Myndigheten för psykologiskt försvar

ANDREAS EDEVALD OCH JOHAN ÖSTERBERG

Janusansiktet

– en metafor för informations-påverkan och kognitiva bias

MPF SKRIFTSERIE 9/2025

9/2025

ISBN: 978-91-989646-6-0

Produktion: Gritti

Tryck: Scandinavian Print Group

Bilden på framsidan är framställd med hjälp av AI.

Språkmodeller har använts som ett redaktionellt stöd vid språklig bearbetning av texten. Analys, källvärdering, bedömningar och slutsatser har gjorts av författarna, som också ansvarar för textens innehåll i dess helhet.

Innehåll

Förord

”Det moderna kriget drabbar allt och alla. Frontlinjen går bildligt talat i varje enskild medborgares medvetande.” Detta kunde 1950-talets svenskar läsa i den utredning om psykologiskt försvar som publicerades 1953 (SOU 1953:29, s. 18). Och de orden är fortfarande lika aktuella i dag som då de skrevs för drygt 70 år sedan.

Det informations- och kommunikationssamhälle som vi lever i dag är naturligtvis helt annorlunda och vi ser nu enormt snabba teknologiska framsteg särskilt inom området artificiell intelligens. Psykologisk krigföring, informationspåverkan, informationskrig, eller vilket begrepp man nu vill använda, handlar dock, precis som 1953, fortfarande om att påverka hur vi uppfattar verkligheten, vad vi tror och litar på och hur vi utifrån det agerar och fattar beslut. Det är därför viktigt att förstå hur vi människor tar till oss information och hur vi formar tankar till för oss sammanhängande system.

AI-system kan förstärka vår förståelse, effektivisera beslutsfattande och lösa komplexa problem, men de kan också förvanska information, förstärka kognitiva bias och utnyttjas för otillbörlig informationspåverkan. Först när vi förstår våra sårbarheter kan vi på allvar påbörja ett arbete med att stärka vår digitala motståndskraft och stå rustade inför den informationspåverkan som främmande makt riktar mot Sverige.

Syftet med den här rapporten är att bidra med en nyanserad förståelse för den dubbelhet som präglar AI-utvecklingen och att ge beslutsfattare, forskare och allmänhet verktyg för att navigera i vår tids informationsmiljö. Författare är Johan Österberg och Andreas Edevald, som båda är anställda vid Myndigheten för psykologiskt försvar men som här skriver i egenskap av fristående forskare. Det är författarna som ensamma ansvarar för rapportens innehåll och slutsatser.

Stockholm, augusti 2025

MAGNUS HJORT

Generaldirektör

Sammanfattning

Sammanfattning

I dagens samhälle där informationsflödet är konstant och ofta överväldigande, uppstår naturliga sårbarheter som kan utnyttjas av motståndare med strategiska avsikter. Genom psykologisk påverkan – där desinformation, rykten och riktad propaganda används – kan en aktör undergräva tillit, splittra opinioner och skapa oro hos befolkningen. Dessa angrepp sker inte sällan i gråzonen mellan fred och konflikt, där effekterna är svåra att upptäcka innan de fått fäste. Föreliggande studie beskriver den dubbla naturen av upprepning i mänsklig kognition, vilket framhäver upprepningens roll både för lärande, men även för spridande av desinformation. Studien granskar hur upprepning kan befästa sann kunskap, men också göra individer mottagliga för desinformation genom kognitiva snedvridningar som exempelvis den illusoriska sanningseffekten och bekräftelsebias. Vidare kan artificiell intelligens (AI) fungera som ett ”Janusansikte”-verktyg, förmögen att både förstärka desinformation genom massgenerering och algoritmisk förstärkning, samt fungera som en kraftfull lösning för att motbevisa falska berättelser och främja kritiskt tänkande. För att motverka de negativa effekterna av upprepning och AI, krävs en medveten ansträngning för att övervinna snedvridningar och främja medie- och informationskunskap.

Abstract

This study examines the inherent duality of repetition in human cognition, drawing parallels to the Roman god Janus, who looks to both the past and the future. Repetition is foundational to learning and memory (”Repetition is the mother of knowledge”), yet it also makes individuals and societies vulnerable to deception (”If a lie is repeated often enough, it becomes true”). Furthermore, the study explores key cognitive biases—illusory truth effect, confirmation bias, motivated reasoning, and knowledge neglect—that explain why repeated false information can gain traction. Critically, the rise of Artificial Intelligence (AI) is presented as a ”Janus-faced” tool, capable of vastly amplifying both the positive and negative impacts of repetition. While AI can exacerbate the spread of disinformation through mass generation, targeted influence, and algorithmic reinforcement, it also holds significant promise as a tool for debunking, critical thinking education, and understanding cognitive vulnerabilities. The core message is that confronting these challenges requires a conscious effort to overcome cognitive biases, robust media literacy, and the ethical, interdisciplinary application of AI.

