inte, som leverantörer av förmedlingstjänster utför för att särskilt upptäcka, identifiera och hantera olagligt innehåll eller information som är oförenlig med deras villkor, och som tillhandahålls av tjänstemottagare. Det omfattar åtgärder som påverkar innehållets tillgänglighet, synlighet eller åtkomlighet, exempelvis nedprioritering, demonetisering, avaktivering av åtkomst, borttagning samt åtgärder som påverkar användarens möjlighet att tillhandahålla information, exempelvis uppsägning eller avstängning av ett konto.⁵

Mer kortfattat kan det beskrivas som plattformarnas åtgärder för att upptäcka, identifiera och hantera olagligt innehåll eller innehåll som bryter mot tjänstens villkor, inklusive borttagning, begränsning, nedprioritering eller kontosanktioner.

Inom forskningen finns flera olika definitioner av moderering men en som ofta citeras är Grimmelmans definition från 2015. Den är kortare och betonar snarare syftet än praktiken: “the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse” (Grimmelmann, 2015).

En annan ofta citerad definition betonar i stället modereringens komplexitet och systemkaraktär:

the detection of, assessment of, and interventions taken on content or behaviour deemed unacceptable by platforms or other information intermediaries, including the rules they impose, the human labour and technologies required, and the institutional mechanisms of adjudication, enforcement, and appeal that support it (Gillespie et al., 2020 s. 2).

Moderering är alltså inte en enkelt avgränsad och definierad företeelse, inte *en* enda praktik, utan en flerdimensionell process som sker på olika nivåer och utförs av olika aktörer, involverande människor, institutioner, organisationer och tekniska system. Det omfattar inte bara att avgöra vad som får finnas kvar, utan också hur innehåll upptäcks, klassificeras, rangordnas, förstärks eller begränsas, vilka användare som får synlighet, och hur tvister om beslut hanteras (Badouard & Bellon, 2025; Caplan, 2023).

6.1.1 Moderering som ett flerdimensionellt system

Carretta beskriver i sin avhandling hur innehållsmoderering idag kan ses som ett treparts-system. Vad som får yttras i offentligheten har reglerats sedan länge, men på olika sätt. I den pre-digitala eran, det vill säga innan internet, kan regleringen beskrivas som ett dualistiskt system där, kortfattat, staten var regulator och den som yttrade sig blev reglerad. Den passiva publiken, de som lyssnade eller läste, ingick inte i modellen. I och med framväxten av internet, de sociala medierna och publikens nya roll som

⁵ [Article 3, the Digital Services Act \(DSA\)](#).

aktiva användare förändrades detta. En tredje part – plattformsföretagen – klev in varmed systemet för reglering transformerades från dualistiskt till pluralistiskt. I fråga om innehållsmoderering online utgörs systemet av (minst) tre parter:

- Plattformsföretagen som äger och styr tjänsterna – de digitala arenorna – som användarna nyttjar för att yttra sig och konsumera innehåll. Detta är privata företag och relationen mellan dem och användarna styrs av kontrakt, *T&C* (*Terms and Conditions*) inklusive *Community Guidelines* som reglerar hur plattformarna/tjänsten får användas och vilket innehåll som är tillåtet.
- Användarna som nyttjar plattformarna för att yttra sig, sprida sitt eget och andras innehåll samt ta del av olika slags innehåll. De kan även anmäla olagligt innehåll till plattformen och polisen.
- Regulatorerna på nationell eller överstatlig nivå, t.ex. EU. Här ingår också icke-statliga organisationer eller sammanslutningar som, under vissa omständigheter, kan dela regleringsansvaret med statliga aktörer och samarbeta med branschen. Exempelvis kan civilsamhällesorganisationer bidra i framtagandet av *best practices*.

Regleringen av plattformarnas innehåll är således dels privat självreglering, dels statlig men däremellan finns också samreglering (*co-regulation*) tillsammans med, eller genom, tredje parter som civilsamhällesorganisationer vilka bistår företagen och staten på olika sätt. Ett exempel är svenska Källkritikbyrån som verkar som oberoende tredjeparts-fakta-granskare inom Metas plattformar Facebook och Instagram. Uppdraget innebär att Källkritikbyrån självständigt granskar innehåll som sprids på plattformarna och där det finns skäl att ifrågasätta riktigheten. Granskningarna genomförs enligt etablerade journalistiska metoder och resultaten, inklusive tillvägagångssätt och resonemang, publiceras öppet på Källkritikbyråns webbplats så att allmänheten kan ta del av bedömningen. Meta kan välja att märka det aktuella inlägget och hänvisa användare till den fullständiga faktagranskningen, men det är Meta som tar beslut om och ansvarar för alla tekniska och modererande åtgärder. Tredjeparts-granskarens roll är alltså endast rådgivande.

Regleringen av plattformarnas innehåll sker således genom en sorts trepartsmodell: plattformsföretagen, staten/överstatliga organ och tredjeparter såsom branschorganisationer, civilsamhällesorganisationer m.fl. Det är, som Carretta beskriver det, ett regulatoriskt kontinuum där parterna gemensamt formar plattformarnas innehåll och därmed det offentliga samtalet.

Eftersom plattformarna regleras av rådande lagstiftning på statlig och överstatlig nivå (EU) är de ålagda att moderera innehåll som klassificeras som olagligt. De modererar användargenererat innehåll på statens vägnar och är i den bemärkelsen en mellanhand, eller förlängd arm, mellan staten och användarna. Samtidigt har de möjligheten att genom sina *T&C* reglera användningen och moderera innehållet

ytterligare så att det är i enlighet med plattformarnas egna användarvillkor. Det är således både reglering ”of platforms” och ”by platforms” (Carretta, 2025).

6.1.2 Lagligt, olagligt eller oönskat innehåll

Innehållet på plattformarna kan delas in i tre kategorier:

1. Olagligt innehåll
2. Lagligt innehåll
3. Lagligt men oönskat innehåll (”lawful but awful content”, Carretta, 2025 s. 57).

Den första kategorin måste plattformsföretagen ta bort. Kortfattat är principen att det som är olagligt offline också är olagligt online. Den andra kategorin är lagligt och oproblematiskt innehåll medan den tredje kategorin är mer problematisk. Detta är innehåll som inte är olagligt men som plattformarna ändå vill moderera då det strider mot deras *T&C* och *Community Guidelines*. Plattformarna har friheten att formulera egna regler för att hantera och ta bort lagligt innehåll från sina tjänster om de anser att innehållet inte är önskvärt, och det är särskilt detta utrymme som blivit omdiskuterat (Carretta, 2025 s. 60ff). Vad som betraktas som ”oönskat” är till stor del en normativ fråga där svaren varierar mellan olika kulturella kontexter, mellan olika gemenskaper och ibland även från person till person. Att formulera riktlinjer som ska accepteras och följas i många olika länder kan vara utmanade (Gerrard, 2022). Plattformarna kan dessutom moderera innehåll baserat på kommersiella överväganden för att inte förlora intäkter från annonsörer. Moderering är ett verktyg för att forma en produkt/tjänst som går att sälja till annonsörer och användare. All moderering handlar alltså dels om laglighet, dels om affärsmässighet (Gorwa et al., 2020; Carretta, 2025; Lindgren, 2024).

6.1.3 Innehållsmoderering i praktiken

Moderering sker i praktiken genom en hybrid av människor och maskiner: användarrapportering, mänskliga moderatorer, maskininlärningsystem, expertpaneler, faktagranskare, *trusted flaggers*, interna policyteam och ibland externa organ som *Oversight Board* eller tvistlösningsinstanser enligt DSA. I de tidiga internet-forumen var det ofta deltagarna själva som åtog sig rollen som moderatorer för sitt ”community” och även idag förekommer detta i viss mån, på t.ex. Reddit, men det är alltmer sällsynt (Gorwa et al., 2020). Det är mer förekommande att företagen anställer personer (*in-house* eller *outsourcing*) för att manuellt gå igenom och moderera innehåll och de kan också träna upp automatiserade system för moderering. En kombination av automatiserad och mänsklig moderering verkar vara det vanligaste på de stora plattformarna idag, där automatiserade system hittar och prioriterar innehåll medan människor hanterar svåra gränsfall (Gorwa et al., 2020; Gerrard, 2022).

All moderering är dock plattformsspecifik (Gerrard, 2022). Hur mänsklig och automatiserad moderering kombineras varierar alltså mellan plattformarna och tittar

man på de offentliga dokument som Meta, Tiktok, Reddit och Flashback publicerat finns både likheter och skillnader. Både Meta och Tiktok använder regelstyrd moderering utifrån *Community Standards/Guidelines*, teknik för automatiserad upptäckt och prioritering av innehåll, men även stora team av mänskliga moderatorer. Båda har rapporteringsfunktioner från användare. Tiktok uppger att de modererar innehåll både med avseende på om det ska tas bort/vara kvar och hur stor räckvidd innehållet ska få genom att rekommenderas i *For You*-flödet.⁶ Även Meta arbetar med nedprioritering i stället för borttagning av innehåll i vissa fall.⁷

Reddit beskriver sin modell som en lager-på-lager-moderering som är community-driven. Det innebär att det finns både globala Reddit *Rules* och lokala subreddit-regler, där frivilliga moderatorer (*mods*) gör en stor del av det dagliga modereringsarbetet, medan Reddits egna anställda (*admins*) hanterar plattformsövergripande regelbrott. Här används även automatiserad innehållsmoderering.⁸ Reddit skiljer sig därmed från Meta och Tiktok genom att mycket av modereringen är decentraliserad till communities. Även svenska Flashback skiljer sig från Meta och Tiktok genom att fungera mer som ett klassiskt forum med regelverk och moderatorer.⁹

Denna beskrivning av ett urval plattformar är ytterst översiktlig och skulle man gå ner mer på detaljnivå kommer allt fler skillnader mellan plattformarna att framträda. Det gäller inte bara hur modereringen görs utan även hur olika plattformar förhåller sig till moderering av innehåll i relation till olika slags problem, såsom näthat, upphovsrättsbrott och desinformation. Schaffner med kollegor (2024) har gjort en systematisk genomgång av policytexter gällande moderering från 43 engelskspråkiga plattformar som visar på en betydande variation mellan plattformar i både struktur och innehåll i reglerna, vilket gör modereringslandskapet svåröverskådligt och rättssäkerheten ojämn. Det går alltså inte att dra slutsatser om plattformarnas moderering generellt eftersom de sätt på vilka modereringen fungerar varierar dem emellan. Detta är viktigt att ha i åtanke eftersom det innebär att även den enskildes ställning i förhållande till plattformarna och dennes möjlighet att komma till uttryck där varierar. Tittar man närmare på vad automatiserad innehållsmoderering innebär i praktiken blir det ännu mer komplicerat (se avsnitt 6.1.5).

6.1.4 Innehållsmoderering i praktiken – i Sverige

Mediemyndigheten har kontaktat Meta, Google, Tiktok, X, Reddit och Flashback för att inhämta uppgifter om modereringsåtgärder som berör användare i Sverige under 2025. Syftet har varit att få en samlad bild av omfattning, typer av åtgärder samt hur beslut fattas och följs upp. Svar har kommit in från Tiktok och Flashback. X har

⁶ tiktok.com/euonlinesafety/en/.

⁷ [Sharing results: Community Standards Enforcement Report | Transparency Center](#)
[More Speech and Fewer Mistakes.](#)

⁸ [Content Moderation, Enforcement, and Appeals – Reddit Help.](#)

⁹ [Regler - Flashback Forum.](#)

avstått från att besvara frågorna, medan övriga tillfrågade plattformar inte har återkommit.

Tiktok uppger att plattformen under 2025 genomförde ett mycket stort antal modereringsåtgärder som berörde svenska användare. Åtgärderna avsåg huvudsakligen enskilt innehåll, medan en mindre andel riktades mot konton. Plattformen redovisar också att modereringen av videor i hög grad sker automatiserat. Däremot har motsvarande uppgifter inte redovisats för alla andra typer av åtgärder. Tiktok beskriver vidare att användare i vissa fall kan få information om och överklaga modereringsbeslut via plattformens egna funktioner. Sammantaget ger svaren en relativt omfattande kvantitativ bild av modereringen, men redovisningen är inte fullständig i alla delar.

Flashback beskriver en moderering som i större utsträckning är knuten till forumets struktur, innehållskvalitet och relevans i trådar. Plattformen redovisar uppgifter om borttagna inlägg och flyttade trådar samt uppger att flera olika modereringsåtgärder används, bland annat radering av inlägg, flytt och låsning av trådar samt tillfälliga och permanenta avstängningar. Däremot har Flashback inte lämnat sammanställd statistik om automatisering, överklaganden eller ändrade beslut. Underlaget ger därför en mer begränsad och delvis kvalitativ bild av modereringen än den som Tiktok redovisar.

Sammantaget uppvisar Tiktok och Flashback tydligt olika sätt att moderera, vilket bekräftar plattformarnas ovan nämnda offentliga beskrivningar. Det speglar även skillnader i plattformarnas struktur och innehållstyper. Tiktoks moderering sker i mycket stor skala och är i hög grad automatiserad, medan Flashbacks moderering framstår som mer begränsad och i huvudsak manuellt driven.

6.1.5 Automatiserad innehållsmoderering

Automatiserad innehållsmoderering innebär att plattformarna, med eller utan hjälp av AI, skapar system som lär sig känna igen olagligt, eller lagligt men oönskat, innehåll och förhindrar att det laddas upp på plattformen eller tas bort så fort det upptäcks på plattformen (Carretta, 2025 s. 67f). Systemen klassificerar användargenererat innehåll baserat på matchning och prediktion (Gorwa et al., 2020). Detta sker bland annat genom att jämföra innehåll med tidigare klassificerat olagligt/lagligt oönskat innehåll, regelbaserade filter för t.ex. blockerade ord och fraser eller förbjudna länkar samt maskininlärning för textanalys där modellen inte bara tittar på enstaka ord utan på sammanhang, ton med mera, liknande tekniken bakom stora språkmodeller. Det finns även system som kan analysera bilder och video för att t.ex. upptäcka nakenhet och olika slags symboler. Ljud kan transkriberas och analyseras som text. En viktig teknik är hash-matchning där ett digitalt ”fingeravtryck” av en bild, video eller fil skapas i avsikt att snabbt upptäcka om samma eller liknande material laddas upp igen. Detta används bland annat för att identifiera upphovsrättsskyddat material, känt olagligt material och tidigare borttaget innehåll. Systemen kan också analysera beteenden i relation till innehåll och upptäcka mönster som kan tyda på spam eller botar, såsom

massutskick av identiska meddelanden, automatiserade gillningar eller delningar. Modereringssystem kan också väga in vem/vilket konto som publicerat innehållet och kontots historik samt andra användares rapportering av innehåll och konton (Gorwa et al., 2020; Carretta, 2025; Drolsbach & Pröllochs, 2024; Karabulut et al., 2022; Gerrard, 2022; Saleous et al., 2025).

Under senare år har AI använts alltmer inom innehållsmoderering då det hjälper plattformarna att få bort problematiskt innehåll snabbt på ett resurseffektivt sätt. Givet den enorma mängden material som laddas upp på plattformarna varje dag har automatisering genom AI-baserade algoritmer blivit nödvändig (Carretta, 2025 s. 70). Att moderera allt material manuellt låter sig inte göras och det finns dessutom problem förknippat med mänsklig moderering, inte minst bristande arbetsvillkor för moderatorerna och rapporter om psykisk ohälsa till följd av den exponering de utsatts för (Lindgren, 2024; Steiger et al., 2021). Men även den algoritmiska modereringen har brister och problem.

Innehållsmoderering är generellt svårgenomträngligt och inte så lätt att få grepp om. Detsamma kan sägas om AI och när de kombineras uppstår det som brukar kallas för *the black box problem* eller ”The best-kept secret in tech” (Gerrard, 2022). Även om system och processer kan beskrivas på en övergripande nivå är det mycket svårt, eller omöjligt, att göra det i det enskilda fallet. Användaren vet vad den har laddat upp och kan se om det gillas, delas eller, tvärtom, tas bort men inte hur systemet har processat innehållet (Carretta, 2025; Gorwa et al., 2020). Denna brist på transparens gäller även för forskare, journalister och andra som borde kunna granska processerna och det är bland annat för att motverka denna bristande transparens som EU har tagit fram DSA (se vidare avsnitt 10, bilaga). Trots DSA kommer det att kvarstå svårigheter för forskare att studera systemen som tekniska objekt eftersom de betraktas som affärshemligheter och är tekniskt komplicerade på sätt som kräver specialistkompetens för att förstå. De är dessutom under ständig förändring (Gandini et al., 2022). Det betyder dock inte att systemen inte alls kan studeras och diskuteras kritiskt.

Den automatiska modereringen är inte, och kan inte bli, helt ofelbar. AI tränas på material för att lära sig skilja mellan innehåll som sannolikt är olagligt och/eller oönskat och ska tas bort, respektive innehåll som får vara kvar på plattformen. Här finns risk för ”falska positiva”, där legitimt innehåll så att säga testas positivt för olagligt/oönskat och tas bort på felaktiga grunder, eller tvärtom: risk för ”falska negativa”, där skadligt innehåll släpps igenom (Carretta, 2025 s. 71ff; Gerrard, 2022; Gorwa, 2020; Peterson-Salahuddin, 2024).

Den automatiska modereringen kan även ha svårigheter med ironi, satir och citat eftersom kontexten kring ett ord, fras eller uttryck försvinner eller feltolkas av AI. Att ord, fraser eller uttryck inte ses i sitt sammanhang innebär även svårigheter med att tolka kulturella och politiska nyanser vilket bland annat kan bidra till orättvis eller

ojämlik behandling av olika grupper av användare (Gorwa et al., 2020; Peterson-Salahuddin, 2024; Steen et al., 2023; Balendra, 2025).

En annan välkänd brist är bias i träningsdata. Eftersom AI tränas på befintliga, människoskapade data är den behäftad med samma brister och problem och kan reproducera dessa. Exempelvis fördomar och stereotypa föreställningar om olika grupper eller kulturer (Lindgren, 2024; Peterson-Salahuddin, 2024).

Även mindre språk kan vara problematiska för automatiserad moderering eftersom det finns mindre material att träna AI på (Peterson-Salahuddin, 2024; Balendra, 2025).