Nyckelord: Repetition, kognitiva bias, artificiell intelligens, informationspåverkan

Ordlista

AI	Teknik som gör det möjligt för datorer att utföra uppgifter som normalt kräver mänsklig intelligens, som exempelvis att förstå språk, känna igen mönster eller fatta beslut.
Generativa språkmodeller	AI-system som kan skapa nytt innehåll, som texter, bilder, videoklipp eller ljud, utifrån enkla instruktioner eller data.
Desinformation	Felaktig eller manipulerad information som sprids med flit för att skada en person, organisation eller ett land. Till exempel skapandet och delandet av påhittade nyheter.
Misinformation	Falsk information som inte nödvändigtvis har skapats för att orsaka skada. Det kan handla om missförstånd, eller att någon läser en artikel på nätet, tror att det är fakta – och delar för att informera sina vänner, utan att vara medveten om att det som delas är falskt.
Propaganda	Budskap som sprids i syfte att påverka individers eller organisationers attityder och handlande till förmån för avsändaren. Propaganda bygger ofta på desinformation och vilseledning och är också ett vapen som används i krig.
Narrativ	En sammanhängande ram som ordnar fakta, händelser och tolkningar till en meningsfull helhet och därigenom formar hur mottagare förstår och värderar berättelsen.
Otillbörlig informationspåverkan	Otillbörlig informationspåverkan är när främmande makt eller andra yttre hotaktörer på ett skadligt sätt försöker påverka, störa och styra det offentliga samtalet i Sverige. Några av medlen i informationspåverkan är till exempel desinformation, vilseledning och propaganda.
Motståndskraft	Motståndskraft är samhällets och individernas gemensamma förmåga att stå emot, hantera och återhämta sig från kriser, yttre påtryckningar och otillbörlig informationspåverkan, så att våra grundläggande funktioner, den fria åsiktsbildningen och handlingsfriheten bevaras.
Psykologiskt försvar	Det psykologiska försvaret är samhällets gemensamma förmåga att upptäcka och stå emot otillbörlig informationspåverkan som riktas mot Sverige från främmande makt.
Ekokammare	En informationsmiljö där likasinnade personer upprepar och förstärker samma budskap, medan motstridiga perspektiv filtreras bort.
Konspirationsteori	En föreställning att en händelse eller situation ytterst beror på en hemlig, samordnad komplott av mäktiga aktörer som verkar i det fördolda.

Inledning

I dagens samhälle där informationsflödet är konstant och ofta överväldigande, uppstår naturliga sårbarheter som kan utnyttjas av motståndare med strategiska avsikter. Genom psykologisk påverkan – där desinformation, rykten och riktad propaganda används – kan en aktör undergräva tillit, splittra opinioner och skapa oro hos befolkningen. Dessa angrepp sker inte sällan i gråzonen mellan fred och konflikt, där effekterna är svåra att upptäcka innan de fått fäste. Med anledning av detta är det centralt att identifiera hur dessa hot manifesteras och hur de kan motverkas, för att skydda både individens frihet och samhällets motståndskraft i stort.

Sedan urminnes tider har människor försökt förstå hur vi lär oss, vad vi tror på och varför vi tror på det? Under detta sökande har folkliga visdomsord uppstått som fångar observationer om lärande och tro. Två ofta citerade talesätt belyser repetitionens dubbla roll: *”Repetition är kunskapens moder”* som betonar att upprepad exponering för korrekt information främjar inlärning, medan *”Om man upprepar en lögn tillräckligt ofta blir den sann”* varnar för att även felaktigheter kan upplevas som sanna om de upprepas. Dessa uttryck pekar på en central paradox i mänsklig kognition – att upprepning både kan befästa genuin kunskap och cementera missuppfattningar. En historisk koppling kan göras till berättelsen om Janus, och janusansiktet. Begreppet janusansikte har sitt ursprung i den romerska mytologin och bygger på guden Janus, en central gestalt i den romerska religionen. Janus var gud för övergångar, förändring, tidens gång samt dörrar och portar (Beard et al., 1998). Det som särskilde Janus från andra gudar var hans två ansikten: ett vänt mot det förflutna och ett mot framtiden. Denna avbildning gör honom till en symbol för dualitet och övergång. Janus gav namn åt januari, årets första månad, eftersom han symboliserar skiftet mellan det gamla och det nya året. I nutida bemärkelse har janusansikte kommit att symbolisera dubbelhet, och används ofta för att beskriva en person som visar olika sidor beroende på situation, men även en symbol för människans dubbla natur och komplexitet.

Artificiell intelligens (AI) kan tjäna som ett verktyg till janusansiktet genom att bidra till att producera, förstärka och sprida innehåll i antagonistiskt syfte, likväl som att AI kan användas för goda och källkritiska ändamål. Det blir därför viktigare än någonsin att förstå de kognitiva bias som kan göra oss som individer, och i förlängningen vårt samhälle, sårbara för hur vi tolkar potentiellt felaktig information och hur vi kan använda oss av dessa insikter i kombination med AI för att stärka vår motståndskraft.