Forskare pekar således på flera problem med algoritmisk moderering vilka sammantagna försvårar insyn och begriplighet, genererar frågor om hur rättvisa systemen egentligen är samt hur de riskerar att dölja och avpolitisera beslutsprocesser som i allra högsta grad är politiska (Gorwa et al., 2020). Forskningen visar även att *hur* moderering görs har betydelse för om modereringsbesluten uppfattas som rimliga eller orättvisa. Surveyundersökningar och experimentella studier har visat att modereringsbeslut uppfattas som mer legitima om de fattas med transparenta processer, expertis och fungerande överprövning än om besluten enbart görs av AI, men resultaten är över lag tämligen spretiga här (Martel et al., 2025; Pan et al., 2022; Bhandari et al., 2021; Jhaver et al., 2023; Molina & Sundar, 2022).

För användarna kan modereringen medföra flera konsekvenser, bland annat (Gerrard, 2022):

- Inlägg ges etiketter/innehållsvarningar
- Innehåll tas bort
- Avstängning från plattformen
- Avstängning från partnerprogram och dylikt (demonetisering)
- *Hashtag bans* (vissa hashtags förbjuds och tas bort)
- *Shadowbanning/visibility reduction* (konton eller inlägg tas bort från rekommendationssystem eller får begränsad synlighet på andra sätt).

Att bli avstängd från plattformen kan betraktas som den hårdaste konsekvensen av moderering men även avstängning från partnerprogram kan ge kännbara ekonomiska konsekvenser för användare som har intäkter därifrån. Att helt ta bort innehåll är också en ”hård” form av moderering medan den sista punkten, *shadowbanning*, är mindre ingripande – om än inte oproblematiskt, vilket rapporten återkommer till i avsnitt 6.2.1. Innan dess beskrivs och diskuteras systemen som tvärtom syftar till att lyfta fram och synliggöra vissa innehåll och konton.

6.2 Algoritmisk kurering och rekommendationssystem

Rekommendationssystem definieras i DSA (artikel 3 s) som ett helt eller delvis automatiserat system som används av en onlineplattform för att i sitt gränssnitt föreslå specifik information till tjänstemottagare eller prioritera sådan information, inbegripet till följd av en sökning som initierats av tjänstemottagaren eller genom att på annat sätt fastställa den relativa ordningen eller framträdande placeringen för den information som visas.¹⁰ Eftersom rekommendationssystemen styr vilket innehåll som får framträdande placering, eller tvärtom, kan de betraktas som ytterligare en aspekt av plattformsmodereringen (Gillespie, 2022).

Rekommendationssystem fungerar i praktiken som prediktions- och rangordningssystem: de försöker förutse vilket innehåll en användare sannolikt kommer att titta på, klicka på, gilla, kommentera, dela eller stanna kvar vid. Det gör att systemet inte bara visar ”relevant” innehåll, utan prioriterar innehåll som bedöms skapa fortsatt uppmärksamhet och interaktion. Rekommendationssystemen är alltså inte en enda ”algoritm”, utan ett samspel mellan rankingsmodeller, tidigare användarbeteende, innehålls egenskaper, kontots tidigare genomslag och andra plattformsspecifika variabler (Corsi, 2024; Guess et al., 2023; Huszár et al., 2022).

Eftersom plattformarnas rekommendationssystem av affärsmässiga skäl inte är offentliga vet vi inte exakt hur de fungerar. Det går därmed inte att studera direkta effekter av rekommendationssystemen. Forskningen på detta område är alltså, av praktiska skäl, starkt begränsad. Det finns dock en del experimentella studier där man påverkat systemen (t.ex. Huszár et al., 2022) eller laborerat med olika input för att testa vad systemen prioriterar (t.ex. Corsi, 2024). Resultaten visar främst att engagerande innehåll och konton med stor räckvidd gynnas. Inlägg från konton med många följare och hög interaktion har således större sannolikhet att rankas högt av systemet (Corsi, 2024). Det finns även studier som visar att systemen tenderar att prioritera politiskt högerlutande nyhetskällor och inlägg (Huszár et al., 2022; González-Bailón et al., 2023) men det går inte att dra generella slutsatser av dessa studier eftersom resultaten sannolikt kommer att variera mellan plattformar, tidpunkter och kontexter.

Innan rekommendationssystemen och de personliga flödena togs fram var flödena oftast kronologiska. Det vill säga att det senaste inlägget hamnade först, högst upp på sajten, och visades på ungefär samma sätt för alla. Detta gav goda förutsättningar för synlighet och uppmärksamhet. Omkring 2013 förändrades detta när plattformarna övergick till den mer personliga algoritmiska kureringen av innehåll (Williams, 2025). Den sannolikt främsta anledningen till systemförändringarna är den så kallade uppmärksamhetsekonomi som plattformarna verkar inom. Eftersom utbudet av

¹⁰ [Article 3, the Digital Services Act \(DSA\)](#).

innehåll idag är enormt råder det hård konkurrens om publikens uppmärksamhet och varje plattform eller annan aktör får arbeta hårt för att fånga och behålla publikens uppmärksamhet. Rekommendationssystemen är alltså utformade för att göra just detta (Zuboff, 2021).

I en amerikansk experimentell studie av Instagram och Facebook ändrade forskarna flödet för ett urval deltagare, från personligt, algoritmstyrt till kronologiskt. Resultaten visade att deltagarnas tid och aktivitet på plattformarna minskade och att innehållet de exponerades för förändrades. Andelen politiskt innehåll ökade, liksom andelen otillförlitligt material, men ideologiskt blev innehållet mer blandat. Trots dessa förändringar kunde studien inte belägga några signifikanta effekter på deltagarnas attityder under de tre månader som studien pågick (Guess et al., 2023). En annan storskalig amerikansk experimentell studie på Facebook kom till liknande slutsatser. En förändring av flödet (i detta fall en begränsning av inlägg från politiskt likasinnade) påverkade inte deltagarnas attityder (Nyhan et al., 2023). Trots att själva exponeringen ändrades tydligt förefaller evidensen för direkta kortsiktiga attitydeffekter vara svag. Rekommendationssystem formar således vilka röster som får höras och vilka informationsmiljöer individer rör sig i, men det är empiriskt svårare att visa att de på kort tid ändrar politiska övertygelser hos genomsnittsanvändaren.

Det betyder dock inte att systemen saknar betydelse. För det första kan små exponeringsskillnader bli samhälleligt betydelsefulla om de verkar på miljontals användare över lång tid. För det andra kan vissa grupper vara mer påverkbara än genomsnittet. Här skulle det behövas longitudinella studier som följer användare under längre tid, men även i olika länder och under olika förhållanden för att man ska kunna säga något mer precist om effekter på människors attityder och uppfattningar. För det tredje kan effekterna ligga i plattformarnas dagordningsfunktion (*agenda-setting*) snarare än i omedelbara förändringar av politiska preferenser eller attitydskalor: vilka sakfrågor och samhällsproblem som uppmärksammas mest i mediernas flöden kan påverka vad människor uppfattar som viktigt (Einarsson et al., 2025). En slags algoritmisk *nudging* (Yu et al., 2024). För den enskildes möjlighet att bilda opinion i en fråga är detta högst relevant.

6.2.1 Dold nedprioritering (shadowbanning)

Begreppet *shadowbanning* har använts sedan åtminstone början av tjugohundratalet och syftade då på mekanismer som osynliggjorde inlägg för alla användare av ett internetforum utom för den som postat inlägget. Begreppet fick större spridning 2018 när konservativa Twitter-användare i USA anklagade plattformen för att medvetet begränsa vissa republikaners synlighet (Savolainen, 2022; Delmonaco et al., 2024) och under de senaste åren har ett forskningsfält börjat växa fram. Idag används begreppet för att beskriva hur plattformarna på olika sätt begränsar vissa inläggs eller kontons/användares synlighet, utan att informera användaren om det. Inlägget kan

alltså finnas kvar, men det syns sämre i sökningar, rekommendationer, hashtaggar, kommentarsfält eller flöden. Det kan beskrivas som motsatsen till den förstärkning av innehåll och konton/användare som rekommendationssystemen syftar till, men det är inte en bieffekt av rekommendationssystemet utan sker medvetet. Forskningen betonar dock att begreppet är omstritt och att plattformar sällan använder termen själva, utan i stället talar om *visibility reduction*, *demotion*, *borderline content policies* eller liknande (Risius & Blasiak, 2024; Delmonaco et al., 2024).

Shadowbanning, eller dold nedprioritering, är inte *en* teknik utan en samling av tekniker varav fyra huvudtyper, enligt Risius & Blasiak (2024): *ghost bans* (innehåll blir osynligt för alla andra än kontoinnehavaren, Jaidka et al., 2023), *search bans* (kontot/innehållet försvinner ur plattformens sökindex), *search suggestion bans* (kontot föreslås inte i sökningar, Goldman, 2021) och *downtiering* (innehållets synlighet reduceras eller begränsas utan att tas bort, Jaidka et al., 2023).

Utifrån plattformarnas perspektiv kan dessa tekniker användas för att hantera lagligt men oönskat innehåll, alltså sådant innehåll som inte är olagligt eller tydligt bryter mot plattformens användarvillkor men som plattformen ändå inte vill ge stor exponering. Det kan gälla misstänkt spam eller innehåll som kan skada människors hälsa och välbefinnande, exempelvis drogrelaterat innehåll eller innehåll som kan främja ätstörningar eller självskadebeteende (Chancellor et al., 2017; Thorell et al., 2026). Shadowbanning kan alltså vara motiverat.

En annan fördel med att nedprioriteringen sker dolt är att plattformarna kan hantera oönskat material utan att avslöja exakt vilka tekniker de använder och därmed förhindra att potentiellt illvilliga aktörer lär sig kringgå teknikerna (Risius & Blasiak, 2024). Forskare som studerat online-gemenskaper kopplade till ätstörningar och självskadebeteende har dock sett hur användare utvecklar ett alternativt språk för att kringgå automatiserad upptäckt och nedprioritering, så kallat *algospeak*¹¹ (Lookingbill & Le, 2024; Gerrard, 2018). Detta har även iakttagits i andra marginaliserade grupper (Steen et al., 2023).

Samtidigt har bristen på transparens kring *shadowbanning* – inklusive diskussionen om det överhuvudtaget existerar och plattformarnas förnekelse därav – givit upphov till både folkteorier om systemen och personligt lidande bland användare som upplever sig vara diskriminerade av algoritmerna. Savolainen (2022) beskriver hur användare utvecklar en sorts “algorithmic folklore” (folkteorier) för att försöka förstå och förklara varför räckvidd, sökbarhet eller exponering plötsligt minskar. Register med flera (2023) visar på liknande sätt hur Instagram-användare kollektivt skapar mening kring den upplevda algoritmiska oförutsägbarheten, osäkerheten och risken att

¹¹ En vokabulär med kodord som används för att undvika automatiserad moderering av inlägg på sociala medier (t.ex. *sewerslide* i stället för *suicide*). Det möjliggör fortsatt spridning av skadligt innehåll trots säkerhetsåtgärder på plattformarna. (Thorell et al., 2026).

“straffas”, vilket leder till vaksamhet och försök att blidka algoritmen. Karizat med flera (2021) menar att Tiktok-användare utvecklar folkteorier om att plattformen filtrerar bort vissa sociala identiteter – de icke algoritmiskt privilegierade – vilket påverkar hur de uttrycker sig och betar sig på plattformen för att anpassa sig till (idén om) de algoritmiskt privilegierade och undvika algoritmisk marginalisering.

Cotter (2023) använder begreppet *black box gaslighting* för att beskriva den epistemiska konflikt som uppstår när användare upplever att de har utsatts för dold synlighetsstyrning men plattformen nekar till att shadowbanning existerar och i stället hänvisar till teknisk komplexitet och slumpen. Eftersom plattformarna är de enda som har insyn i systemen har de kunskapsövertag och tolkningsföreträde varmed användarens erfarenhet ogiltigförklaras. Själva tvisten om huruvida shadowbanning existerar blir en del av plattformarnas maktutövning, menar Cotter (2023).

De asymmetriska makt- och kunskapsförhållandena mellan användare och plattformar har även föranlett att användare gått samman och kollektivt försökt experimentera och testa folkteorier om hur algoritmerna fungerar för att ”bevisa” förekomsten av shadowbanning: *collaborative algorithm investigation* (Delmonaco et al., 2024). Det finns alltså en hel del kvalitativ forskning om hur användare upplever, tolkar och förhåller sig till dold synlighetsstyrning och den visar hur användare försätts i tillstånd av osäkerhet och ovisshet där de försöker förstå hur de egna erfarenheterna hänger samman med regler och mekanismer som de aldrig fullt ut får tillgång till.

Flera studier lyfter även att den dolda synlighetsstyrningen ses som ojämlik eller orättvist fördelad då marginaliserade grupper upplever sig råka ut för det i högre grad än andra. Plattformarnas modereringspraktiker kritiserar för att vara diskriminerande då de i tysthet nedprioriterar innehåll baserat på användares ras/etnicitet, kön, sexuella läggning och funktionsnedsättningar, med både ekonomiska och psykologiska konsekvenser för användarna (Are, 2022; Delmonaco et al., 2024; Kojah et al., 2025; Willcox, 2025).

Men det finns även exempel på motståndshandlingar bland användare. Förutom tidigare nämnda *collaborative algorithm investigation* och *algospeak* använder forskare begreppet *algorithmic resistance* för att beskriva hur användare aktivt försöker motarbeta algoritmer genom att exempelvis ändra profilegenskaper, dela tips om hur man kan undgå dold nedprioritering eller dela, gilla och på andra sätt försöka förstärka ett innehålls synlighet som man misstänker har blivit automatiskt nedprioriterat (Delmonaco et al., 2024; Velkova & Kaun, 2019; Karizat et al., 2021). *The black box problem* och bristen på transparens kan således skapa tillitsproblem och dragkamp mellan användare och plattformar samtidigt som det reser berättigade frågor om rättvisa och makt.

Det finns ett fåtal empiriska studier som försökt belägga shadowbanning med vetenskapliga metoder. En storskalig analys av Twitter-konton fann mönster i

nedprioriteringen som indikerar att den inte sker slumpmässigt eller är resultat av någon tillfällig ”bugg” (Le Merrer et al., 2021). Jaidka och kollegor (2023) granskade omkring 25 000 amerikanska Twitter-konton och fann att shadowbans över lag var relativt sällsynta, men att de inte var slumpmässiga: konton med bot-liknande beteende drabbades oftare, liksom nya konton med lågt socialt inflytande. Verifierade konton nedprioriterades mer sällan, liksom konton med många följare och tweets som brukar skapa stort engagemang (stort socialt inflytande). Även vissa typer av politiska eller kränkande svar på inlägg blev oftare nedprioriterade. Studierna visar därmed att dold synlighetsstyrning kan analyseras systematiskt, men det verkar bara ha gjorts på Twitter. Detta förklaras sannolikt av att Twitter historiskt har erbjudit öppna API:er och datatillgång, vilket gör kvantitativa studier möjliga. De kvalitativa användarstudierna som tidigare nämnts handlar oftare om Instagram och Tiktok.

6.3 Innehållsmodereringens för- och nackdelar för det demokratiska samtalet

Frågan om innehållsmodereringens konsekvenser för demokratin har fått mycket uppmärksamhet inom forskningen och den samlade bilden pekar på att innehållsmoderering är demokratiskt nödvändig men demokratiskt riskfylld. På en övergripande nivå ger modereringen plattformsföretagen stor makt över offentligheten utan att de behöver ta fullt ansvar eller ge full insyn i sina system (Gandini et al., 2025; Aytac, 2024; Caplan, 2023; Reimer & Peter, 2021). Å andra sidan kan större statlig inblandning förvandla moderering från skydd till repression. Det är alltså en tvetydig bild, men några fördelar för det demokratiska samtalet verkar det råda konsensus om:

- Moderering kan minska hat, trakasserier och desinformation som annars kan tysta eller skrämja bort deltagare från offentlig debatt (Vasist et al., 2023; Gerrard, 2022; Kashyap & Formosa, 2023).
- Moderering kan skydda särskilt utsatta grupper och därmed göra samtalet tryggare, mer tillgängligt, inkluderande och jämlikt, särskilt för grupper som annars riskerar att drabbas av hat och trakasserier. Vissa normativa studier argumenterar för att plattformarna har ansvar för att ingripa mot innehåll som skadar andra eller gör det svårare för människor att delta i samtalet på lika villkor (Peterson-Salahuddin, 2024; Kashyap & Formosa, 2023).
- Moderering kan förbättra samtalets kvalitet genom att begränsa innehåll som är manipulativt, grovt vilseledande och destruktivt. Moderering är ett sätt att skydda det demokratiska samtalet mot falsk eller manipulativ information, särskilt om det

sprids storskaligt eller förstärks algoritmiskt (Acosta Navas, 2025; Dwork et al., 2024; Scharfbillig et al., 2026).

- Genom att använda dold nedprioritering och inte berätta exakt hur systemen fungerar minskar risken att olika aktörer lär sig hur man bäst kringgår modereringssystemen för att exempelvis sprida desinformation (Carretta, 2025; Risius & Blasiak, 2024).

Forskningen lyfter dock även flera nackdelar och problem för det demokratiska samtalet. Ett problem är risken för överblockering eller underblockering av innehåll. Plattformarna har en juridisk skyldighet att ta bort illegalt innehåll och använder AI-baserade verktyg för att identifiera och blockera sådant innehåll. Systemen används även för att identifiera och blockera den mer flytande kategorin av oönskat innehåll. När plattformarna agerar proaktivt tillämpar de ibland en ”block-first, question-later approach” vilket riskerar leda till överblockering av lagligt innehåll och därmed inskränkning av användarnas rättigheter till information och yttrandefrihet (Carretta, 2025 s. 76).

Som nämnts kan algoritmisk moderering även missa kontextens betydelse och drabba marginaliserade röster oproportionerligt. Flera empiriska studier visar att automatiserad moderering ofta är dålig på kontext, ironi, subkulturella uttryck och minoritetsspråk, vilket leder till både falska positiva och falska negativa (Carretta, 2025). Algoritmer kan alltså både missa verkligt skadligt innehåll och samtidigt felaktigt stoppa legitimt innehåll, särskilt från marginaliserade grupper (Gorwa et al., 2020; Steen et al., 2023; Golunova, 2025).

Detta kan undergräva det demokratiska samtalets legitimitet och dess funktion som självreglerande mekanism, enligt bland annat Williams (2025). En fundamental princip i demokratiska samhällen är antagandet att det egna systemet och beslutsfattandet inte är ofelbart – både människor och system kan ibland hamna snett och därför behövs självreglerande mekanismer, så kallade ”checks and balances” som kan korrigera. En fri press och yttrandefrihet är exempel på sådana mekanismer eftersom det möjliggör granskning och debatt där olika idéer, argument och perspektiv kan mötas och stötas mot varandra för att underbygga ett rationellt beslutsfattande (Acosta Navas, 2025; Williams, 2025). Detta har bland annat beskrivits i den tyske filosofen Habermas välkända teori om ”the public sphere” (1962/1984). En annan fundamental princip är att det offentliga samtalet ska vara inkluderande, att alla medborgare ska ha samma rättigheter att yttra sig och delta i debatten (samt rösta) för att beslutsfattandet ska betraktas som legitimt. Om vissa grupper diskrimineras algoritmiskt och om inte alla grupper har möjlighet att göra sin röst hörd på lika villkor undergrävs det demokratiska samtalets legitimitet (Williams, 2025). Politiskt eroderar det legitimiteten för det demokratiska beslutsfattandet eftersom principen om jämlikt deltagande sätts ur spel (Williams, 2025).

Det riskerar också att försvåra det offentliga samtalets funktion som korrigerande mekanism, om vissa grupperns perspektiv inte får ta plats. Det offentliga samtalet blir fattigare, samtidigt som andra delar av publiken går miste om information och ny kunskap (Williams, 2025).

Den korrigerande funktionen kan dock även sättas ur spel om desinformation och propaganda släpps igenom och om botar används för att förstärka vissa perspektiv som då framstår som vanligare, eller populärare, än vad de egentligen är, en ”bandwagon-effekt” (MPF, 2018). Eftersom personer som upplever att de tillhör en majoritet är mer benägna att uttrycka sina åsikter än andra kan detta påverka personers vilja yttra sig. Personer som inte håller med om den åsikt som framstår som majoritetens tenderar att vara mindre benägna att yttra sig och härmed skapas en tystnadsspiral.¹² Botar och annan teknisk manipulation av informationsflöden i avsikt att skapa sken-majoriteter kan alltså dels ge en helt felaktig bild av den faktiska opinionen, dels påverka den enskildes vilja att delta i opinionsbildning, antingen positivt eller negativt (Gandini et al., 2025). För lite moderering kan – precis som snedvriden moderering – skapa problem (Acosta Navas, 2025).

Ett annat problem är att plattformsföretagens primära syfte inte är att facilitera den demokratiska debatten utan att generera ekonomisk vinst. Innehållsmodereringen sker alltså till stor del av kommersiella skäl (Carmi, 2020; Carretta, 2025; Zuboff, 2021; Williams, 2025) och det som fångar människors uppmärksamhet och håller dem kvar på plattformen är inte nödvändigtvis det som är bäst för demokratin – exempelvis rationella argument och saklighet (Alnemr, 2025; Gandini et al., 2025).

Rekommendationssystemens underliggande logik är att skapa engagemang eftersom det dels skapar innehåll till plattformarna vilket kan dra in fler användare, dels kan ge signaler om vad som intresserar användarna. Denna information kan sedan användas för att styra innehåll så att det blir mer anpassat till individens preferenser, det vill säga personliga flöden. En konsekvens av personliga flöden är att individen primärt exponeras för innehåll som är i linje med det hen brukar intressera sig för, varför befintliga intressen och åsikter snarare befasts än utmanas. Om det offentliga samtalet ska bidra till demokratin förutsätter det att olika åsikter, perspektiv och argument möts, vilket inte underlättas av personliga flöden (Williams, 2025).

Vissa forskare och debattörer menar att den algoritmiska styrningen dragen till sin spets kan skapa filterbubblor där intressen och uppfattningar bekräftas och förstärks (Pariser, 2012) men förekomsten av filterbubblor har även ifrågasatts (Dahlgren, 2021). Snarlika begrepp är diskursiva enklaver (Alnemr, 2025), epistemiska filter (Acosta Navas, 2025) och ekokammare som syftar på sociala miljöer där människor med liknande intressen och åsikter möts och bekräftar varandra. Forskare har pekat på att algoritmer och rekommendationssystem kan bidra till att skapa dessa miljöer,

¹² [NE - tystnadsspiralen](#).

även om de kanske inte skapar hermetiskt tillslutna bubblor (Nguyen, 2020; Törnberg & Törnberg, 2024; Williams, 2025; Gandini et al., 2025; Acosta Navas, 2025). Forskare har även visat att rekommendationssystemen kan bidra till affektiv polarisering genom att förstärka negativa föreställningar och känslor gentemot andra grupper (Acosta Navas, 2025; Staal Bruun Overgaard, 2024). En modell- och simuleringsstudie på Twitter-data har testat om synlighetsbegränsande tekniker kan användas för att förskjuta opinioner utan att märkbart förändra plattformens policy. Resultaten visar att selektiv synlighetsbegränsning teoretiskt kan påverka opinioner, öka eller minska polarisering och samtidigt framstå som neutral för externa observatörer. Den visar alltså inte att en viss plattform gör detta, men ger empiriskt modellbaserat stöd för att tekniska justeringar skulle kunna ha betydande effekter på opinionsbildning (Chen & Zaman, 2024). Eftersom den automatiserade modereringen potentiellt kan ha kraftfulla effekter har åtskilliga forskare efterlyst en mer ansvarsfull hållning från plattformarnas sida. Det finns också forskare som föreslår alternativa sätt att designa plattformarna och de algoritmiska systemen så att de kan bättre främja meningsfull debatt, kunskapsutbyte och användarnas förmåga till eftertanke och rationellt övervägande, eller deras epistemiska resiliens (Gillespie, 2022; Acosta Navas, 2025). Demokratifrämjande moderering kan således inte bara handla om vad som tas bort.

Att modereringssystemen är så ogenomskådliga, svåra att förstå och bevisa är också problematiskt. Dels för att det försvårar möjligheten att ifrågasätta och diskutera besluten och den makt som plattformarna de facto utövar, vilket indirekt stärker deras position än mer. Plattformarnas sätt att fungera, deras underliggande logik, naturaliseras (Gillespie, 2022). Dels då det minskar tilliten till plattformarna och kanske i förlängningen till det offentliga samtalet. Även om plattformarna inte ensamma utgör hela den offentliga arenan har de stor betydelse, vilket innebär att uppfattningar om plattformarna kan spilla över på det demokratiska systemet och på den enskildes vilja att engagera sig och delta där (Dumbrava, 2021). Ytterligare en konsekvens av modereringssystemens ogenomskådlighet är att den genomsnittliga användaren idag inte har möjlighet att avgöra i vilken grad eller på vilket sätt den information hen tar del av är modererad. Som ”vanlig användare” är det i princip omöjligt att avgöra hur den egna användarupplevelsen har kurerats och om man, så att säga, gått miste om väsentlig information man har rätt att ta del av (Carretta 2025, s. 80).

Samtidigt är det svårt att se exakt hur plattformarna ska redogöra för den automatiserade modereringen och i synnerhet den dolda nedprioriteringen, eftersom det är så komplext. Vad som prioriteras ner (eller upp) beror på åtskilliga faktorer och är kumulativa effekter. Det är dessutom i princip omöjligt att veta hur ett inlägg skulle ha färdats på en plattform om den inte hade modererats alls – det finns så att säga ingen ”normalfördelning” att jämföra med. Det plattformarna däremot kan redogöra tydligare för – och förändra – är principerna, intentionerna bakom modereringen (Gillespie, 2022).

Den övergripande bilden i litteraturen är således att innehållsmoderering varken är entydigt demokratifrämjande eller entydigt demokratihämmande. Dess demokratiska värde beror på vad som modereras, hur det görs, vem som beslutar och vilka rättssäkerhetsmekanismer som finns. Med andra ord: det demokratiska idealet i litteraturen är inte “ingen moderering”, utan “legitim moderering”.

Sammantaget kan det ha flera konsekvenser för den enskildes opinionsbildning, men det beror på vem den enskilde är och vilken position personen i fråga har. Socialt inflytelserika användarkonton med många följare och stort engagemang löper lägre risk för dold nedprioritering. För den som positionerat sig på en plattform kan den algoritmiska synlighetsstyrningen vara positiv. Forskningen visar även att användare kan aktivt manipulera eller använda sig av rekommendationssystemen för att synliggöra en fråga eller politiskt perspektiv, så kallad algoritmisk aktivism, eller hashtag-aktivism. Det kan handla om att försöka undgå borttagning av innehåll, som genom tidigare nämnda *algospeak*, men även olika sätt att optimera spridning av innehåll, till exempel om ett mycket litet antal kärnaktörer producerar budskap som sedan förstärks av en stor grupp *amplifiers* som gillar, delar, kommenterar etc. Ett fåtal kärnaktörer kan så att säga sätta agendan medan majoriteten fungerar främst som förstärkare. Detta kan ge synlighet åt perspektiv som annars försvinner i flödena, samtidigt som det (precis som bot-arméer) kan ge intryck av ett brett stöd men i praktiken skapa ojämlig förstärkning i opinionsbildningen (Chong, 2024; Gandini et al., 2025).

För andra användare kan den algoritmiska synlighetsstyrningen vara mer destruktiv, särskilt för marginaliserade grupper. Flera forskare påpekar att *shadowbanning* inte bara handlar om att få begränsad synlighet på plattformarna utan bör förstås som en form av digital tystnad med potentiella effekter på självkänsla, social bekräftelse och psykiskt välbefinnande (Thomas & Manalil, 2025; Are, 2022).

De algoritmiska processer genom vilka publiker och opinioner uppstår, vidmakthålls och förändras kan således ha olika utfall för olika grupper eller personer. Den faktiska möjligheten att delta i samhällsdebatten påverkas dessutom av andra användares beteenden och attityder samt den egna medie- och informationskunnigheten, vilket tas upp i de följande avsnitten (7 och 8).

7 Hur påverkar hot, hat och trakasserier viljan att delta i den digitala samhällsdebatten?

Det är välkänt att samtalsklimatet på sociala medier kan vara hårt. I Sverige har undersökningar visat att både förtroendevalda och journalister utsätts för hot och trakasserier online vilket kan få konsekvenser för viljan att publicera sig och delta i det offentliga samtalet (Björkenfeldt, 2025; Brå, 2025:17). Enligt *Politikernas trygghetsundersökning* utsattes cirka 18 procent av de förtroendevalda 2024 för hot och påhopp i sociala medier (Brå, 2025:17). I en studie av svenska riksdagsledamöter fann Erikson, Håkansson och Josefsson (2023) att skillnaden mellan kvinnor och män gällande hur mycket de utsätts för nättrakasserier är liten. Män som utsattes för trakasserier var däremot något mer benägna att lämna politiken än kvinnor, medan kvinnor i högre utsträckning idkade själv censur och blev mer försiktiga med vad de delade på nätet. Konsekvenserna av utsattheten kan alltså se olika ut.

Det finns färre undersökningar om allmänhetens utsatthet på nätet och i vilken grad det påverkar enskilda personers vilja att delta, men det finns några aktuella svenska exempel.

Internetstiftelsens *Valspecial 2026* visar att 30 procent av de tillfrågade deltagarna i Novus webbpanel (18 år eller äldre) har diskuterat politik på internet under det senaste året. En större andel av männen (53 procent), än kvinnorna (25 procent) har diskuterat politik på nätet, liksom en större andel yngre, förstagångsväljare och politiskt intresserade personer. Totalt 5 procent uppger att de har försökt påverka andra att ändra politiska åsikter på internet under det senaste året och 7 procent uppger att de själva har ändrat åsikt till följd av politiska diskussioner på internet (Internetstiftelsen, 2026).

Det är en relativt liten andel som har upplevt hat och hot i dessa sammanhang. Totalt 8 procent har ”blivit utsatta för hård kritik, hat eller hot när de har delat eller uttryckt politiska åsikter på nätet under det senaste året”¹³, vilket är på samma nivå som inför förra valet år 2022. Utsattheten är dock högre i de grupper som brukar diskutera politik på nätet. Det är också en större andel, drygt 20 procent, som har ”avstått från att uttrycka sina politiska åsikter på nätet för att undvika hård kritik, hat eller hot” och ungefär lika många har ”avstått från att engagera sig politiskt på grund av samtalsklimatet på nätet”. Stiftelsen finner samtidigt att det är vanligare att personer som har diskuterat någon politisk sakfråga också har avstått från att uttrycka sina

¹³ Det kan diskuteras huruvida ”hård kritik” ska blandas ihop med ”hat och hot”. Kritik kan vara hård men saklig och välgrundad, medan ”hat och hot” brukar användas för att beskriva personangrepp och kränkningar som syftar till att skrämja, tysta, sära eller förminska en annan person.

åsikter för att undvika hård kritik, hat eller hot, vilket de menar ”tyder på att många i första hand begränsar sin medverkan i politiska diskussioner på nätet snarare än att avstå från dem helt”. Bland personer som har politiska åsikter klart till höger har en större andel avstått från politiska diskussioner för att undvika hård kritik, hat eller hot. Samtidigt har en större andel personer med politiska åsikter klart till vänster avstått från att engagera sig politiskt på grund av samtalsklimatet på nätet jämfört med personer med åsikter klart till höger (Internetstiftelsen, 2026).

Brottsförebyggande rådets nationella trygghetsundersökning 2025 visar att 21 procent av befolkningen ofta har ”avstått från att skriva något eller lägga upp bilder eller filmer på internet, på grund av oro för att bli utsatt för trakasserier eller hot”. Andelen har legat på ungefär samma nivå sedan 2018. De finner totalt sett ingen skillnad mellan kvinnor och män, däremot att det är vanligare bland äldre än bland yngre att avstå. Om man tittar på härkomst och utbildningsnivå är skillnaderna förhållandevis små (Brå, 2025:15).

Mediemyndigheten har, tillsammans med övriga nordiska länder, genomfört en enkätundersökning om befolkningens medie- och informationskunnighet. Här ingick en fråga om respondenterna ”avstått från att kommentera inlägg på sociala medier av rädsla för påhopp” och de svenska resultaten visar att 26 procent har avstått från detta varje vecka eller varje dag.

En skillnad mellan undersökningarna är att valspecialen ställer frågor specifikt om politiska diskussioner på internet medan Brå:s undersökning gäller aktiviteter på internet i allmänhet. Trots detta blir resultaten näraliggande: omkring 20 procent har avstått från att lägga upp material på nätet, uttrycka sina politiska åsikter eller engagera sig politiskt på grund av oro för att bli utsatt på något sätt. Resultaten i Mediemyndighetens undersökning, som endast gäller att kommentera inlägg, ligger något högre.

7.1 Konsekvenser och åtgärder på individ- och strukturell nivå

I en systematisk litteraturöversikt av internationell forskning om näthatets drivkrafter och politiska konsekvenser visar Kentmen-Cin (2025) att det finns konsekvenser på både individnivå och strukturell nivå. På individnivå främjar näthat (*online hate speech*) fördomar, rädsla, misstro och isolering, och leder ofta till känslor av maktlöshet, minskat socialt förtroende i samhället samt minskat politiskt deltagande. På strukturell nivå fungerar näthat som en slags grindvakt som förstärker stereotyper, normaliserar diskriminering, minskar acceptansen för minoritetsgrupper och kan avskräcka dessa från att driva kontroversiella frågor och göra politisk karriär. Forskningen visar att regelverket rörande näthat varierar mellan länder och mellan plattformar och efterlyser tydligare och mer konsekventa lagar och regler på området. EU:s *Digital*

Services Act lyfts fram som ett bra initiativ då det syftar till att harmonisera regelverket, men samtidigt betonas att även andra insatser behövs, såsom tekniska och utbildningsbaserade. Näthat borde, enligt den samlade forskningen, sanktioneras hårdare men det behövs även insatser för att stärka användares digitala kompetens och motståndskraft för att komma till rätta med problemen (Kentmen-Cin, 2025). Både drivkrafter och konsekvenser finns på såväl individ- som strukturell nivå och därför behöver även åtgärderna ligga på olika nivåer.

8 **Medie- och informationskunnighetens betydelse för opinionsbildning**

Medie- och informationskunnighet (MIK) avser individers förmåga att söka, analysera, kritiskt granska och använda information från olika medier. MIK omfattar även förståelse för mediernas funktion i samhället samt förmågan att hantera olika typer av innehåll, till exempel påverkan och desinformation.

MIK är ett paraplybegrepp för kunskap och färdigheter som är relevanta för vår tids medie-, informations- och kommunikationsmiljöer, om mediernas roll i samhället, om hur och under vilka villkor de arbetar (Carlsson, 2023).

Medier har stor betydelse för vår inhämtning av kunskap, hur vi formar och sprider åsikter, individuellt och i samhället i stort. I ett digitaliserat medielandskap påverkas opinionsbildningen av både traditionella medier, sociala medier, politisk kommunikation och olika former av informationspåverkan (Broda & Strömbäck, 2024; Lilja, Eklund & Tottie, 2024).

MIK-begreppet har ofta likställts med källkritik. Forskarna Haider och Sundin menar att det i tillägg till kritik också måste finnas ett visst mått av tillit och förtroende för att kunna förhålla sig till informationen som möter användaren. De omnämner det som källtillit, att med hjälp av kunskap om hur samhällets institutioner och mediasystemet fungerar kunna göra en bedömning angående vilka källor det går att ha rimlig tillit till (Haider & Sundin, 2022). Haider och Sundin menar att bedömning av information inte bara handlar om att vara kritisk, utan också om hur människor skapar tillit till källor i exempelvis digitala miljöer. Källtillit kan därför förstås som en rimlig och medveten tillit till vissa informationskällor, snarare än blind tilltro eller ett överdrivet ifrågasättande (ibid). Enligt U. Carlsson är både källkritik och källtillit att betrakta som ett resultat av MIK (Carlsson, 2023).

Medie- och informationskunnighet kan ha betydelse på både individnivå och samhällsnivå. På individnivå stärker MIK förmågan att bilda välgrundade åsikter. På samhällsnivå bidrar MIK till en mer motståndskraftig och informerad offentlighet (Nordiska ministerrådet, 2025/2026).

8.1 **MIK:s betydelse för tolkning och värdering av information**

Medie- och informationskunnighet har betydelse för hur individer tolkar och värderar information. Forskning visar att personer med dessa kunskaper har generellt bättre

förmåga att granska källor, bedöma trovärdighet och förstå informationens sammanhang (Ashley, Maksl & Craft, 2017; El Emary & Alhebi, 2025).