Kognitiva bias/snedvridningar

I denna rapport används oftare termen kognitiva snedvridningar istället för det engelska cognitive bias. Detta val speglar en strävan efter språklig tydlighet och precision i en svensk kontext, där begreppet bättre beskriver de systematiska och ofta omedvetna avvikelser från rationellt tänkande som påverkar människors bedömningar och beslut.

Trots goda intentioner är människans tänkande långt ifrån objektivt. Vi tenderar att fastna i kognitiva genvägar, mentala tumregler som sparar tid och energi, men som också öppnar upp för feltolkningar. I föreliggande rapport diskuterar vi kognitiva bias som relaterar till syftet med rapporten. Det finns betydligt fler snedvridningar, men vi har valt att använda nedanstående med hänsyn till studiens syfte. En av de mest studerade kognitiva snedvridningarna är den *illusoriska sanningseffekten*, först dokumenterad av Hasher et al., (1977). Effekten innebär att vi tenderar att betrakta upprepade påståenden som mer trovärdiga, inte för att de är bättre underbyggda utan för att de känns bekanta. Det spelar ingen roll om påståendet är sant eller falskt. Pennycook et al., (2018) visade hur denna effekt bidrar till att falska nyheter kan spridas online medan Fazio et al., (2015) fann att inte ens förkunskap skyddar oss. Vi kan känna igen en lögn, veta att den är falsk men ändå börja tro på den efter upprepning.

En annan kognitiv snedvridning är *bekräftelsebias*, vår benägenhet att söka, tolka och minnas information som bekräftar våra redan existerande uppfattningar. En studie genomförd av Wason (1960) visade hur testdeltagare systematiskt ignorerade information som kunde motbevisa deras egna hypoteser. Senare forskning av Lord et al., (1979) och Nickerson (1998) bekräftar att denna tendens är både djup och svår att bryta, och att den påverkar hur vi tolkar ny information även när vi konfronteras med motbevis. Bekräftelsebias fungerar ofta omedvetet och drivs av hjärnans strävan efter kognitiv effektivitet. Det är helt enkelt lättare att ta in det som passar in i det vi redan tror. Utöver detta finns ett besläktat men mer emotionellt färgat fenomen: *motiverat tänkande* (motivated reasoning). Här handlar det inte bara om att föredra bekräftande information, utan om att aktivt anpassa sitt tänkande för att skydda sina önskningar, värderingar eller sin självbild. Kunda (1990) visade hur vi inte bara tolkar information selektivt, utan också omformar våra slutsatser för att passa det vi vill tro på. Detta hänger nära samman med Festingers (1957) teori om kognitiv dissonans, där vi försöker minska det obehag som uppstår när vi ställs inför information som strider mot våra uppfattningar. Istället för att ompröva våra ståndpunkter tenderar vi att omtolka fakta, bortförklara motargument eller helt enkelt söka tröst i våra tidigare övertygelser.

Medan bekräftelsebias kan ses som en kognitiv genväg, är motiverat tänkande en psykologisk försvarsmekanism där båda gör oss sårbara, men på olika sätt. Ibland har vi kunskapen, men använder den inte. Det kallas *kunskapsförsumlighet* (knowledge neglect), ett begrepp som fångar hur vi tenderar att förbise relevant information, särskilt när vi är fokuserade på att snabbt förstå eller reagera snarare än att tänka kritiskt. Kahneman och Tversky (1973) visade hur vi ofta litar på den information som är lättast tillgänglig, snarare än den som är mest relevant. Marsh och Umanath (2014) har i sin tur visat hur svårt det kan vara att återaktivera korrekt kunskap, även när den faktiskt finns i minnet. Kognitiva bias är alltså systematiska mönster i hur människor tänker, som leder till att vi ibland gör bedömningar och fattar beslut på ett irrationellt eller förvrängt sätt. Dessa tankefällor uppstår ofta som en följd av begränsningar i vår mentala kapacitet. Vi förenklar för att orka tänka, men påverkas också av känslor, tidigare erfarenheter och förutfattade meningar. Bias kan prägla både vardagliga beslut och komplexa omdömen inom politik, ekonomi, vetenskap och utbildning. I en tid av intensiva informationsflöden utgör kognitiva bias därför inte bara personliga svagheter, utan även sårbarheter som kan exploateras av antagonistiska aktörer som vill genomföra och sprida otillbörlig informationspåverkan.