En undersökning om hur svenska gymnasieelever bedömer digitala nyheter visar att elevers förmåga att granska källor, jämföra information och värdera trovärdighet varierar, färdigheter som är centrala för digital medborgarkompetens (Nygren & Guath, 2022). Digital kompetens, inkluderande teknisk förmåga att använda medier, ingår i MIK-begreppet men utöver det handlar det också om att kunna förstå och värdera information på ett sätt som ligger till grund för välgrundade åsikter (Nygren & Guath, 2022; Mediemyndigheten, 2024b).

En kvalitativ studie av El Emary och Alhebi visar att individer med hög mediekunnighet i större utsträckning analyserar politisk information kritiskt och är mer uppmärksamma på källors trovärdighet. Detta påverkar i sin tur vilka uppfattningar de utvecklar i samhällsfrågor (El Emary & Alhebi, 2025).

8.2 MIK som skydd mot desinformation och otillbörlig informationspåverkan

Falsk eller vilseledande information som sprids med avsikt att påverka mottagaren benämns som desinformation. Sådan information kan påverka opinionsbildningen genom att förändra människors verklighetsuppfattning, skapa misstro eller förstärka polarisering (Broda & Strömbäck, 2024; FOI, 2022).

Systematiska översikter visar att utbildningsinsatser inom medie- och informationskunnighet ofta förbättrar individers förmåga att identifiera falsk eller vilseledande information och kan fungera som en skyddsfaktor mot desinformation (Anstead et al., 2025; Droog et al., 2025).

Förmågan att skilja mellan tillförlitliga nyheter och desinformation påverkas av faktorer som ålder, utbildning och kön. Insatser med fokus på mediekunnighet förbättrade deltagarnas förmåga att upptäcka desinformation vilket minskar risken att opinioner formas på felaktiga eller manipulerade grunder (Wagnsson & Östervall, 2025; Anstead et al., 2025).

Totalförsvarets forskningsinstitut (FOI) har i en enkätstudie visat att medborgare har störst förtroende för faktabaserade och källkritiska bemötanden av vilseledande information. Detta visar att saklighet och transparens är viktiga för att motverka påverkan och stärka motståndskraften i samhället (FOI, 2022).

I en systematisk litteraturöversikt framgår att forskningen om desinformation, missinformation och falska nyheter har vuxit kraftigt, men att forskningsfältet fortfarande är fragmenterat. De framhåller behovet av mer samlad kunskap om hur

falsk och vilseledande information sprids, upptäcks och påverkar samhällen (Broda & Strömbäck, 2024).

8.3 MIK, medborgerlig kompetens och demokrati

Ett sätt att förstå medie- och informationskunnighet är som en del av bredare medborgerlig kompetens. Medborgerlig kompetens handlar om de kunskaper och färdigheter som behövs för att kunna delta som informerad medborgare i ett demokratiskt samhälle, i demokratiska processer (Lilja, Eklund & Tottie, 2024; Nordiska ministerrådet, 2025/2026).

Lilja, Eklund och Tottie visar i en studie om medborgerlig kompetens och desinformation att kunskapen om desinformation varierar i befolkningen. Studien visar också att många efterfrågar mer kunskap och fler insatser för att stärka medborgares förmåga att identifiera och bedöma information (Lilja, Eklund & Tottie, 2024).

Insatser för att stärka befolkningens medie- och informationskunnighet utförs i de flesta länder. För att få en uppfattning om kunskapsnivån i befolkningen genomförde de nordiska mediemyndigheterna 2025 en undersökning i ländernas befolkning 16 år och uppåt (Nordiska ministerrådet, 2025/2026). Frågor ställdes bland annat om respondenternas attityder till inlägg och nyheter som publiceras på olika plattformar, till exempel om respondenterna tycker att de ger relevant/viktig information om aktuella händelser, om de är trovärdiga och om de stärker yttrandefriheten och demokratin. Resultaten visar att svenska respondenter instämmer i något lägre grad i påståendena när de gäller sociala medier än när de gäller nyheter i tv eller dagstidningar:

Tabell: Svenska respondenters genomsnittliga grad av instämmande gällande olika mediers funktion, 16+ år

	Sociala medier	Nyheter i tv	Nyheter i dagstidningar
...de ger mig relevant/viktig information om aktuella händelser i Sverige och resten av världen	2,9	3,9	3,7
...de är trovärdiga källor till information	2,6	3,7	3,5
...de ger en rättvis bild av vad som sker i Sverige och resten av världen	2,8	3,5	3,4
...de låter olika människor komma till tals	3,2	3,5	3,4
...de stärker yttrandefriheten och demokratin	3,1	3,7	3,6

Anm: 1 instämmer inte alls – 5 instämmer i högsta grad

En annan fråga i undersökningen gällde kännedom om användares möjlighet att anmäla inlägg som anses bryta mot användarvillkoren (t.ex. desinformation eller olagligt innehåll). Mer än en fjärdedel av de svenska respondenterna kände inte till detta:

Tabell: Andel som känner till möjlighet att anmäla inlägg samt andel som själv har anmält, 16+ år (procent)

Känner till och har anmält	31%
Känner till men har inte anmält	41%
Känner inte till	27%

På sociala medieplattformar finns inte, så som hos redaktionella nyhetsmedier, någon ansvarig utgivare eller redaktion som granskar det som publiceras. Även om plattformarna modererar innehållet och kan ta bort innehåll som är olagligt eller oönskat vilar ansvaret för innehållet som publiceras även på den enskilde användaren. Att skapa ett klimat där olika åsikter får ta plats blir där en gemensam uppgift, något en medie- och informationskunnig individ bör ha viss kännedom om. Nedan följer

resultat från de svenska respondenterna vad gäller kännedom om användarvillkor och vem som har ansvaret för olika typer av innehåll samt deras attityder till vad som är lämpligt agerande online.

Tabell: Svenska respondenters genomsnittliga grad av instämmande gällande användarvillkor och ansvar, 16+ år

Jag anser att man bör fråga om lov innan man taggar familjemedlemmar/vänner eller publicerar bilder eller filmer på andra än mig själv i sociala medier	4,2
Jag själv har ansvar för att det jag lägger upp på mitt sociala medie-konto följer lagar och regler	4,4
Det är viktigt att säga ifrån i kommentarsfältet om man ser att andra utsätts för hets/påhopp i sociala medier	4,4

Anm: 1 instämmer inte alls – 5 instämmer i högsta grad.

8.4 MIK och politiskt deltagande

Opinionsbildning handlar inte enbart om hur människor själva tar till sig och bedömer information och formar uppfattningar utan också om hur människor uttrycker sina åsikter och deltar i samhällsdebatten. Medie- och informationskunnighet har enligt olika studier samband med politiskt engagemang och deltagande samt ökad exponering för olika perspektiv. Det tyder på att MIK kan stärka individers förmåga att delta i demokratiska samtal (Ashley, Maksl & Craft, 2017; Kahne, Lee & Feezell, 2012).

I en studie av Holt, Shehata, Strömbäck och Ljungberg framgår att både politisk användning av sociala medier och uppmärksamhet på politiska nyheter i traditionella medier kan öka politiskt engagemang över tid. Studien bygger på svenska paneldata från riksdagsvalet 2010 och visar att sociala medier kan fungera som en utjämnande faktor, särskilt vad gäller yngre medborgares politiska deltagande (Holt, Shehata, Strömbäck & Ljungberg, 2013).

Sambandet mellan digitala medier och politiskt deltagande har även studerats i en senare undersökning av forskarna Dimitrova, Shehata, Strömbäck och Nord. I deras panelstudie framgår att digital medieanvändning endast hade svaga effekter på politisk kunskap, men tydligare effekter på politiskt deltagande (Dimitrova, Shehata, Strömbäck & Nord, 2014).

8.5 Algoritmisk medvetenhet

Medie- och informationskunnigheten (MIK) omfattar även digital kompetens och en viss förståelse för hur innehåll och information produceras och sprids på nätet. Förståelse för hur algoritmer fungerar har betydelse för att exempelvis kunna ta ställning till vad för personlig information man delar i de spår vi lämnar när vi är online samt hur det kan påverka det innehåll som visas i flödet, det användaren får rekommenderat i exempelvis sociala medier, sökmotorer och strömningstjänster. Här ingår även en förståelse för att algoritmer inte är neutrala (Haider & Sundin, 2017). Medvetenheten om algoritmer är ofta ojämn, delvis felaktig (baserad på anekdoter och folkteorier snarare än teknisk kunskap) och beroende av erfarenhet. Forskningsfältet växer snabbt och begreppen används fortfarande på lite olika sätt. *Algorithmic literacy*, algoritmisk kunnighet avser en bredare förståelse för hur de fungerar och hur man kan förhålla sig kritiskt till dem. Människor har även egna vardagsteorier om hur algoritmer fungerar, *folk theories of algorithms* (Oeldorf-Hirsch & Neubaum, 2023).

Eslami med kollegor fann i en studie med 40 Facebook-användare 2015 att för sex av tio (62,5 procent) var algoritmisk filtrering av Facebooks nyhetsflöde okänt. När deltagarna fick se hur algoritmen påverkade flödet reagerade många med överraskning och irritation. Studien visar också att ökad medvetenhet om algoritmer kunde leda till mer aktivt användarbeteende och större känsla av kontroll (Eslami et al., 2015).

Användarnas egna vardagsteorier om hur algoritmer fungerar skapas när algoritmer är svårbegripliga, när användare försöker förstå algoritmiska flöden genom erfarenhet i vardagssituationer. En algoritmisk medvetenhet växer då fram praktiskt, genom användning och observation, snarare än genom formell utbildning. Användarens egna teorier om vad som påverkar algoritmen, vilka kan vara delvis korrekta, felaktiga eller mycket förenklade, får effekter på användarens agerande, ibland strategiskt för att styra flödet. Även felaktiga teorier kan således få verkliga konsekvenser för hur människor använder digitala plattformar (Eslami et al., 2016).

I en norsk studie identifierades fem vanliga vardagsteorier: algoritmer ses som begränsande, praktiska, reducerande, svårgripbara och exploaterande. En central slutsats är att människor inte är passiva inför algoritmers påverkan. Forskarna beskriver en digital irritation, att användare känner frustration över algoritmer, även när de inte fullt ut förstår hur algoritmerna fungerar. Algoritmisk medvetenhet innehåller på så sätt känslor, vardagliga tolkningar samt politiska eller samhälleliga funderingar om datainsamling, övervakning och plattformars makt (Ytre-Arne & Moe, 2021).

Oeldorf-Hirsch och Neubaum (2023) genomförde en studie bland sociala medieanvändare i USA och Tyskland. De fann att yngre användare, personer med högre utbildning och personer som använder sociala medier oftare än genomsnittet tenderar att ha högre algoritmisk medvetenhet. I studien framkom att de amerikanska

användarna hade högre algoritmisk medvetenhet och mer positiva attityder till algoritmer än tyska användare, troligen eftersom de använde sociala medier mer.

Ålder, utbildning, internetfärdigheter och användningsvanor tycks spela roll. Det gör att algoritmisk kunnskap också blir en fråga om digital ojämlikhet, inte bara individuell kunnskap.

Hur människor lär och får medvetenhet om algoritmer har även undersökts i en litteraturstudie omfattande 19 vetenskapliga artiklar. Där framkom att transparens är viktig för synen på algoritmer och hur de kan hjälpa användaren att påverka resultatet i önskvärd riktning. Vidare framgick att användare främst lär genom erfarenhet vilket ger stöd till tidigare nämnda forskningsresultat samt att utbildning och ett reflekterande, aktivt deltagande också har betydelse. Infrastruktur, kontext och sociala sammanhang visade sig även bidra till den ökade medvetenheten. I flera av undersökningarna i litteraturstudien framkom att de användare som hade en algoritmisk medvetenhet bättre kan bedöma när algoritmerna kan vara till nytta eller inte. Genom denna medvetenhet kan personliga inställningar justeras för att exempelvis kunna få ut mesta möjliga av systemet och samtidigt skydda personlig information (Norrby, 2022). Internationell forskning har även visat att vissa grupper i samhället kan ha svårare att ta till sig och förstå hur algoritmer och innehållsmoderering fungerar. Plattformarnas användningsvillkor är generellt svårlästa och oöverskådliga vilket leder till att vissa grupper snarare förlitar sig på anekdotiska bevis och folkteorier om hur innehållsmoderering fungerar än tar reda på vilka villkor som gäller. Detta skapar i sin tur misstro mot de sociala medierna, både gällande hur plattformarna faktiskt modererar innehåll man lägger upp och gällande deras vilja och förmåga att informera om detta (Mayworm et al., 2024). Med ojämn kunnskap om algoritmer i befolkningen riskerar digitala klyftor att öka och fördjupas när algoritmer blir allt vanligare i det digitala samhället.

För att undersöka befolkningens kunskaper om vad som påverkar en sökning online innehöll den nordiska mediekunnskapsundersökningen¹⁴ ett enkelt test där respondenten fick besvara frågan om vad som avgjorde sökträffarnas ordning. En skärmdump med en Googlesökning på dunjacka visades. Syftet med frågan var att undersöka respondentens medvetenhet om algoritmer.

¹⁴ <https://www.norden.org/en/publication/nordic-media-literacy-survey>.

Tabell: Vad beror ordningen i sökträffarna på? 16+ år (procent)

	16–24 år	25–34 år	35–44 år	45–54 år	55–64 år	65+ år
Den bästa jackan är överst	21%	18%	8%	6%	4%	3%
Det är slumpen som avgör ordningen	10%	11%	7%	3%	5%	4%
Någon har betalat för att deras jacka ska komma överst	44%	48%	65%	65%	60%	54%
Den träff som flest personer har klickat på är överst	21%	17%	14%	16%	16%	13%
Vet ej	4%	5%	6%	9%	15%	26%

Trots att den översta träffen i sökresultatet är märkt med texten ”sponsrad” anser knappt hälften av de yngre vuxna (under 35 år) att detta är orsak till placeringen. Ungefär en femtedel i denna åldersgrupp anser att ”den bästa jackan” listats överst. I den äldsta åldersgruppen, 65 år och äldre, är osäkerheten kring vad som avgör ordningen som störst, en fjärdedel svarar ”vet ej” på frågan. Resultatet kan ses som ett exempel på variationer i förståelse för bakomliggande påverkan av informationsflödet och att olika kunskaper behöver förstärkas i olika åldersgrupper.

8.6 Begränsningar och komplexitet i sambandet mellan MIK och opinionsbildning

När det gäller möjligheter och förutsättningar för opinionsbildning är de olika kunskaper och förmågor som omfattas av medie- och informationskunnighet centrala men sambandet är komplext. Opinionsbildning formas av strukturer som algoritmer, kommersiella plattformar och individualiserade flöden, inte enbart av individens kompetens (Haider & Sundin, 2019). Samma källkritiska verktyg kan användas både för att stärka vetenskapligt framtagen kunskap och för att ifrågasätta etablerad

kunskap (t.ex. i konspirationsteorier) som skapar tvivel och misstro. Det handlar för den enskilde därför om att veta var trovärdig information går att hitta och att ha tillit till dessa källor och förtroende för institutioner (t.ex. vetenskapligt framtagen forskning) som en utgångspunkt för att kunna bilda sig åsikter (Sundin & Haider, 2018). Medie- och informationskunnighet påverkar snarare kunskap, kritiskt tänkande och förmågan att värdera information mer direkt än vad den påverkar specifika politiska åsikter (Ashley, Maksl & Craft, 2017; Droog et al., 2025).

Det är därför viktigt att inte se medie- och informationskunnighet som en enkel lösning på problem kopplade till desinformation och opinionsbildning. Däremot visar forskning att MIK är en central förutsättning för att individer ska kunna orientera sig i ett komplext informationslandskap, öka förutsättningarna för politiskt deltagande och samhällsengagemang samt stärka demokratisk motståndskraft (Anstead et al., 2025; Broda & Strömbäck, 2024; Kahne, Lee & Feezell, 2012; Holt et al., 2013; Lilja, Eklund & Tottie, 2024).

9 Utvärdering och forskning om DSA

Under internets första decennier kännetecknades både EU:s och USA:s reglering av en återhållsam syn på ansvar för innehåll på nätet. Internet behandlades inte i första hand som ett traditionellt medieområde, utan som en infrastruktur eller förmedlingstjänst av information. Denna grundsyn innebar att onlineplattformar i stor utsträckning var fredade från ansvar för innehåll som användare publicerade eller laddade upp (Lundqvist, 2024).

Den grundläggande ansvarsfriheten för förmedlingstjänster består även efter EU:s förordning om digitala tjänster (*Digital Services Act*, DSA) som antogs den 19 oktober 2022 och trädde i kraft den 16 november 2022. Den tillämpas i sin helhet sedan den 17 februari 2024. Reglerna i DSA syftar bland annat till att skydda användare från vilseledande information och minska olagligt innehåll online (se vidare avsnitt 10, bilaga).

Om en användare misstänker att en leverantör av förmedlingstjänster har överträtt DSA kan denne lämna in klagomål till den nationella samordnaren för digitala tjänster, vilket i Sverige är Post- och Telestyrelsen (PTS). PTS noterar i sin *Verksamhetsrapport 2025* att antalet anmälningar ökar stadigt, vilket kan vara ett uttryck för att allt fler känner till DSA och rätten att lämna in klagomål. Skälen till klagomålen var främst brist på motivering av beslut gällande begränsning eller nedstängning av konton på sociala medier (PTS, 2026).

DSA har utvärderats och diskuterats inom forskningen, civilsamhället och myndigheter sedan några år tillbaka. Eftersom förordningen trädde i kraft för endast drygt två år sedan finns det inte så mycket empirisk forskning om effekterna av DSA utan mer teoretisk utifrån bland annat juridiska, etiska och medievetenskapliga perspektiv. Forskningen studerar olika delar av DSA, bland annat kraven på transparens och huruvida plattformarna lever upp till dem, kraven på moderering och hur det samspelar med grundläggande rättigheter såsom yttrandefrihet och demokratiska värden. I detta avsnitt har vi avgränsat materialet till vetenskapliga artiklar som:

- Utvärderar och diskuterar transparenskravet med särskilt fokus på *DSA Transparency Database*
- Diskuterar utformningen av DSA samt dess styrkor och svagheter i förhållande till användarnas grundläggande rättigheter och opinionsbildning.