Artificiell intelligens (AI)

I en värld där teknologiska framsteg snabbt omvandlar vårt sätt att tänka och interagera, har AI blivit ett centralt verktyg inom många områden. Från att förbättra beslutsfattande till att optimera arbetsprocesser, har AI potentialen att förändra samhället på djupet. Men samtidigt som vi omfamnar dessa framsteg, uppstår viktiga frågor om de risker som följer med AI:s användning, särskilt när det gäller farorna med repetitiva mönster och de kognitiva snedvridningar som kan påverka både maskiner och människor. Genom att förstå dessa risker kan vi bättre navigera de etiska och praktiska utmaningarna som följer med den allt mer automatiserade världen. Repetition, både i algoritmer och mänskligt tänkande, kan skapa ett ekosystem där tidigare misstag förstärks istället för att åtgärdas. Kognitiva snedvridningar, både hos AI och i våra egna beslut, kan leda till allvarliga feltolkningar som ytterst påverkar vårt beslutsfattande. AI kan också förstärka förutfattade meningar genom de rekommendationssystem som styr mycket av hur vi konsumerar information i de digitala miljöerna. Dessa algoritmer (drivna av maskininlärning på användardata) är programmerade för att öka engagemanget och tiden som vi spenderar på plattformen, vilket ofta innebär att de visar oss sådant vi troligen gillar eller håller med om. Konsekvensen blir en automatiserad bekräftelsebias för att säkerställa att vi aldrig lämnar flödet. Vi matas med likartade perspektiv som de vi redan interagerat med, vilket ytterligare bekräftar våra åsikter. De digitala plattformarnas design kan alltså leda till att felaktigheter eller ensidiga narrativ fortsätter dyka upp i vårt flöde om vi en gång visat intresse för dem. Detta märks exempelvis i YouTubes och TikToks rekommendationer såväl som i Facebooks nyhetsflöde. När någon tittar på ett konspiratoriskt videoklipp, rekommenderas snart fler i samma anda. Det är en indirekt form av AI-förstärkt repetition. Algoritmen ser till att du påminns om och återkommer till samma budskap gång på gång. Den sociala bevisföringen (likes, delningar) synliggör dessutom att ”många andra” verkar gilla innehållet, vilket ökar den illusoriska sanningseffekten, starkare bekräftelsebias och större selektiv exponering och motiverat tänkande.

Syfte

Syftet med föreliggande studie är att beskriva hur repetition kan påverka inlärning och attityder, samt att koppla de inledningsvis nämnda talesätten till illusorisk sanningseffekt, bekräftelsebias och relaterade kognitiva snedvridningar. Vilken roll kan AI och framförallt generativa språkmodeller spela i att förstärka eller motverka dessa fenomen i pedagogiska sammanhang? Rapporten är strukturerad enligt följande. Först följer ett teoretiskt ramverk kring repetition och kognitiva snedvridningar, därefter en analys av repetitionens makt i praktiken, med exemplifiering av hur falska uppgifter kan få fäste och hur detta kan motverkas. Slutligen behandlas AI:s dubbla roll i informationslandskapet innan en avslutande diskussion.

Repetition och inlärningens mekanismer – teoretiskt ramverk

Talesättet *”Repetition är kunskapens moder”* sammanfattar en grundläggande princip inom pedagogik och kognitionsvetenskap, upprepad exponering för information underlättar inlärning och minnesbildning. Genom repetition förstärks minnesspår, vilket gör det lättare att återkalla information vid ett senare tillfälle. Klassisk experimentell forskning av Ebbinghaus (1885 i Ruger & Bussenius 1913) har visat att övning ger färdighet. Ju fler gånger vi repeterar fakta eller färdighet, desto mer robust blir vår kunskap. Modern forskning (Cepeda et al., 2006 och Roediger & Karpicke, 2006) bekräftar att schemalagda repetitioner (t.ex. utspädd repetition eller spaced repetition) motverkar glömska och förbättrar långtidsminnet. Detta är den positiva sidan av repetitionens makt, den hjälper oss att lära och bevara korrekta uppgifter. En student som övar multiplikationstabellen om och om igen, eller en pilot som simulerar flygsituationer upprepade gånger, bygger upp en djupare och mer lättillgänglig kunskapsbas. Repetition leder också till ökad familjaritet, vilket gör informationen mer lättillgänglig och flytande att bearbeta kognitivt, något som ofta korrelerar med upplevd säkerhet i det man lärt sig.

Det är dock viktigt att poängtera att repetition i sig inte gör en uppgift sann eller korrekt; den förbättrar bara vår förmåga att komma ihåg och använda informationen. Om det vi repeterar är felaktigt riskerar vi att befästa misstag istället för korrekt kunskap. Här överlappar repetitionens domän med området för kognitiva snedvridningar – särskilt illusorisk sanningseffekt – som belyser hur familjaritet kan misstas för sanning. Med andra ord, upprepad information kan kännas sann för att den är bekant, även om den i realiteten är falsk. Detta leder oss in på de kognitiva bias som är relevanta för hur vi tar till oss upprepad information.

Illusorisk sanningseffekt

Illusorisk sanningseffekt innebär att upprepade påståenden tenderar att uppfattas som mer trovärdiga än påståenden vi hör för första gången, oavsett det faktiska sanningsvärdet. Forskning har visat att en lögn som upprepas tillräckligt ofta kan uppfattas som ”sanning” av mottagaren, precis som det folkliga talesättet varnar för. Detta fenomen observerades vetenskapligt för första gången på 1970-talet i studier där testdeltagare fick utvärdera triviala fakta. Fraser som hade presenterats tidigare bedömdes som mer korrekta än nya fraser, trots att deltagarna inte fick någon ny evidens för riktigheten. Hasher et al., (1977) var bland de första att dokumentera denna effekt, och senare studier (Pennycook, Cannon & Rand, 2018 och Fazio, Rand, & Pennycook, 2019) har visat att detta även gäller för mer komplex information såsom politiska påståenden och nyhetsrubriker.

Den centrala mekanismen bakom illusorisk sanningseffekt är familjaritet och kognitiv lätthet. När vi hör eller ser något om och om igen blir bearbetningen i hjärnan smidigare. Vi känner igen det, vilket ger ett intryck av att ”detta har jag hört förut, alltså måste det ligga något i det”. I grunden är upprepning en praktisk strategi som hjärnan använder för att spara energi. Genom att automatisera kunskap, som att kunna alfabetet eller cykla, behöver vi inte aktivt bearbeta samma information varje gång vi möter den. Denna automatisering är en evolutionär överlevnadsmekanism. Världen är komplex, och utan mentala genvägar skulle vi snabbt bli överväldigade av informationsflödet. Denna känsla av bekanthet kan dock också misstas för en indikation på sanning, särskilt om vi inte aktivt minns varifrån informationen kommer eller om vi uppfattade den som pålitlig från början. I själva verket kan vi falla offer för det som Kahneman (bl.a. 2003) kallar ”System 1-tänkande”; snabbt, automatiskt och associativt – där vi förlitar oss på mentala genvägar som ”det som är bekant är sant”. ”System 2-tänkande”, det långsammare, mer analytiska tänkandet som skulle kunna ifrågasätta påståendet, aktiveras inte alltid spontant när informationen redan känns sann.

Illusorisk sanningseffekt drabbar över både åldrar och kunskapsnivåer. Fazio et al., (2015) visade att inte ens förkunskaper skyddar mot effekten. I studien såg man att deltagare som tidigare känt till det rätta svaret ändå började uppfatta felaktiga påståenden som mer sanna efter upprepad exponering. Effekten tyder på att känslan av bekantskap som repetition skapar ibland trumfar vårt logiska minne, även när vi egentligen vet bättre. I den bemärkelsen kan man säga att kunskap inte alltid skyddar mot upprepade lögner. Till och med triviala fel, exempelvis att blanda ihop bibliska gestalter, kan leta sig in i minnet om de presenteras i förbigående och sedan upprepas. Ett klassiskt exempel är experimentet om Mosesillusionen (Erickson & Mattson, 1981). När försökspersoner fick frågan *”Hur många djur av varje slag tog Moses med sig på arken?”* svarade de flesta ”två”, trots att de egentligen visste att det var Noa som enligt berättelsen byggde arken. Efteråt, när samma personer fick frågan *”Vem var det som byggde och seglade arken?”* svarade många felaktigt ”Moses”, vilket tyder på att de hade tagit till sig den felaktiga uppgiften genom den initiala exponeringen. Den felaktiga informationen blev bekant och integrerades i deras minne – ett exempel på hur repetition (eller i detta fall bara en enda ouppmärksammad exponering, uppföljd av frågor som stärkte det) kan plantera en oriktig uppfattning. Vidare tycks repetition ge störst effekt initialt, ökningen i upplevd sanning är starkast under de första exponeringarna, varefter kurvan planar ut (Hassan & Barber, 2021). Denna insikt

har praktisk betydelse. Det antyder att om otillbörlig informationspåverkan kan bemötas tidigt, innan den hunnit upprepas många gånger, kan man förhindra att felaktiga idéer får fäste. Om en lögn redan har ekat genom samhället ett dussintal gånger kan det däremot vara betydligt svårare att rubba den etablerade tilltron.

Sammanfattningsvis visar illusorisk sanningseffekt att repetition i sig själv kan vilseleda vår sanningsbedömning. Den utgör den psykologiska förklaringen bakom talesättet om den upprepade lögnen som *”blir sann”*. I praktiken blir givetvis inte lögnen bokstavligen sann, men för människors upplevelse kan den kännas sann, vilket ibland är nog för att få reella konsekvenser (t.ex. om väljare tror på en falsk nyhet och ändrar sitt beteende utifrån den). Illusorisk sanningseffekt är en av de mest centrala kognitiva snedvridningarna i dagens informationsmiljö, särskilt i samband med otillbörlig informationspåverkan och spridning av propaganda.

Bekräftelsebias och selektiv exponering

En annan nyckelfaktor som påverkar hur vi tar till oss upprepad information är bekräftelsebias; tendensen att söka upp, tolka och minnas information på ett sätt som bekräftar våra redan existerande föreställningar eller hypoteser. Peters (2022) beskriver hur bekräftelse är den troligtvis mest diskuterade och beforskade kognitiva snedvridningen.