9.1 Transparenskravens genomslag

I september 2023 lanserade Europeiska kommissionen *DSA Transparency Database* som en offentlig databas för innehållsmoderering inom EU. Databasen samlar så kallade *statements of reasons* det vill säga standardiserad information som plattformar enligt DSA ska lämna för varje beslut att ta bort innehåll eller på annat sätt begränsa dess spridning. Uppgifterna omfattar bland annat vilken åtgärd som vidtagits, skälen för beslutet samt dess rättsliga eller avtalsmässiga grund.

Syftet med databasen är att möjliggöra ökad insyn i plattformarnas modereringsbeslut och att skapa bättre förutsättningar för granskning av hur innehåll hanteras online. Databasen gör det möjligt att följa modereringsbeslut i nära realtid och att analysera dessa genom sök- och nedladdningsfunktioner.

Detta avsnitt beskriver i vilken utsträckning transparenskraven och databasen har bidragit till ökad insyn i plattformarnas innehållsmoderering, samt vilka begränsningar som kvarstår.

9.1.1 Ökad insyn – men begränsad tillförlitlighet

DSA har inneburit ett tydligt steg framåt när det gäller insyn i plattformarnas innehållsmoderering. Genom *DSA Transparency Database* och krav på transparensrapporter finns det nu mer systematisk och offentlig information om modereringsåtgärder än tidigare.

Flera studier visar dock att dessa verktyg ännu inte ger en fullt tillförlitlig, jämförbar eller heltäckande bild av plattformarnas faktiska moderering. Samtidigt begränsas genomslaget av egenrapportering, varierande datakvalitet och otydlig rapporteringspraxis (Groesch et al., 2026; Trujillo et al., 2025; Shahi et al., 2025).

9.1.2 Fortsatt stora skillnader mellan plattformar

Analyserna visar att det finns tydliga skillnader mellan plattformarna både vad gäller omfattning och inriktning på modereringen. Skillnaderna gäller bland annat hur mycket innehåll som modereras, vilka typer av innehåll som prioriteras och vilka åtgärder som vidtas. Detta tyder på att modereringspraktikerna fortsatt är plattformsspecifika, trots att de rapporteras inom ett gemensamt EU-ramverk (Drolsbach & Pröllochs, 2024; Trujillo et al., 2025; Shahi et al., 2025).

Plattformarnas modereringsgrunder anges ofta i breda kategorier. Det försvårar en mer exakt bedömning av vilka överträdelser som ligger bakom åtgärderna och försvårar jämförelser mellan plattformar. Vidare framträder missinformation och desinformation relativt svagt i de studerade datamängderna. Detta kan spegla både faktiskt fokus i modereringen och hur plattformarna klassificerar sina åtgärder (Drolsbach & Pröllochs, 2024; Trujillo et al., 2025).

9.1.3 **Automatisering är en viktig del i modereringen, men uppgifterna är svåra att tolka**

Som tidigare beskrivits (i avsnitt 6) använder plattformarna olika typer av modereringsåtgärder för att ta bort, begränsa, upp- eller nedprioritera innehåll och användarkonton. Studierna visar att automatisering spelar en central roll i innehållsmodereringen. Samtidigt skiljer sig plattformarna åt i hur användningen av automatiserade system rapporteras. Forskningen pekar på att dessa uppgifter i flera fall är svåra att förena med plattformarnas övriga rapportering, vilket innebär att uppgifter om automatisering bör tolkas med försiktighet (Drolsbach & Pröllochs, 2024; Trujillo et al., 2025; Groesch et al., 2026; Leerssen, 2023).

9.1.4 **Databrister försvårar jämförelser och analys**

Flera studier visar att databasen präglas av brister i datakvalitet och rapportering. Viktiga uppgifter saknas ofta och kategorier används på olika sätt. Vidare finns exempel på duplikat, inkonsekventa tidsstämplar och poster där modereringsåtgärder förefaller ha registrerats innan innehållet publicerades. Detta betyder att databasen, trots ett betydande informationsvärde, i flera avseenden ger ett otillräckligt underlag för jämförelser och mer långtgående slutsatser om plattformarnas moderering (Trujillo et al., 2025; Shahi et al., 2025; Groesch et al., 2026).

Databasen har också mer grundläggande begränsningar (Groesch et al., 2026):

- den innehåller inte det modererade innehållet i sig, vilket innebär att det inte är möjligt att bedöma om enskilda modereringsbeslut är korrekta, proportionerliga eller rättsenliga
- den omfattar endast fall där modereringsåtgärder har vidtagits, vilket innebär att det saknas information om innehåll som granskats men inte åtgärdats
- den innehåller i begränsad utsträckning uppgifter om innehållets räckvidd och spridning, vilket försvårar bedömningen av modereringens faktiska effekter.

Forskningen visar också att det finns avvikelser mellan uppgifter i DSA-databasen och i plattformarnas egna transparensrapporter, bland annat vad gäller hur modereringsåtgärder och graden av automatisering redovisas. Detta medför att egenrapportering inte är tillräcklig för att säkerställa kvalitet och ansvarsskyldighet, och att kompletterande oberoende granskning är nödvändig (Trujillo et al., 2025; Shahi et al., 2025).

Plattformarna uppfyller ofta de formella kraven på rapportering, men utnyttjar i begränsad utsträckning möjligheten att tillhandahålla mer detaljerad och jämförbar information (Trujillo et al., 2025; Shahi et al., 2025). Flera studier betonar således att det finns osäkerheter i datakvalitet, rapporteringspraxis och metod vilket

sammantaget innebär att resultaten från databasen bör tolkas med stor försiktighet och att man bör vara varsam med att dra mer långtgående slutsatser (Groesch et al., 2026; Trujillo et al., 2025; Drolsbach & Pröllochs, 2024).

9.1.5 Begränsade förändringar i moderering under valperioden

Den mest omfattande studien av moderering som skedde i anslutning till Europaparlamentsvalet 2024 visar inte några tydliga tecken på brett förändrade modereringsmönster under perioden.

Undantaget är LinkedIn, där en ökning av moderering av valrelaterad missinformation har identifierats. I övrigt är slutsatsen att det antingen inte skedde någon tydlig anpassning eller att databasens utformning gör sådana förändringar svåra att observera (Shahi et al., 2025).

9.1.6 Begränsad insyn i rekommendationssystemens påverkan

Transparens om plattformarnas moderering handlar inte enbart om borttagning av innehåll och avstängningar av konton. DSA syftar också till att skapa insyn i åtgärder som påverkar innehållets synlighet och räckvidd, till exempel nedprioritering och andra former av begränsad spridning.

Studierna visar att sådana åtgärder är svåra att definiera, observera och utvärdera. Detta hänger samman med plattformarnas rekommendations- och rankingssystem. Det innebär att transparenskraven är viktiga, men att de ännu endast delvis fångar plattformarnas strukturella påverkan på vilka innehåll som får genomslag (Leerssen, 2023; Söderlund et al., 2024).

I detta sammanhang skiljer forskningen mellan horisontell och vertikal transparens. *Horisontell transparens* avser den insyn som användare och allmänhet får i plattformarnas regler och modereringsåtgärder. *Vertikal transparens* avser den insyn som möjliggörs genom tillsyn, revision och forskartillgång.

Den vertikala transparensen framhålls som särskilt viktig när det gäller rekommendationssystem. Dessa system påverkar i hög grad vilket innehåll som exponeras och kan ge upphov till risker som utvecklas över tid, exempelvis för informationsflöden, autonomi och spridning av skadligt innehåll (Söderlund et al., 2024).

Det betonas även ett behov av så kallad *socioteknisk transparens*. Detta avser insyn inte bara i de tekniska systemen, utan också i hur dessa samverkar med användare, innehåll och samhälleliga förhållanden.

Sammanfattningsvis innebär detta att meningsfull granskning av plattformarnas påverkan kräver mer än information om enskilda modereringsbeslut. Det kräver dessutom strukturerad tillsyn och oberoende analys av hur systemen fungerar i praktiken (Gausen et al., 2025; Söderlund et al., 2024).

9.2 DSA:s betydelse för grundläggande rättigheter och opinionsbildning

I detta avsnitt beskrivs forskning om DSA:s skrivningar och hur regelverket generellt påverkar användarens ställning i förhållande till plattformarna och för opinionsbildning i samhället. Endast vetenskapliga artiklar publicerade mellan 2023 och 2026 har inkluderats.

9.2.1 En procedurell regleringsmodell som kan stärka grundläggande rättigheter

En genomgång av forskningen visar först och främst att DSA betraktas som en procedurell (eller processuell) regleringsmodell på så sätt att lagen framför allt ställer krav på processer: transparens, motiveringar, klagomål, tillsyn, riskbedömningar och ansvarsfördelning. Flera forskare betonar att denna procedurella ansats är viktig eftersom den inte direkt fastställer vad människor ska få uttrycka, men ändå försöker begränsa skadliga inslag i den digitala offentligheten genom att reglera hur plattformarna ska hantera risker (Carretta, 2025; Husovec, 2024; Eder, 2024; Palumbo, 2024; Caruso, 2025; Reviglio, Bleyer-Simon & Verza, 2024).

Forskningen ger för övrigt en dubbel bild. Å ena sidan kan DSA stärka enskildas opinionsbildning genom att göra plattformarnas moderering mer transparent, möjlig att ifrågasätta och binda den tydligare till grundläggande rättigheter, särskilt yttrandefrihet. Bland annat artikel 14, som innebär att plattformar vid tillämpning av sina användarvillkor ska ta vederbörlig hänsyn till användarnas grundläggande rättigheter, innebär att regler som tidigare varit rent privata – plattformarnas egna användarvillkor – nu i större utsträckning kopplas till offentliga rättighetsprinciper (Quintais et al., 2023; Wolters & Gellert, 2023). Flera organisationer som arbetar med mänskliga rättigheter har under många år argumenterat för att plattformarna borde inkorporera principer om mänskliga rättigheter i sina användarvillkor men det är först i och med DSA som de åläggs att ta hänsyn till dessa i sin verksamhet. För en enskild användare kan detta, enligt Quintais et al., innebära att plattformar inte längre kan moderera innehåll enbart utifrån kommersiella eller interna policyintressen utan måste väga in rättighetsaspekter. Både EU:s och FN:s konvention om mänskliga rättigheter reglerar den vertikala relationen mellan staten och individen, inte den horisontella

relationen mellan privata aktörer, såsom företag och individer. Forskarnas bedömning är att artikel 14 indirekt kopplar den horisontella relationen till grundläggande rättigheter, det vill säga att rättighetsskyddet får betydelse även i relationen mellan användare och privata plattformar. Därmed stärks individens position i förhållande till plattformarna.

Detta innebär att DSA kan stärka enskildas opinionsbildning genom att användare får bättre möjligheter att förstå och ifrågasätta plattformars beslut. Om en persons inlägg tas bort, nedprioriteras eller kontot begränsas kan DSA:s regler om motivering och klagomål bidra till större rättssäkerhet. Samtidigt påpekar flera forskare att formella möjligheter att klaga inte är tillräckligt om besluten exempelvis sker med automatiserade system och svårbegripliga regler. De formella möjligheterna måste omsättas i praktiken (Carretta, 2025; Eder, 2024; Golunova, 2025; Quintais et al., 2023).

Det finns även forskare som hävdar att skyddet för grundläggande rättigheter, såsom det nu är formulerat i DSA, inte är tillräckligt. Exempelvis Carretta argumenterar för ”a fundamental rights approach to content moderation” genom att skärpa artikel 14 från att kräva att plattformarna tar *vederbörlig hänsyn* till grundläggande rättigheter (”due regard for fundamental rights”) till krav på *tillbörlig aktsambet* (”due diligence”) (Carretta, 2025 s. 386f). Den nuvarande skrivningen om *vederbörlig hänsyn* innebär att plattformarna måste respektera användarnas grundläggande fri- och rättigheter enligt EU:s stadga om de grundläggande rättigheterna. Det vill säga att de måste väga in användarens rättigheter i sitt eget agerande, vid t.ex. modereringsbeslut. Det är en tämligen passiv roll medan *tillbörlig aktsambet* är en mer aktiv och systematisk process där man utreder och minimerar risker innan ett beslut fattas, det vill säga en aktiv form av riskhantering och riskminimering. Detta skulle innebära att plattformarna måste utgå från användarnas grundläggande rättigheter när de utformar sina tjänster, modereringspraktiker och användarvillkor (T&C). Det skulle så att säga bygga in användarnas rättigheter i plattformarnas själva struktur och funktionssätt.

9.2.2 Risk för övermoderering

Å andra sidan pekar flera forskare på risker, i synnerhet att DSA kan leda till ökad borttagning eller nedprioritering av innehåll, särskilt om plattformar överreagerar på krav att hantera olagligt eller samhällsskadligt innehåll, eller om offentliga aktörer indirekt påverkar privata modereringsregler (Quintais et al., 2023; Galantino, 2023; Palumbo, 2024; Golunova, 2025). Bland annat Galantino (2023) argumenterar för att DSA:s koppling till desinformationsreglering riskerar att skapa otydliga ansvarsförhållanden, särskilt när offentliga aktörer uppmuntrar privata plattformar att agera mot lagligt men oönskat innehåll (Galantino, 2023). Desinformation är ett omstritt och svårdefinierat begrepp och DSA definierar inte desinformation i sina bindande artiklar (Ó Fathaigh, Bujs & van Hoboken, 2025). Detta innebär att tillämpningen kan medföra risker för borttagning utan tillräcklig grund. Samtidigt är DSA:s ”röda linje”, enligt Husovec, att kommissionen inte blir en

”sanningsmyndighet”: riskhanteringen ska avse systemiska risker och processer, inte ge myndigheter frihet att avgöra vilka åsikter som är tillåtna (Husovec, 2024).

För enskildas opinionsbildning innebär detta att DSA kan bidra till att minska manipulation, falska konton, koordinerade påverkanskampanjer och annan systematisk informationspåverkan. Samtidigt betonar forskningen vikten av tydliga gränser så att insatser mot desinformation inte används för att tysta minoritetsuppfattningar eller politiskt obekväma åsikter (Ó Fathaigh, Buijs & van Hoboken, 2025; Galantino, 2023; Husovec, 2024).

9.2.3 Systemrisker, algoritmer och den praktiska yttrandefriheten

Som tidigare nämnts påverkas opinionsbildning inte bara av om ett inlägg får finnas kvar eller tas bort utan också av synlighet: vilka inlägg som hamnar högt i flödet, vilka som får spridning och vilka som i praktiken blir osynliga. Reviglio, Bleyer-Simon och Verza beskriver detta som plattformars ”opinion power”, det vill säga deras förmåga att påverka individuell och offentlig åsiktsbildning genom design, användarvillkor, innehållsmoderering och rekommendationssystem. Detta flyttar fokus från klassisk yttrandefrihet – rätten att säga något – till frågan om vilka röster som faktiskt blir hörda i digitala miljöer (Reviglio, Bleyer-Simon & Verza, 2024; Eder, 2024; Riemer & Peter, 2021). Teoretiskt skulle man kunna säga att DSA ställer krav på att även dold nedprioritering av innehåll ska redovisas men eftersom det i praktiken är tekniskt komplicerat och svårt att identifiera kommer det sannolikt inte stoppa *shadowbanning*. Motsatsen, förstärkning av innehåll eller *shadow-promotion* förblir också svårupptäckt (Reviglio, Bleyer-Simon & Verza, 2024, s 127).

Ytterligare faktorer i sammanhanget är informationsasymmetrin och ojämlikheten i fråga om teknisk expertis som är svår att reglera bort:

These asymmetries, along with the platforms’ business models that prioritize user engagement – a strategy that seem to contribute to most of the unintended and harmful consequences of social media – remain largely unaddressed by current regulations (Reviglio, Bleyer-Simon & Verza, 2024, s 133).

Även Golunova (2025) argumenterar för att DSA är otillräckligt när algoritmer används för moderering i massiv skala. DSA kräver att plattformarna ska formulera sina regler tydligt och begripligt och att dessa ska tillämpas konsekvent och icke-diskriminerande. Men när det är AI-baserade algoritmer som tillämpar reglerna blir det varken tydligt och begripligt eller konsekvent och icke-diskriminerande. Detta beror på de problem som beskrevs i avsnitt 6, såsom att algoritmiska system bygger på generella och förbestämda kriterier som i praktiken inte kan ta hänsyn till och väga in alla rättigheter på individuell nivå. Automatiserade system kan inte heller känna igen eller bedöma den faktiska betydelsen av ett innehåll och har svårigheter med att anpassa sig till sammanhang som snabbt ändras, vilket riskerar leda till att systemet

förstärker exkludering i stället för att skydda utsatta grupper och yttrandefriheten. Eftersom DSA inte fokuserar på tekniken, hur regler implementeras av AI, är DSA otillräckligt enligt Golunova (2025). DSA fångar inte heller in skaleffekten av automatiserad moderering. Automatiserade system påverkar miljontals inlägg varför små brister kan få stora systematiska konsekvenser. DSA kräver att mycket stora plattformar årligen gör systemriskbedömningar (se avsnitt 10) men Golunova (2025) menar att dessa inte är tillräckliga då de inte fångar in själva modereringssystemens design. DSA angriper så att säga konsekvenserna men inte tillräckligt de tekniska orsakerna (AI-baserade algoritmer). DSA är ett viktigt steg för att utöka plattformarnas ansvar och transparens, men kräver ytterligare konkretisering för att skydda yttrandefriheten, enligt Golunova (2025).

Golunova framför även kritik av artikel 17 i DSA, som bland annat fastställer att aktörer måste motivera sina beslut att begränsa innehåll. Hon framför att om man tittar på den bokstavliga lydelsen av artikel 17, så kan den tolkas som att den endast kräver att leverantören ska informera om automatiserade medel har använts, men inte exakt *hur* dessa medel påverkat de faktiska besluten. Det kan vara svårt att utifrån sådana motiveringar effektivt invända mot algoritmiska beslut. Även om det senare finns möjligheter att överklaga besluten, både internt och externt till kontrollorgan, så kan det vara svårt för enskilda att förstå varför beslut togs och hur de går att undvika i framtiden.

Eder (2024) framhåller, precis som Golunova (2025), att DSA:s individuella rättsmedel inte fullt ut fångar den strukturella dimensionen. En användare kan klaga på att ett inlägg tagits bort, men det är betydligt svårare att förstå eller ifrågasätta hur rekommendationssystem, nedprioritering eller algoritmisk förstärkning påverkar informationsmiljön som helhet. Eder (2024) framhåller att systemriskbedömningarna är ett viktigt verktyg här då de, trots vissa brister, kastar ljus över plattformarnas samhälleliga och strukturella effekter på det offentliga samtalet. Kravet på systemriskbedömningar beskrivs av Eder (2024) som ett paradigmskifte, från en modell där enskilda användare själva måste klaga på enskilda beslut, till en modell där plattformar måste bedöma hur deras tjänster påverkar samhälleliga värden som offentlig debatt, val, hälsa och säkerhet. Ansvaret förflyttas därmed från individen till plattformarna (Eder, 2024).