Bekräftelsebias gör att vi söker upp och litar mer på information som bekräftar våra åsikter, samtidigt som vi bortser från eller nedvärderar sådant som säger emot dem. Detta fenomen interagerar med repetition på så vis att människor ofta väljer att repetera (eller exponeras för) information som stöder deras egna ståndpunkter. Det kan ske antingen frivilligt eller genom sin sociala omgivning och media (Wason, 1960, Lord et al., 1979 och Nickerson, 1998). I en värld av ökande informationsmängd kan vi inte ta in all information, så vi filtrerar. Bekräftelsebias verkar som ett omedvetet psykologiskt filter: vi klickar oftare på nyheter som låter som något vi redan håller med om, vi umgås gärna med likasinnade och formar så kallade ekokammare både off- och online. I dessa ekokammare upprepas liknande ståndpunkter om och om igen och motsägande perspektiv trängs bort. Den som redan tror på en viss konspirationsteori tenderar exempelvis att söka sig till forum och grupper där just den teorin diskuteras och bekräftas dagligen, snarare än att aktivt exponera sig för motargument (Garrett, 2009). Således ser vi hur bekräftelsebias kan förstärka repetitionens effekt, inte bara upprepas information, utan dessutom filtreras den så att man bara hör den version som bekräftar den egna världsbilden. Detta skapar en självförstärkande loop: tro leder till selektiv exponering, som leder till mer repetition av det man redan tror, vilket ytterligare förstärker tron.

Människor generaliserar så gärna att de ofta ’vet’ i förväg vad de kommer att gilla och vad de kommer att ogilla. De utvecklar antaganden och förväntningar, som delvis bestämmer deras framtida utvärderande reaktioner. Psykologisk forskning beskriver ett närliggande begrepp kallat biased assimilation eller snedvriden assimilering (Lord & Taylor, 2009). Det innebär att när människor konfronteras med blandade bevis – vissa som stödjer deras ståndpunkt, medan andra motsäger den – tenderar de att lägga större vikt vid stödjande bevis och kritisera eller avfärda motstridiga bevis. I praktiken kan det leda till att samma information tolkas helt olika beroende på betraktarens utgångspunkt. En studie av Lord et al., (1979) fann att personer som var för respektive emot dödsstraff, efter att ha läst samma blandade vetenskapliga rapporter, blev ännu mer övertygade om sin ursprungsposition. De hade alltså bara assimilerat det material som stödde deras tes och rationaliserat bort resten. På så sätt kan konfrontation med motstridig information paradoxalt nog leda till att felaktiga uppfattningar cementeras ännu mer, särskilt om personen ”överlever” konfrontationen genom att avfärda den.

Vid starka övertygelser, såsom konspirationsteorier, ser man ofta hur bekräftelsebias och biased assimilation resulterar i att motbevis tolkas om som ytterligare bekräftelse. Lewandowsky et al., (2013) noterade att konspirationsteoretiker kan skapa självtätande *(self-sealing)* narrativ, där varje motargument tolkas som ytterligare bevis för konspirationens existens. (”Beviset för att månfärden ägt rum är i själva verket en del av mörkläggningen” etc.). En studie av Costello et al., (2024) beskriver hur konspirationsteorier aktiverar biased assimilation där anhängare synar kritiska uppgifter mer noga än de okritiskt accepterar information som bekräftar konspirationen. Därigenom upprätthålls tron, eftersom bara den ”önskade” informationen upprepas och ges förtroende. Med andra ord, bekräftelsebias får oss att repetera det välbekanta (och önskade) budskapet för oss själva och varandra, medan vi undviker upprepad exponering för det som utmanar oss.

Detta fenomen förstärks av dagens moderna medie-strukturer. På sociala medier uppstår lätt fragmenterade informationsmiljöer där olika grupper har sina egna versioner av ”sanning” som cirkulerar internt. Algoritmer kan förstärka detta genom att rekommendera innehåll baserat på vad man tidigare interagerat med, vilket ofta innebär mer av samma slag. Ekokammare och filterbubblor i de digitala miljöerna är pådrivna manifestationer av bekräftelsebias på samhällsnivå. Enligt

Del Vicario et al., (2016) binds och isoleras likasinnade gemenskaper online, vilket underlättar spridningen av felaktigheter och försvårar spridningen av korrigerande fakta. Om en felaktig nyhet delas inom en sluten krets, riskerar den att upprepas där utan att någon utifrån ifrågasätter den, och ju mer den upprepas, desto mer befäst blir dess status som ”sanning” i den kretsen.

Sammantaget kan bekräftelsebias beskrivas som en kognitiv fälla där vi, oftast omedvetet, låter våra önskningar och förutfattade meningar styra vilket informationsflöde vi utsätter oss för. Konsekvensen blir att vi får en skev bild av verkligheten, där repetitiva budskap från den egna gruppen eller övertygelsen dominerar vårt medvetande. Repetitionens makt kombinerad med bekräftelsebias kan således skapa en illusion av konsensus och sanning. Man hör något ”överallt” (inom sin krets) och tolkar avsaknaden av motröster som att det inte finns några hållbara invändningar. Detta understryker vikten av medvetenhet om kognitiva snedvridningar och ett aktivt källkritiskt förhållningssätt för att bryta det självbekräftande repetitiva mönstret.

Kunskapsförsumlighet och motiverat tänkande

Förutom illusorisk sanningseffekt och bekräftelsebias finns ytterligare kognitiva snedvridningar som kan få oss att ta till oss felaktig information vid upprepning. En sådan är motiverat tänkande (motivated reasoning). Det är nära besläktat med bekräftelsebias och syftar på att emotionella eller ideologiska drivkrafter kan styra vårt rationella tänkande. Vi har lättare att tro på sådant vi vill ska vara sant. Motiverat tänkande kan leda till att vi omedvetet viktar evidens utifrån vad som tjänar våra syften eller värderingar, snarare än utifrån objektiv styrka. I konspirationsteorier, till exempel, finns ofta underliggande psykologiska motiv – behov av kontroll, enkla förklaringar på komplexa händelser, misstro mot auktoriteter – som gör att anhängare är motiverade att försvara teorin. Detta driver dem att resonera på ett sätt som skyddar den ursprungliga tron (t.ex. genom biased assimilation). Som Costello et al., (2024) påpekar utgör konspirationsteoretikers motstånd mot motbevis ett exempel där rationell uppdatering av föreställningar ersätts av psykologiska försvarsmekanismer. Följden blir att sakliga korrigeringar traditionellt setts som verkningslösa: ”man kan inte resonera folk ur en position som de inte har resonerat sig in i” (fritt efter Swift, citerat i Costello et al.).

En annan mekanism är kunskapsförsumlighet (knowledge neglect), som berördes ovan i samband med Mosesillusionen (Erickson & Mattson, 1981). Kunskapsförsumlighet innebär att människor ofta misslyckas med att utnyttja relevant förkunskap i stunden när de tar emot ny information, särskilt om fokus ligger på att förstå eller reagera snarare än att kritiskt utvärdera. Vår hjärna prioriterar ofta snabb förståelse över noggrann faktakoll. Som följd kan vi råka ”lära in” ny information som är falsk, trots att vi egentligen vet bättre, bara för att vi inte kopplade in rätt kunskap i rätt ögonblick. Om den felaktiga informationen sedan dyker upp igen (repetition), finns risken att vi minns den som om den vore sann eftersom vi redan inkorporerat den. Vellani et al., (2023) visade hur försöksdeltagare var mer benägna att dela påståenden de tidigare utsatts för, och förhållandet mellan upprepning och delande medierades av upplevd sanning. Så, upprepning av desinformation snedvred bedömningar av påståenden och gjorde att deltagarna valde att sprida desinformation. Kunskapsförsumlighet kan ses som upptakten. Den första exponeringen där försvaret brister, och illusorisk sanningseffekt som uppföljaren; mekanismen som vid upprepning fördjupar misstaget.

Den teoretiska insikten är att repetitionens makt över våra sinnen är stor men inte ödesbestämd. Genom utbildning i källkritik och faktagranskning kan individer bli bättre rustade att känna igen och stå emot kognitiva snedvridningar. Samtidigt utgör den stora och snabba informationsmiljön en utmaning för förmågan att upprätthålla ett kritiskt förhållningssätt. Repetition och kognitiva bias påverkar hur människor tar till sig information inom områden som utbildning, nyhetskonsumtion, propaganda och sociala medier, och AI-tekniker spelar en allt större roll i denna dynamik.

Lärande och repetition

I både pedagogiska sammanhang och vardagligt lärande är de positiva effekterna av repetition tydligt dokumenterade. Studenter förbättrar sina färdigheter genom övning och återkommande exponering: lingvister repeterar glosor, musiker övar skalor, och vetenskapliga begrepp återkommer i olika former för att befästas. En särskilt effektiv metod är distribuerad repetition; att sprida ut övningstillfällena över tid, vilket bevisligen förbättrar långtidsminnet genom att underlätta överföringen av information från korttids- till långtidsminnet (Cepeda et al., 2006).