Detta är viktigt eftersom skador på opinionsbildning ofta inte drabbar en individ isolerat utan publiken och offentligheten som helhet. Om en plattform förstärker polariserande, falskt eller manipulerat innehåll kan det påverka hela det offentliga samtalet utan att någon enskild användare enkelt kan visa att just hens rättighet har kränkts. Systemrisker påverkar alla, inte bara en enskild användare. Därför menar Eder att DSA:s systemriskmodell kan ge ett mer ändamålsenligt ansvar för plattformarnas strukturella makt (Eder, 2024).

För enskilda innebär detta att DSA:s betydelse inte bara ligger i skydd mot borttagning, utan också i möjligheten att få insyn i och skydd mot osynlig styrning av

informationsflöden. Men den praktiska opinionsfriheten beror på om användare kan förstå och påverka de system som avgör vilken information de möter (Reviglio et al., 2024; Eder, 2024; Golunova, 2025).

9.2.4 Tillsynens betydelse

Inom DSA:s tillsynssystem har nationella *Digital Services Coordinators*, DSC, centrala roller. DSC är nationella administrativa myndigheter som ansvarar för att genomföra DSA i varje medlemsland. Dessa myndigheter hanterar klagomål och formar normer kring plattformars moderering, exempelvis genom betrodda anmälare, uppförandekoder och tillsynsåtgärder. För stora plattformar (VLOP) har EU-kommissionen exklusiv kompetens, men de nationella myndigheterna kan fortfarande utöva en betydande roll i enskilda fall. Avsikten med en decentraliserad tillsynsstruktur är att skydda medborgarnas rättigheter, men det kan leda till att tillämpningen av DSA varierar mellan medlemsstater (van de Kerkhof, 2025).

Om olika nationella tillsynsmyndigheter har olika syn på yttrandefrihet, olagligt innehåll eller samhällsrisker kan användare i praktiken få olika starkt skydd beroende på medlemsstat. Detta kan, enligt van de Kerkhof (2025) innebära en risk för ojämnhet i skyddet av yttrandefriheten inom EU (van de Kerkhof, 2025).

För enskildas opinionsbildning innebär detta att DSA:s effekter inte bara avgörs av plattformarnas policyer, utan också av myndigheternas kompetens, oberoende, resurser, prioriteringar och tolkningar. En välfungerande tillsyn kan stärka användarnas skydd mot godtycklig moderering, medan en svag eller otydlig tillsyn kan göra skyddet oförutsägbart. van de Kerkhof föreslår därför en ökad centralisering och tydligare riktlinjer för de nationella tillsynsmyndigheterna (van de Kerkhof, 2025; se även Eder, 2024).

Samtidigt bygger DSA in fler aktörer i andra delar av tillsynen, vilket kan balansera de statliga och överstatliga organens inflytande. I framtagandet av *guidelines* och *best practices* samt i systemriskutvärderingar ska oberoende experter och civilsamhällesorganisationer involveras vilket innebär att de kan komma att påverka hur risker definieras, identifieras och eventuellt även hur riskerna ska hanteras. Här kan potentiellt skapas ”goda cirklar” som både stärker civilsamhällets inflytande och hanteringen av systemrisker (Eder, 2024).

Både Eder (2024) och Carretta (2025) menar att kommissionen borde främja ett system där flera aktörer, såsom civilsamhället, är involverade i systemriskbedömningar och andra processer för att tillsammans successivt förfina och förankra regleringen bland berörda parter. Carrettas analys av regleringen ”of” och ”by platforms” visar att det finns behov av ett mer konsistent och effektivt system för ”checks and balances” som är applicerbart på olika slags plattformar, men givet problemområdets komplexitet måste utformningen av ett sådant system involvera en mångfald av aktörer och perspektiv (Carretta, 2025 s. 374).

9.2.5 Samlad analys: vad betyder DSA för enskildas opinionsbildning?

Sammantaget visar forskningen att DSA kan stärka enskildas opinionsbildning på tre huvudsakliga sätt. För det första kan DSA stärka användares ställning gentemot plattformar genom krav på tydligare användarvillkor, motiveringar, klagomål och hänsyn till grundläggande rättigheter. För det andra kan DSA göra plattformarnas strukturella makt mer synlig genom krav på transparens, riskbedömningar och genom tillsyn. För det tredje kan DSA bidra till att motverka systematisk manipulation av informationsmiljön, exempelvis koordinerade desinformationskampanjer (Quintais et al., 2023; Eder, 2024; Ó Fathaigh et al., 2025). Man kan beskriva det som att DSA försöker flytta maktbalansen något från plattformarna till användare och tillsynsmyndigheter genom att ställa krav på processer som bland annat möjliggör offentlig granskning. Därmed kan DSA stärka förutsättningarna för yttrandefrihet, informationsfrihet och pluralism i digitala miljöer (Quintais et al., 2023; Caruso, 2025).

Samtidigt visar forskningen att DSA kan innebära risker för enskildas opinionsbildning. En risk är att plattformar, av rädsla för sanktioner eller offentlig kritik, begränsar mer innehåll än nödvändigt. En annan risk är att gränsen mellan offentlig reglering och privat moderering blir oklar för användare och att tillämpningen kan variera mellan olika medlemsländer. En tredje risk är att algoritmisk nedprioritering och andra synlighetsbeslut förblir svåra att upptäcka och ifrågasätta (Palumbo, 2024; Galantino, 2023; Golunova, 2025).

Åtskilliga forskare betonar att DSA är ett viktigt steg i riktning mot att stärka användarnas position och rättigheter i det digitala medielandskapet. Exempelvis Carretta skriver i slutdiskussionen till sin avhandling att:

The EU has taken the lead with the strictest rules and strongest protection of fundamental rights online (as it did already in 2018 with the GDPR for data protection). The DSA forces global online platforms to assess any actual or foreseeable negative effects on the exercise of fundamental rights, thus prioritising users' rights over their right to conduct a business and generate profits. In other words, the new European obligations for platform governance, content moderation, design and deployment of AI systems (especially for AOCM) place the protections of users' fundamental rights at the forefront. This explains why Europe has been and still is, the predominant regulator of global commerce and why it can still successfully export certain norms and lead regulatory efforts. (Carretta, 2025 s. 393)

I genomgången av forskningslitteraturen till detta uppdrag har vi inte funnit någon som ifrågasätter själva ambitionen med DSA, däremot ifrågasätts, som nämnts, om regleringen är fullt tillräcklig och hur den ska tillämpas i praktiken. Genomslaget för transparenskravet begränsas av brister i datakvalitet, jämförbarhet och verifierbarhet samt mer grundläggande begränsningar i hur transparensdatabasen är utformad och vilken typ av information som samlas in. Egenrapportering, otydliga kategorier,

varierande rapporteringspraxis och begränsad tillgång till kompletterande data innebär att transparensen ännu inte är tillräcklig för säker uppföljning eller robust analys av plattformarnas faktiska agerande.

Vilken betydelse DSA kommer att ha för enskildas opinionsbildning kommer bero på hur plattformarna tolkar sina skyldigheter, hur tillsynsmyndigheter agerar, hur effektiva användarnas klagomöjligheter blir, hur forskartillgång och krav på mer enhetlig och precis rapportering utvecklas och hur riskbedömningar kring exempelvis desinformation genomförs i praktiken (Groesch et al., 2026; Söderlund et al., 2024; Trujillo et al., 2025; Shahi et al., 2025; Leerssen, 2023; Gausen et al., 2025; Drolsbach & Pröllochs, 2024; Eder, 2024; van de Kerkhof, 2025; Ó Fathaigh et al., 2025). Flera forskare pekar på plattformarnas affärsmodeller som det grundläggande problemet för det demokratiska samtalet och enskildas opinionsbildning i det digitala medielandskapet. Eftersom DSA inte påverkar affärsmodellerna kommer regleringen inte åt grundproblemet, enligt detta perspektiv, medan andra menar att skärpningar av skrivningarna i DSA ändå skulle kunna stärka användarnas position ytterligare.

Då DSA trädde i kraft för endast drygt två år sedan är det för tidigt för att dra några slutsatser om regleringens effekter på den digitala mediemiljön och enskildas opinionsbildning. DSA garanterar inte i sig en friare eller mer pluralistisk opinionsbildning, men den samlade forskningen pekar på att DSA skapar verktyg som kan bidra till det. Om transparens, klagomål, tillsyn och riskbedömningar fungerar väl kan DSA stärka individens möjlighet att uttrycka sig, ta del av information och förstå plattformars beslut. Om reglerna däremot tillämpas otydligt eller defensivt kan de leda till övermoderering och fortsatt informell, dold styrning (Husovec, 2024; van de Kerkhof, 2025; Eder, 2024). DSA bör därför förstås som en rättslig infrastruktur vars demokratiska värde beror på om den tillämpas transparent, proportionerligt och med starkt skydd för yttrandefrihet (Palumbo, 2024; Husovec, 2024; Ó Fathaigh et al., 2025).

10 Bilaga. Transparens och reglering: DSA

Innan DSA beslutades höll flera medlemsstater på att införa nationella krav på aktsamhet för digitala tjänster och hur de skulle ta itu med olagligt innehåll, desinformation och andra samhällsrisker. Ett syfte med DSA var att harmonisera dessa regler inom EU och att motverka att kraven skiljde sig åt mellan medlemsstaterna. Reglerna skulle därför fastställas på unionsnivå, för att innovativa digitala tjänster skulle kunna växa fram och skalas upp i EU. De gemensamma reglerna skulle både vara enhetliga, effektiva och proportionerliga. Det fanns också ett behov av rättslig klarhet för plattformar och användare med respekt för grundläggande rättigheter, med tanke på den snabba tekniska utvecklingen.¹⁵

DSA innehåller regler för det som kallas förmedlingstjänster. Det är digitala tjänster som kopplar samman användare med varor, tjänster eller information. DSA gäller för leverantörer av förmedlingstjänster vars tjänster riktar sig till fysiska eller juridiska personer i en eller flera medlemsstater. Reglerna gäller oavsett var leverantören har sitt etableringsställe.

DSA delar in plattformar och förmedlingstjänster i olika kategorier: enbart vidarefordran (*mere conduit*), cachning, samt värdtjänster. Onlineplattformar är en slags värdtjänst. Onlineplattformar och sökmotorer som har 45 miljoner eller fler aktiva tjänstemottagare i EU i genomsnitt per månad kan utses av kommissionen till mycket stora onlineplattformar (VLOP) eller mycket stora online-sökmotorer VLOSE (jfr artikel 33). DSA gäller för leverantörer av förmedlingstjänster i olika storlek, men de mest långtgående skyldigheterna är inriktade mot VLOP och VLOSE.¹⁶ Mindre företag är undantagna från många regler. Exempel på VLOP är Facebook, Instagram, Tiktok och X. Google Search är exempel på VLOSE.¹⁷

Svenska forum och andra aktörer som inte är VLOP eller VLOSE kan ändå behöva följa vissa krav, beroende på deras storlek och vilken tjänst de tillhandahåller. Det kan exempelvis gälla krav på tydliga villkor, kontaktvägar, funktioner för att anmäla olagligt innehåll, motivering av vissa modereringsbeslut och, i vissa fall, funktioner för att klaga. Däremot gäller de mest långtgående kraven, till exempel krav på bedömning och hantering av systemrisker, främst VLOP och VLOSE.

I detta avsnitt beskrivs kraven för förmedlingstjänster översiktligt, från de grundläggande krav som gäller brett, till de mer långtgående krav som gäller för de mycket stora aktörerna. Flera av de krav som tas upp i avsnittet gäller alltså inte bara

¹⁵ eur-lex.europa.eu/legal-content/sv/TXT/PDF/?uri=CELEX:52020PC0825.

¹⁶ Uppgifter hämtade från Myndigheten för digital förvaltnings webbplats den 1 juni 2026.

¹⁷ Uppgifter hämtade från Europeiska unionens officiella webbplats den 26 maj 2026.

VLOP och VLOSE, utan även andra typer av förmedlingstjänster. Avsnittet beskriver inte exakt vilka krav som gäller på varje nivå. I stället följer avsnittet DSA:s struktur: från mer generella krav till de mest långtgående skyldigheterna.

10.1 DSA och skyldigheter för VLOP och VLOSE

VLOP och VLOSE måste ha transparenta allmänna villkor kring bland annat deras moderering och se till att ta hänsyn till bland annat yttrandefriheten när de genomför begränsningar för användningen av tjänsten (jfr artikel 14). VLOP och VLOSE ska också inkomma med rapporter om sin innehållsmoderering minst var sjätte månad (jfr artikel 15, 42). Rapporterna ska följa en standardiserad mall och redovisa plattformarnas arbete med innehållsmoderering och ärendehantering. Dessa grundläggande skyldigheter ger användare en viss nivå av insyn och transparens.

VLOP ska bland annat tillhandahålla mekanismer som gör att användare kan anmäla innehåll, motivera de begränsningar de gör av innehåll, ha ett internt system för hantering av klagomål, informera om tvistlösning utanför domstol och hantera anmälningar från vissa särskilt betrodda anmälare med prioritet. Betrodda anmälare kan exempelvis vara civilsamhällesorganisationer, eller andra enheter, och har vissa krav på aktsamhet och objektivitet. De kan arbeta med många olika frågor, exempelvis barnskydd eller bedrägeribekämpning. I Sverige finns i dagsläget två betrodda anmälare: ECPAT Sverige och Svenska Stöldskyddsföreningen. I Sverige tilldelas organisationer status som betrodd anmälare av Post- och telestyrelsen.

Vidare ska onlineplattformar också vidta särskilda åtgärder för att skydda mot missbruk av tjänsten, genom att under en rimlig tidsperiod och efter att först ha gett en varning också tillfälligt avbryta tillgång till tjänsten för aktörer som ofta lägger upp olagligt innehåll (jfr artikel 23). De har också krav på sig om mer omfattande transparensrapportering, de ska utforma sina gränssnitt så att de inte exempelvis vilseleder eller manipulerar användare och de ska informera om huvudparametrarna för deras rekommendationssystem (jfr artiklarna 24, 25 och 27).

Dessutom ska de minst en gång om året genomföra systematiska riskbedömningar av sina tjänster i EU. När riskerna har identifierats måste plattformarna vidta skäliga och effektiva åtgärder för att minska dem. Det kan innebära förändringar i rekommendationssystem, gränssnitt eller andra delar av tjänsternas design. Skyldigheten riktas alltså mot plattformen på systemnivå, inte enbart mot enskilda modereringsbeslut. Riskbedömningen ska bland annat titta på faktiska eller förutsägbara negativa effekter på yttrandefriheten och informationsfriheten.

VLOP och VLOSE ska också beakta hur deras rekommendationssystem och andra algoritmiska system, innehållsmoderering och allmänna villkor kan påverka riskerna.

De ska sedan vidta åtgärder för att minska dessa risker, genom att anpassa utformningen av tjänsten, deras villkor, system för innehållsmoderering eller rekommendationssystem. Om en tjänst har ett rekommendationssystem som består i att profilera någon, ska de också erbjuda ett alternativt rekommendationssystem som inte bygger på profilering (jfr artikel 34, 35 och 38).

Om en aktör inte följer förordningen kan det få vissa konsekvenser. För VLOP och VLOSE har kommissionen en möjlighet att ge aktören böter på 6 procent av den globala årsomsättningen. Om aktören inte deltar i tillsynen, och exempelvis vägrar svara eller lämnar felaktiga uppgifter, kan kommissionen dessutom ge en avgift på 1 procent av den totala årliga inkomsten eller globala omsättningen, eller ett löpande vite för att få dem att följa vissa beslutade åtgärder (jfr artikel 74–76 DSA).

Sammantaget innehåller DSA flera olika parallella skyldigheter som påverkar hur plattformarna designar system, rekommendationssystem och system för anmälningar.

10.2 DSA och självreglering

10.2.1 Frivilliga åtaganden och system för interna klagomål

DSA innehåller också bestämmelser som direkt tar sikte på och främjar frivilliga åtaganden. Bland annat ska kommissionen samråda med Europeiska nämnden för digitala tjänster (nedan kallad nämnden) om utveckling och genomförande av frivilliga standarder som fastställs av relevanta europeiska och internationella standardiseringsorgan rörande bland annat elektronisk inlämning av anmälningar enligt artiklarna 16 och 22 och standarder för mallar och processer för användarvänlig kommunikation med tjänstemottagare om begränsningar till följd av allmänna villkor (jfr artikel 44).

Kommissionen och nämnden ska uppmuntra och främja utarbetandet av frivilliga uppförandekoder på unionsnivå för att bidra till en korrekt och konsekvent tillämpning av DSA, med beaktande av i synnerhet utmaningarna med att hantera olika typer av olagligt innehåll och systemrisker. Sedan ska kommissionen och nämnden bland annat övervaka, utvärdera och offentliggöra hur uppförandekodernas mål uppnås (jfr artikel 45).

Kommissionen ska stödja och främja utvecklingen av frivilliga standarder, som sedan fastställs av standardiseringsorgan. Kommissionen och DSA-nämnden ska också uppmuntra och främja utarbetande av frivilliga uppförandekoder (artiklarna 44–47).

Det finns också krav på att onlineplattformar ska ha system för att kunna lämna in klagomål och ett system för en intern hantering av dessa klagomål. Den som klagat kan också välja ett organ för tvistlösning utanför domstol, hos ett organ som

certifierats särskilt för detta ändamål (artikel 20–21). Detta ersätter inte rätten att också få saken prövad av en domstol.

10.2.2 Metas tillsynsråd

Plattformarna har även utvecklat egna system för normbildning, prövning och självreglering. Ett centralt exempel är Metas tillsynsråd, *The Oversight Board*, som lanserades 2020 (Thorslund, 2023). Tillsynsrådet har till uppgift att överpröva modereringsbeslut på Facebook, Instagram och Threads. Det består av medlemmar från olika delar av världen och ska bedöma anmälda inlägg för att avgöra vad som ska tas bort och vad som får finnas kvar (Myndigheten för press, radio och tv, 2023). Rådet är ett självständigt beslutande organ och oberoende från Meta.

Användare kan anmäla innehåll direkt till rådet eller via de anmälningsfunktioner som finns på respektive plattform. Tillsynsrådet prioriterar ärenden som bedöms kunna påverka många användare globalt, som är betydelsefulla för det offentliga samtalet eller som väcker viktiga frågor om Metas användarvillkor (Thorslund, 2023). Rådet ska också bedöma om Metas borttagning av innehåll respekterar användarnas mänskliga rättigheter och yttrandefrihet.

Två fall där rådet överprövade Metas borttagning av innehåll gäller användare i Sverige. Det första ärendet är från 2023 och handlade om våld mot kvinnor där en person i ett inlägg på Instagram muntligen fördömde detta.¹⁸ I sitt beslut rekommenderade rådet Meta att uppdatera sina användarvillkor gällande hatiskt beteende, samt uppdatera sin interna vägledning till granskare för att säkerställa att den här typen av inlägg inte togs bort av misstag. Det andra ärendet är från 2021 och handlade om ett inlägg på Facebook där en svensk journalist rapporterade om sexuellt våld mot två barn.¹⁹ Meta ansåg att inlägget stred mot deras användarvillkor om sexuellt utnyttjande av barn, övergrepp mot barn och nakenbilder av barn. Enligt rådet var det dock tydligt att användaren rapporterade om en fråga av allmänt intresse och att journalisten fördömde sexuellt utnyttjande av barn.

10.3 Transparens om villkor och begränsningar

Artikel 14 reglerar villkor och transparens hos förmedlingstjänsterna, i synnerhet hur plattformarnas användarvillkor ska utformas och kommuniceras. Enligt artikel 14 ska leverantörer i sina allmänna villkor beskriva vilka begränsningar som de kan komma att tillämpa på information som användare har delat, utifrån bland annat leverantörens policy, förfaranden, åtgärder och verktyg för innehållsmoderering. Det

¹⁸ Tillsynsrådets beslut 2023–002-IG-UA.

¹⁹ Tillsynsrådets beslut 2021–016-FB-FBR.

vill säga hur de använder algoritmer för att fatta beslut eller hur de arbetar med mänsklig, manuell granskning och hur arbetsordningen ser ut för att hantera klagomål. Informationen ska vara klar och begriplig, användarvänlig och lättillgänglig – även för barn om tjänsten i huvudsak riktas till eller används av minderåriga. När leverantören sedan tillämpar och kontrollerar att reglerna följs ska det ske med hänsyn till användarnas grundläggande rättigheter, såsom yttrandefrihet. Artikel 14 kräver således att digitala tjänster är öppna, tydliga och rättvisa kring sina användarvillkor – särskilt vad gäller innehållsmoderering – och att de kommunicerar dessa regler på ett begripligt och rättssäkert sätt för alla användare.

Om en leverantör tar bort eller begränsar innehåll som en användare delat ska de, enligt artikel 17, motivera det: om de har begränsat innehållet på grund av att innehållet är olagligt eller strider mot leverantörernas allmänna villkor. Detta inkluderar både när leverantören avlägsnar innehåll en användare har delat, eller på andra sätt begränsar synligheten såsom genom att göra innehållet oåtkomligt. Det gäller också om leverantören skulle hindra användaren från att komma åt tjänsten eller (tillfälligt eller permanent) stänga av kontot. Användaren ska då bland annat få reda på vilka faktiska omständigheter som har legat till grund för beslutet och om automatiserade medel har använts för att upptäcka eller identifiera innehållet eller för att fatta beslutet. Om beslutet grundats på de allmänna villkoren, ska det framgå varför informationen som delats varit oförenlig med dessa. Även denna information ska vara så tydlig och lättbegriplig som möjligt. Plattformarna måste alltså vara öppna om hur och varför de gör modereringsbeslut.

10.4 Borttagning av innehåll och svensk grundlag

Den grundläggande lagstiftningen för svenska medier finns i tryckfrihetsförordningen (1949:105) och yttrandefrihetsgrundlagen (1991:1469). Tryckfrihetsförordningen omfattar tryckta medier, medan yttrandefrihetsgrundlagen omfattar radio, tv, film och video. Medieföretag har dessutom automatiskt grundlagsskydd enligt yttrandefrihetsgrundlagen för vissa publiceringar på internet.²⁰

Sedan 2002 kan även andra aktörer ansöka om ett så kallat utgivningsbevis för sin webbplats och därigenom få motsvarande grundlagsskydd. Ett utgivningsbevis innebär att webbplatsen omfattas av det särskilda yttrandefrihetsrättsliga systemet.

²⁰ Från den 1 januari 2023 gäller nya regler i yttrandefrihetsgrundlagen för webbsändningar. Med massmedieföretag avses förlag, dock inte för endast digitala produkter, tryckeri, nyhetsbyrå eller redaktion för tryckt tidning, radio eller tv.

Enligt Justitiekanslern (JK)²¹ gör DSA inte undantag för webbplatser på internet som omfattas av skyddet i yttrandefrihetsgrundlagen. Grundlagsskyddet gäller svenska webbplatser med utgivningsbevis och för webbplatser som, förenklat uttryckt, drivs av medieföretag. Skyddet innebär bland annat att ansvar för innehållet i en publicering enligt exklusivitetsprincipen endast kan aktualiseras i enlighet med yttrandefrihetsgrundlagen. Utgångspunkten är därför att det allmänna saknar möjlighet att stänga ned eller kräva nedstängning av en webbplats vars verksamhet omfattas av yttrandefrihetsgrundlagens tillämpningsområde. Detsamma gäller blockering eller krav på borttagning av innehåll.

För en enskild aktör bakom en grundlagsskyddad webbplats kan det inte uteslutas att det leder till en potentiell konflikt mellan förelägganden enligt DSA och svensk grundlag, där aktören ytterst kan hänvisa till svensk grundlag och låta bli att agera på ett föreläggande om borttagning av innehåll utan att samtidigt bryta mot unionsrätten. Det finns enligt JK en potentiell osäkerhet kring hur begränsningen i DSA förhåller sig till svensk grundlag.

En annan konflikt rör DSA:s krav på system för hantering av klagomål. När en anmälan om olagligt innehåll har nått en leverantör inträder ett ansvar att skyndsamt agera på anmälan. Mot bakgrund av de kännbara sanktioner som är kopplade till överträdelser av DSA finns enligt JK en tydlig risk för övermoderering där även innehåll som inte är olagligt kan komma att tas bort för att undvika sanktioner. Resultatet kan bli att även lagligt innehåll tas bort och därmed en alltför långtgående begränsning av yttrande- och informationsfriheten.

DSA innehåller inte några bestämmelser om exakt vilket innehåll som är olagligt. Däremot finns det krav på att plattformar ska kunna hantera olagligt innehåll och, för de största aktörerna, att de ska kunna bedöma och motverka systemrisker. DSA innehåller också skydd mot att regleringen används eller tillämpas för långtgående: ansvarsfrihetsreglerna består, allmän övervakningsskyldighet är förbjuden och användare ges rätt till motivering, klagomål och extern tvistlösning. Regleringen är därför inte en enkel begränsning av yttrandefriheten, utan kan beskrivas som ett försök att göra plattformarnas hantering av innehåll mer transparent och kontrollerbar, snarare än en reglering av vad som exakt får sägas online.

²¹ Justitiekanslerns remissyttrande (dnr 2023/7346) på betänkandet *En inre marknad för digitala tjänster – kompletteringar och ändringar i svensk rätt* (SOU 2023:39).

11 Referenser

- Acosta Navas, D. (2025). In Moderation: Automation in the Digital Public Sphere. *Journal of Business Ethics*, 200:325–341.
- Alnemr, N. (2025). Deliberative democracy in an algorithmic society: harms, contestations and deliberative capacity in the digital public sphere. *Democratization*. DOI: 10.1080/13510347.2025.2522920
- Anstead, N., Edwards, L., Livingstone, S., & Stoilova, M. (2025). The Potential for Media Literacy to Combat Misinformation: Results of a Rapid Evidence Assessment. *International Journal of Communication*, 19.
- Are, C. (2022). The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, 22(8), 2002–2019.
- Ashley, S., Maksl, A., & Craft, S. (2017). News Media Literacy and Political Engagement: What’s the Connection? *Journal of Media Literacy Education*, 9(1), 79–98.
- Aytac, U. (2024). Digital Domination: Social Media and Contestatory Democracy. *Political Studies*, 72(1) 6–25.
- Badouard, R., & Bellon, A. (2025). Introduction to the special issue on content moderation on digital platforms, *Internet Policy Review*, in cooperation with Alexander von Humboldt Institute for Internet and Society, Berlin, 14(1), pp. 1-24.
- Balčytienė, A., Bajomi-Lázár P., & Sousa H. (Eds.). (2026). *Digital media shadowing democracy: Technology, communication, and power*. Nordicom, University of Gothenburg.
- Balendra, S. (2025). Meta’s AI moderation and free speech: Ongoing challenges in the Global South. *Cambridge Forum on AI: Law and Governance* (2025), 1, e21, 1–19 doi:10.1017/cfl.2025.5
- Bhandari, A., Ozanne, M., Bazarova, N. N., & DiFranzo, D. (2021). Do You Care Who Flagged This Post? Effects of Moderator Visibility on Bystander Behavior. *Journal of Computer-Mediated Communication*, 26 (2021) 284–300.
- Björkenfeldt, O. (2025). *Journalisters säkerhet – en kunskapsöversikt*. Mediemyndigheten.
- boyd, d. (2026). Social Media Is Now Parasocial Media. *Social Media + Society*, April-June 2026, 1–5.
- Broda, E., & Strömbäck, J. (2024). Misinformation, Disinformation, and Fake News: Lessons from an Interdisciplinary, Systematic Literature Review. *Annals of the International Communication Association*, 48(2), 139–166.

Brottsförebyggande rådet (2025). *Nationella trygghetsundersökningen. Om utsatthet, otrygghet och förtroende*. Rapport 2025: 15.

Brottsförebyggande rådet (2025). *Politikernas trygghetsundersökning. Förtroendevaldas utsatthet och oro för trakasserier, hot och våld under 2024*. Rapport 2025: 17.

Bruns, A. (2008). *Blogs, Wikipedia, Second Life, and Beyond: From Production to Prodisage*. Peter Lang.

Caplan, R. (2023). Networked Platform Governance: The Construction of the Democratic Platform. *International Journal of Communication*, 17 (2023), 3451–3472.

Carlsson, U. (2023). Medie- och informationskunnighet (MIK): Begreppet och kunskapen. I U. Carlsson, P.-M. Wehring, & K. Westling (Eds.), *Medie- och informationskunnighet (MIK) för ett demokratiskt medborgarskap: Rapport från nationell MIK-konferens 2022* (pp. 37–46). Förvaltningen för kulturutveckling, Västra Götalandsregionen.

Carmi, E. (2020). Out with the deviant, in with the social. *Internet Policy review*, 9(4).

Carretta, S. (2025). *Online Content Moderation. The regulatory continuum from EU regulation to platform self-regulation*. [Doktorsavhandling]. Uppsala universitet.

Caruso, C. (2025). Towards the Institutions of Freedom: The European Public Discourse in the Digital Era. *German Law Journal*, 26, pp. 114–137.

Chancellor, S., Kalantidis, Y., Pater, J. A., De Choudhury, M., & Shamma, D. A. (2017). Multimodal Classification of Moderated Online Pro-Eating Disorder Content. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3213–3226.

Chen, Y.-S., & Zaman, T. (2024). Shaping opinions in social networks with shadow banning. *PLoS ONE* 19(3): e0299977. <https://doi.org/10.1371/journal.pone.0299977>

Chong, M. (2024). Twitter and the Affordance: A Case Study of Participatory Roles in the #Marchforourlives Network. *Digital*, 2024, 4, 660–675.

Corsi, G. (2024). Evaluating Twitter's algorithmic amplification of low-credibility content: an observational study. *EPJ Data Science*, 13(18).
<https://doi.org/10.1140/epjds/s13688-024-00456-3>

Cotter, K. (2023). "Shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*, 26(6), 1226–1243.

Dahlgren, P. M. (2021). A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Review*, 42(1), 15–33.

- Delmonaco, D., Mayworm, S., Thach, H., Guberman, J., Augusta, A., & Haimson, O.L. (2024). “What are you doing, TikTok?”: How Marginalized Social Media Users Perceive, Theorize, and “Prove” Shadowbanning. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), Article 154.
- Dimitrova, D. V., Shehata, A., Strömbäck, J., & Nord, L. W. (2014). The Effects of Digital Media on Political Knowledge and Participation in Election Campaigns: Evidence from Panel Data. *Communication Research*, 41(1), 95–118.
- Drolsbach, C. P., & Pröllochs, N. (2024). Content moderation on social media in the EU: Insights from the DSA Transparency Database. *WWW '24: The ACM Web Conference 2024, May 13 - 17, 2024, Singapore, Singapore*.
- Droog, E., Vermeulen, I., van Huijstee, D., Harutyunyan, D., Tejedor, S., & Pulido, C. (2025). Combatting the Misinformation Crisis: A Systematic Review of the Literature on Characteristics and Effectiveness of Media Literacy Interventions. *Communication Research*.
<https://doi.org/10.1177/00936502251363705>
- Dumbrava, C. (2021). *Key social media risks to democracy. Risks from surveillance, personalisation, disinformation, moderation and microtargeting*. Members’ Research Service, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.
- Dwork, C., Hays, C., Kleinberg, J., & Raghavan, M. (2024). Content Moderation and the Formation of Online Communities: A Theoretical Framework. In *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore*, 1307–1317.
- Eder, N. (2024). Making Systemic Risk Assessments Work: How the DSA Creates a Virtuous Loop to Address the Societal Harms of Content Moderation. *German Law Journal*, 25, pp. 1197–1218.
- Einarsson, A. M., Helles, R., & Lomborg, S. (2025). Algorithmic agenda setting: the subtle effects of news recommender systems on political agendas in the Danish 2022 general election. *Information, Communication & Society*, 28(2), 218-238.
- El Emary, I. M. M., & Alhebi, M. A. (2025). The Role of Media Literacy in Shaping Public Opinion and Political Participation in the Digital Era. *IJIS International Journal of Informatics and Information Systems*, 8(3), 148–160. DOI: 10.47738/ijis.v8i3.271
- Erikson, J., Håkansson, S., & Josefsson, C. (2023). Three Dimensions of Gendered Online Abuse: Analyzing Swedish MPs’ Experiences of Social Media. *Perspectives on Politics*, 21(3), 896–912. <https://doi.org/10.1017/S1537592721002048>

Eslami, M. et al. (2015). Reasoning about invisible algorithms in news feeds. *CHI 2015, April 18- 23, 2015, Seoul, Republic of Korea*. CDOI: 10.1145/2702123.2702556

Eslami, M. et al. (2016). First I “like” it, then I hide it: Folk Theories of Social Feeds. *CHI’16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2371–2382. <https://doi.org/10.1145/2858036.2858494>

FOI – Totalförsvarets forskningsinstitut. (2022). *Varning – desinformation! Allmänhetens syn på psykologiskt försvar*. Rapport FOI-R--5264--SE.

Fornusek, M. (2026, 26 januari). Fact-check: Russia’s ‘most successful disinformation campaign’ targets free speech in Europe. *Kyiv Independent*. [Fact-check: Russia's 'most successful disinformation campaign' targets free speech in Europe](#)

Galantino, S. (2023). How Will the EU Digital Services Act Affect the Regulation of Disinformation? *Scripted*, 20(1), 89–129.

Gandini, A., Gerosa, A., Gobbo, B., Keeling, S., Leonini, L., Mosca, L., Orofino, M., Reviglio, U., & Splendore, S. (2022). The algorithmic public opinion: a literature review. Public perception of algorithms in society: accounting for the algorithmic public opinion. *Algocount 2020–2022*.

Gandini, A., Keeling, S., & Reviglio, U. (2025). Conceptualising the ‘algorithmic public opinion’: Public opinion formation in the digital age. *Dialogues on Digital Society* 1–16. DOI: 10.1177/29768640251323147

Gausen, A., Guo, C., & Luk, W. (2025). An approach to sociotechnical transparency of social media algorithms using agent-based modelling. *AI and Ethics*, 5, 1827–1845.

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12), 4492–4511.

Gerrard, Y. (2022). Social Media Moderation: The Best-Kept Secret in Tech. In D. Rosen (Ed.), *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media* (pp. 77–95). Routledge.

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., Roberts, S. T., Sinnreich, A., & Myers West, S. (2020). Expanding the debate about content moderation: scholarly research agendas for the coming policy debates. *Internet Policy Review*, 9(4). DOI: 10.14763/2020.4.1512

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, July-September 2022, 1–13.

Goldman, E. (2021). Content moderation remedies. *Michigan Technology Law Rev*, 28(1), 1–61.

Golunova, V. (2025). Algorithmic content moderation and freedom of expression under the digital services act. *European Review of Public Law*, 37(2), 519–554.

González-Bailón, S., Lazer, D., Barberá, P., Zhang, M., Allcott, H., Brown, T., Crespo-Tenorio, A., Freelon, D., Gentzkow, M., Guess, A. M., Iyengar, S., Kim, Y. M., Malhotra, N., Moehler, D., Nyhan, B., Pan, J., Rivera, C. V., Settle, J., Thorson, E., ... Tucker, J. A. (2023). Asymmetric ideological segregation in exposure to political news on Facebook. *Science*, 381(6656), 392–398. <https://doi.org/10.1126/science.ade7138>

Goodwin, A., Joseff, K., Riedl, M.J., Lukito, J., & Woolley, S. (2023). Political Relational Influencers: The Mobilization of Social Media Influencers in the Political Arena. *International Journal of Communication*, 17(2023), 1613–1633.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, January–June, 1–15.

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law and Technology*, 17(1), 42–109.

Groesch, S., Birrer, A., Just, N., & Saurwein, F. (2026). Big data, small answers: How the DSA Transparency Database falls short of its regulatory objectives. *Telecommunications Policy*, 50(1), Article 103088.

Guess, A. M. et al. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656), 398–404.

Habermas, J. (1962/1984). *Borgerlig offentlighet*. Moderna klassiker, Arkiv förlag.

Haider, J., & Sundin, O. (2017). Algoritmer. Så påverkar de din vardag. *IIS Internetguide*, nr 46. <https://internetstiftelsen.se/app/uploads/2021/01/algoritmer-internetguide.pdf>

Haider, J., & Sundin, O. (2019). How Do you Trust? On Infrastructural Meaning-making and the Need for Self-reflection. In U. Carlsson (Ed.), *Understanding Media and Information Literacy (MIL) in the Digital Age: A question of Democracy* (pp. 107). University of Gothenburg and UNESCO.

Haider, J., & Sundin, O. (2022). Information literacy challenges in digital culture: conflicting engagements of trust and doubt. *Information, Communication & Society*, 25(8), 1176–1191.

Harff, D., & Schmuck, D. (2024). Who Relies on Social Media Influencers for Political Information? A Cross-Country Study Among Youth. *The International Journal of Press/Politics*, 30(3), 841–864.

- Harff, D., & Schmuck, D. (2025). Prevalence, Presentation, and Popularity of Political Topics in Social Media Influencers' Content Across Two Countries. *Political Communication*, 42(3), 351–381.
- Harris, B.C., Foxman, M., & Partin, W.C. (2023). “Don’t Make Me Ratio You Again”: How Political Influencers Encourage Platformed Political Participation. *Social Media + Society*, 9(2). <https://doi.org/10.1177/20563051231177944>
- Holt, K., Shehata, A., Strömbäck, J., & Ljungberg, E. (2013). Age and the Effects of News Media Attention and Social Media Use on Political Interest and Participation: Do Social Media Function as Leveller? *European Journal of Communication*, 28(1), 19–34.
- Husovec, M. (2024). The Digital Services Act’s red line: what the Commission can and cannot do about disinformation. *Journal of Media Law*, 16(1), 47–56.
- Huszár, F., Ktena, S. I., O’Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. *PNAS* 2022, 119(1), e2025334119.
- Internetstiftelsen. (2025). *Svenskarna och internet 2025*.
- Internetstiftelsen. (2026). *Svenskarna och internet: Valspecial 2026*.
- Jaidka, K., Mukerjee, S., & Lelkes, Y. (2023). ‘Silenced on social media: the gatekeeping functions of shadowbans in the American Twitterverse’. *Journal of Communication*, 73(2), 163–178.
- Jenkins, H. (2008). *Konvergenskulturen. Där nya och gamla medier kolliderar*. Daidalos.
- Jhaver, S., Zhang, A. Q., Ze Chen, Q., Natarajan, N., Wang, R., & Zhang, A. X. (2023). Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 289 (October 2023), 33 pages.
- John, N. (2024). Sharing and social media: The decline of a keyword? *new media & society*, 26(4) 1891–1908.
- Jung, H., Dai, W., & Albarracín, D. (2023). How Social Media Algorithms Shape Offline Civic Participation: A Framework of Social-Psychological Processes. *Perspectives on Psychological Science*, 1–14. <https://doi.org/10.1177/17456916231198471>
- Jungselius, B., & Weilenmann, A. (2025). Tracing Change in Social Media Use: A Qualitative Longitudinal Study. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan*, article no. 957, 1–14.
- Kahne, J., Lee, N.-J., & Feezell, J. T. (2012). Digital Media Literacy Education and Online Civic and Political Participation. *International Journal of Communication*, 6, 1–24.

Karabulut, D., Ozcinar, C., & Anbarjafari, G. (2022). Automatic content moderation on social media. *Multimedia Tools and Applications*, 82(3), 4439–4463.

Karizat, N., Delmonaco, D., Eslami, M., & Andalibi, N. (2021). Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), article no. 305, 1–44.

Kashyap, B., & Formosa, P. (2023). The Authority to Moderate: Social Media Moderation and its Limits. *Philosophy & Technology*, 36(78).
<https://doi.org/10.1007/s13347-023-00685-w>

Kojah, S.A., Zhang, B.Z., Are, C., Delmonaco, D., & Haimson, O.L. (2025). “Dialing it Back”: Shadowbanning, Invisible Digital Labor, and how Marginalized Content Creators Attempt to Mitigate the Impacts of Opaque Platform Governance. *Proceedings of the ACM on Human-Computer Interaction*, 9(1). DOI: 10.1145/3701191

Kentmen-Cin, C. (2025). Hate Speech on Social Media: A Systemic Narrative Review of Political Science Contributions. *Social Sciences*, 14(610).
<https://doi.org/10.3390/socsci14100610>

Lanier, J. (2011). *You Are Not a Gadget: A Manifesto*. Alfred A. Knopf.

Leerssen, P. (2023). An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. *Computer Law & Security Review*, 48, April 2023, 105790.

Le Merrer, E., Morgan, B., & Trédan, G. (2021). Setting the record straighter on shadow banning. *IEEE INFOCOM 2021- IEEE Conference on Computer Communications*, 1–10.

Lewis, C. (2014). *Irresistible apps: Motivational design patterns for apps, games, and web-based communities*. Apress.

Lilja, J., Eklund, N., & Tottie, E. (2024). Civic Literacy and Disinformation in Democracies. *Social Sciences*, 13(8), 405.

Lindgren, S. (2024). *Critical theory of AI*. Polity.

Lookingbill, V., & Le, K. (2024). “There’s Always a Way to Get Around the Guidelines”: Nonsuicidal Self-Injury and Content Moderation on TikTok. *Social Media + Society*, 10(2), Article 20563051241254371.

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature human behaviour* 7, January 2023, 74–101.

Lundqvist, B. (2024). En modererad mediemarknad – så ska nya regler jämna ut spelplanen och främja yttrandefriheten. *Sieps – Europapolitisk analys*, juni 2024:17epa. Svenska institutet för europapolitiska studier.

Martel, C., Berinsky, A. J., Rand, D. G., Zhang, A. X., & Resnick, P. (2025). Perceived legitimacy of layperson and expert content moderators. *PNAS Nexus*, 2025, 4, pgaf111.

Mayworm, S., Li, S., Thach, H., Delmonaco, D., Paneda, C., Wegner, A., & Haimson, O. L. (2024). The Online Identity Help Center: Designing and Developing a Content Moderation Policy Resource for Marginalized Social Media Users. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 129, <https://doi.org/10.1145/3637406>

Merz, P., & von Sikorski, C. (2025). Social Media Influencers Can Increase Collective Political Beliefs and Actions: Findings from Experiments and a Quasi-Experimental Field Study. *Human Communication Research*. Hqaf024 <https://doi.org/10.1093/hcr/hqaf024>.

Mediemyndigheten. (2024a). *Mediekonsumtion 2024*. Mediemyndigheten.

Mediemyndigheten. (2024b). *Kartläggning av medie- och informationskunnighet i Sverige*. Mediemyndigheten.

Medieakademin. (2025). *Maktbarometern 2025*. Medieakademin.

Meltwater & We Are Social (2025). *Digital 2026 Global Overview Report*. Datareportal.

Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, 27(4), zmac010.

Morozov, E. (2011). *The net delusion: the dark side of Internet freedom*. PublicAffairs.

Mosco, V. (2004). *The Digital Sublime. Myth, Power, and Cyberspace*. The MIT Press.

Myers West, S. (2022). Social Media as Social Infrastructures. In D. Rosen (Ed.), *The Social Media Debate. Unpacking the Social, Psychological, and Cultural Effects of Social Media* (pp. 5–19). Routledge.

Myndigheten för press, radio och tv. (2023). *Medieutveckling 2023: Mediepolitik*. Myndigheten för press, radio och tv.

Myndigheten för psykologiskt försvar (MPF). (2018). *Att möta informationspåverkan: Handbok för kommunikatörer*. Myndigheten för psykologiskt försvar.

Nationalencyklopedin, opinionsbildning.

<https://www.ne.se/uppslagsverk/encyklopedi/lång/opinionsbildning> (hämtad 2026-06-08), NE Nationalencyklopedin AB.

Nationalencyklopedin, tystnadsspiralen.

<https://www.ne.se/uppslagsverk/encyklopedi/lång/tystnadsspiralen> (hämtad 2026-06-08), NE Nationalencyklopedin AB.

Newman, N. et al. (2025). *Digital News Report 2025*. Reuters Institute for the Study of Journalism.

Nguyen, C.T. (2020). Echo chambers and epistemic bubbles, *Episteme*, 17(2), 141–161. <https://doi.org/10.1017/epi.2018.32>

Norrby, C. (2022). *Algorithmic awareness och algorithmic literacy. En systematisk litteraturstudie av två begrepp* [Examensarbete]. Institutionen för tillämpad IT, Göteborgs universitet.

Nordicom. (2026). *Mediebarometern 2025*. Göteborgs universitet.

Nordicom. (2025). *Mediebarometern 2024 Tema sociala medier*. Göteborgs universitet.

Nordiska ministerrådet. (2025/2026). Nordic Media Literacy Survey. *TemaNord* 2025:572. Nordiska ministerrådet.

Nygren, T., & Guath, M. (2022). Students Evaluating and Corroborating Digital News. *Scandinavian Journal of Educational Research*, 66(4), 549–565.

Nyhan, B. et al. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, 620(7972), 137–144.

Oeldorf-Hirsch, A., & Neubaum, G. (2023). Attitudinal and behavioral correlates of algorithmic awareness among German and U.S. social media users. *Journal of Computer-Mediated Communication*, 28(5), DOI: 10.1093/jcmc/zmad035.

Ó Fathaigh, R., Buijs, D., & van Hoboken, J. (2025). The Regulation of Disinformation Under the Digital Services Act. *Media and Communication*, 13, Article 9615.

Oser, J., & Boulianne, S. (2020). Reinforcement effects between digital media use and political participation: A meta-analysis of repeated-wave panel data. *Public Opinion Quarterly*, 84, Special Issue, 355–365.

Palumbo, A. (2024). A Medley of Public and Private Power in DSA Content Moderation for Harmful but Legal Content: An Account of Transparency, Accountability and Redress Challenges. *JIPITEC*, 15(1), 246–268.

- Pan, C. A., Yakhmi, S., Iyer, T. P., Strasnick, E., Zhang, A. X., & Bernstein, M. S. (2022). Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), Article 82.
- Pariser, E. (2012). *The Filter Bubble. What The Internet Is Hiding From You*. Penguin Books.
- Peterson-Salahuddin, C. (2024). Repairing the harm: Toward an algorithmic reparations approach to hate speech content moderation. *Big Data & Society*, April–June 2024, 1–13.
- Poell, T., Nieborg, D., & van Dijck, J. (2019). Platformisation. *Internet Policy Review*, 8(4). <https://doi.org/10.14763/2019.4.1425>
- Post- och telestyrelsen. (2026). *Verksamhetsrapport 2025 (Annual Activity Report)*. Rapportnummer PTS-ER-2026:14.
- Quintais, J.P., Appelman, N., & Ó Fathaigh, R. (2023). Using Terms and Conditions to apply Fundamental Rights to Content Moderation. *German Law Journal*, 24, 881–911.
- Register, Y., Qin, L., Baughan, A., & Spiro, E.S. (2023). Attached to “The Algorithm”: Making Sense of Algorithmic Precarity on Instagram. In *CHI Conference on Human Factors in Computing Systems (CHI '23), April 23-28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 15 pages.
- Remissyttrande (dnr 2023/7346) från Justitiekanslern på betänkandet SOU 2023:39 *En inre marknad för digitala tjänster – kompletteringar och ändringar i svensk rätt*.
- Reviglio, U., Bleyer-Simon, K., & Verza, S. (2024). Governing Social Media’s Opinion Power: The Interplay of EU Regulations. *MediaLaws*, numero speciale I-2024, 108–135.
- Riedl, M.J., Lukito, J., & Woolley, S.C. (2023). Political Influencers on Social Media: An Introduction. *Social Media + Society*, April-June 2023, 1–9.
- Riemer, K., & Peter, S. (2021). Algorithmic audiencing: Why we need to rethink free speech on social media. *Journal of Information Technology*, 36(4), 409–426.
- Risius, M., & Blasiak, K.M. (2024). Shadowbanning. An Opaque Form of Content Moderation. *Bus Inf Syst Eng* 66(6), 817–829.
- Saleous, H., Gergely, M., & Shuaib, K. (2025). Exploring NLP-Based Solutions to Social Media Moderation Challenges. *Human Behavior and Emerging Technologies*, 2025(1), 9436490.

Savolainen, L. (2022). The shadow banning controversy: perceived governance and algorithmic folklore. *Media, Culture & Society*, 44(6), 1091–1109.

Schaffner, B. et al. (2024). Community Guidelines Make this the Best Party on the Internet. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 16 pages.

Scharfbillig, M., Lewandowsky, S., Altay, S., van Alstyne, M., Kozyreva, A. et al. (2026). *Fractured reality – How democracy can win the global struggle over the information space*. Publications Office of the European Union, Luxembourg.

Sehl, A., & Schützeneder, J. (2023). Political Knowledge to Go: An Analysis of Selected Political Influencers and Their Formats in the Context of the 2021 German Federal Election. *Social Media + Society*, April-June 2023, 1–12.

SFS 1949:105. *Tryckfrihetsförordning*.

SFS 1991:1469. *Yttrandefrihetsgrundlag*.

Shahi, G. K., Tessa, B., Trujillo, A., & Cresci, S. (2025). *A year of the DSA Transparency Database: What it (does not) reveal about platform moderation during the 2024 European Parliament election*. Association for the Advancement of Artificial Intelligence. <https://arxiv.org/abs/2504.06976>

SOU 2007:107. *Opinionsbildande verksamhet och små myndigheter. Delbetänkande av 2006 års förvaltningskommitté*.

SOU 2020:56. *Det demokratiska samtalet i en digital tid. Så stärker vi motståndskraften mot desinformation, propaganda och näthat*.

SOU 2026: 35. *En åldersgräns för barns tillgång till sociala medier. Delbetänkande av Utredningen om en åldersgräns på sociala medier*.

Staal Bruun Overgaard, C. (2024). Perceiving Affective Polarization in the United States: How Social Media Shape Meta-Perceptions and Affective Polarization. *Social Media + Society*, January-March 2024, 1 –12.

Steen, E., Yurechko, K., & Klug, D. (2023). You Can (Not) Say What You Want: Using Algospeak to Contest and Evade Algorithmic Content Moderation on TikTok. *Social Media + Society*, July-September 2023, 1 –17.

Steiger, M., Bharucha, T.J., Venkatagiri, S., Riedl, M.J., & Lease, M. (2021). The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *CHI Conference on Human Factors in Computing Systems (CHI '21)*, May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 14 pages.

- Sundin, O., & Haider, J. (2018). Källkritik, självkritik och källtillit. I U. Carlsson (Red.), *Medie- och informationskunnighet i den digitala tidsåldern – en demokratifråga: Kartläggning, analyser, reflektioner* (s. 59–61). Nordicom.
- Söderlund, K., Engström, E., Haresamudram, K., Larsson, S., & Strimling, P. (2024). Regulating high-reach AI: On transparency directions in the Digital Services Act. *Internet Policy Review* 13(1). <https://doi.org/10.14763/2024.1.1746>
- Thomas, S., & Manalil, P. (2025). Digital silence: the psychological impact of being shadow banned on mental health and self-perception. *Frontiers in Psychology*, 16, 1659272.
- Thorell, L., Dorrestein, M., Wurth, P., & Nutley, S. (2026). *Hur påverkar innehåll och kommunikation på sociala medier ådstörningar och självskadebeteende? Systematiska översikter*. Mediemyndigheten.
- Thorslund, F. (2023). Nya medier, nya remedier – innehållsmoderering och yttrandefrihet på sociala medieplattformar i EU och USA. *Juridisk Publikation*, särtryck ur häfte 1/2023.
- Tillsynsrådets beslut 2021–016-FB-FBR.
- Tillsynsrådets beslut 2023–002-IG-UA.
- Trujillo, A., Fagni, T., & Cresci, S. (2025). The DSA Transparency Database: Auditing self-reported moderation actions by social media. *Proceedings of the ACM on Human-Computer Interaction*, 9(2), 1–28.
- Törnberg, P., & Törnberg, A. (2024). Inside a White Power echo chamber: Why fringe digital spaces are polarizing politics. *New Media & Society*, 26(8), 4511–4533.
- van de Kerkhof, J. (2025). The DSA's Tower of Babel: On Digital Services Coordinators and Freedom of Expression. *European Journal of Risk Regulation*, 16, 1392–1417.
- van Dijck, J. (2013). *The Culture of Connectivity: A Critical History of Social Media*. Oxford University Press.
- Vasist, P. N., Chatterjee, D., & Krishnan, S. (2023). The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-country Configurational Narrative. *Information Systems Frontiers*, April 17, 1–26. <https://doi.org/10.1007/s10796-023-10390-w>.
- Velkova, J., & Kaun, A. (2019). Algorithmic resistance: media practices and the politics of repair. *Information, Communication & Society*, August 2019, 1–18. <https://doi.org/10.1080/1369118X.2019.1657162>

von Sikorski, C., Merz, P., Heiss, R., Karsay, K., Naderer, B., & Schmuck, D. (2025). The Political Role of Social Media Influencers: Strategies, Types, and Implications for Democracy—An Introduction. *American Behavioral Scientist*, 1–17. DOI: 10.1177/00027642251344208

Wagnsson, C., & Östervall, A. (2025). Who Believes in Foreign Disinformation? Evidence from a Disinformation Discernment Exercise in Sweden. *Political Research Exchange*, 7(1). <https://doi.org/10.1080/2474736X.2025.2501039>

Wasike, B. (2023). I Am an Influencer and I Approve This Message! Examining How Political Social Media Influencers Affect Political Interest, Political Trust, Political Efficacy, and Political Participation. *International Journal of Communication*, 17, 3110–3132.

Williams, M. (2025). Social media democracy: How algorithms shape public discourse and marginalise voices. *Journal of Media and Rights*, 3(1), a20. <https://doi.org/10.4102/jmr.v3i1.20>

Willcox, M. (2025). Algorithmic agency and “fighting back” against discriminatory Instagram content moderation: #IWantToSeeNyome. *Frontiers in Communication*, 9:1385869. doi: 10.3389/fcomm.2024.1385869

Wolters, P. & Gellert, P. (2023). Towards a better notice and action mechanism in the DSA. *JIPITEC*, 14(1), 403–419.

Ytre-Arne, B., & Moe, H. (2021). Folk theories of algorithms: Understanding digital irritation. *Media, Culture & Society*, 43(5), 807–824. DOI: 10.1177/0163443720972314.

Yu, X., Haroon, M., Menchen-Trevino, E., Wojcieszak, M., (2024). Nudging recommendation algorithms increases news consumption and diversity on YouTube. *PNAS Nexus*, 3(12), pgae518.

Zuboff, S. (2021). *Övervakningskapitalismen. Vid maktens nya frontlinjer*. Ordfront.