Utöver repetition i sig har även upprepad testning, ofta kallad testeffekten, visat sig vara en kraftfull strategi för att fördjupa lärande. Istället för att enbart läsa om ett ämne har forskare visat att den aktiva återkallelsen av information – att tvinga sig själv att minnas – skapar starkare och mer beständiga minnesspår (Roediger & Karpicke, 2006). En omfattande forskningsöversikt av Dunlosky et al., (2013) rankar just dessa två tekniker, distribuerad repetition och testeffekten, som bland de mest effektiva och evidensbaserade verktygen för

varaktig inlärning. Tillsammans ger dessa resultat tydligt empiriskt stöd för idén att repetition inte bara befäster kunskap, den bygger den. Dock måste repetition inom utbildning vara meningsfull. Ren mekanisk upprepning utan förståelse riskerar att leda till ytinlärning. Effektiv repetition innebär ofta någon form av variation eller aktiv bearbetning, t.ex. att tillämpa ett koncept i olika problem, inte bara läsa samma text om och om igen (Bjork & Bjork, 2011). När repetition kombineras med reflektion och återkoppling, bygger det upp förståelse och kritisk förmåga, inte bara memorering (Dunlosky et al., 2013). På så sätt skiljer sig repetition av sanna, välgrundade fakta i ett lärandesyfte från repetition av missvisande påståenden som vi snart ska diskutera. I en lärandesituation finns (idealiskt) också en lärare eller källa som garanterar att det som repeteras är korrekt, vilket gör att eleven med fog kan stärka sin tillit till informationen genom upprepning.

Även inom forskarsamhället och genom spridning av forskningsrön kan man se att upprepad kommunikation av samma budskap över tid behövs för att det ska slå rot (Lewandowsky et al., 2012). Till exempel har folkhälsoinformation (som riktlinjer kring kost, motion eller vaccin) ofta behövt upprepas genom olika kanaler och under lång tid för att allmänhetens kunskap och beteenden ska påverkas. Här hoppas man att kunskapsaspekten av repetition ska hjälpa sann information att segra över myter och missförstånd. I bästa fall kan upprepad korrekt information tränga ut felaktigheter, särskilt om det felaktiga från början varit mindre spritt. Men när felaktiga uppfattningar redan fått fäste hos människor krävs mer än bara upprepning av fakta. Då behövs även strategier för att hantera det motstånd som bekräftelsebias och motiverat tänkande kan ge upphov till (Hornsey et al., 2018).

Sammanfattningsvis representerar ”repetition är kunskapens moder” repetitionens konstruktiva kraft. Genom att ständigt återkomma till sann och användbar information kan vi förstärka kunskap, färdighet och kompetens. Men denna princip vilar på antagandet att innehållet är verifierat och att mottagaren har intentionen att lära sig. Utanför kontrollerade inlärningsmiljöer, i ett öppet informationsflöde, kan repetition få mer problematiska konsekvenser, vilket talesättet om den upprepade lögnen belyser.

Propaganda och desinformation

Propagandaapparater och antagonistiska aktörer som sprider falska eller vinklade budskap har länge utnyttjat repetition som ett verktyg. Under 1900-talets konflikter med totalitära regimer blev det särskilt tydligt. Historiska och moderna exempel visar att strategisk upprepning av slagord, stereotyper eller rena lögner kan förändra människors uppfattningar i stor skala. Den psykologiska grunden till detta har vi redan identifierat som illusorisk sanningseffekt. När befolkningen matas med samma budskap via radio, tidningar, affischer – eller idag via sociala medier, nyhetskanaler, bloggar – etableras en känsla av att *”alla säger det, då måste det ligga något i det”*. Även lögner kan normaliseras genom ständig upprepning. American Psychological Association (APA 2023) framhåller i en rapport risken med att upprepa felaktiga påståenden, då det riskerar att öka mottagarens tro på det. APA understryker att detta gäller för människor i alla åldrar och kunskapsnivåer, och att till och med media och kända personer som omedvetet sprider en felaktighet kan bidra till att cementera den i folks föreställningsvärld. Av detta skäl rekommenderas att aldrig repetera felaktig information utan att direkt länka det med en tydlig korrigering. Om man till exempel ska bemöta ett falskt rykte, bör man undvika att framföra ryktet om och om igen (eftersom repetition riskerar förstärka det), utan istället lägga fokus på den korrekta informationen, annars kan motverkansförsöket ge bakslag.

I dagens digitala landskap kan en lögn spridas och upprepas snabbare och bredare än någonsin förr. En falsk nyhet kan dyka upp på en obskyr blogg, plockas upp av tusentals användare på X, delas vidare till Facebookgrupper, och inom timmar ha nått miljontals ögon. Varje delning är en form av repetition som legitimerar budskapet i andras ögon. Studier (se exempelvis Center for Countering Digital Hate, 2021; Grinberg et al., 2019; DeVerna et al., 2024) visar att en liten andel användare, så kallade ”superspridare”, står för en oproportionerligt stor andel av delningarna av desinformation. Dessa superspridare, ibland automatiserade botar eller trollfabriker, ser till att meddelandet upprepas konstant i flödet. I ekokammare fortplantas sedan lögnen utan att granskas kritiskt.

När falska narrativ upprepas tillräckligt ofta etableras de ibland som felaktiga konsensusuppfattningar. Folk tror att *”många andra tror på detta, alltså kan det inte vara helt fel”*. Detta kan leda till kollektiva missuppfattningar som är svåra att rucka. Ett exempel är myten om att vacciner orsakar autism, som uppstod ur en diskrediterad studie men som via upprepad spridning i vissa grupper blivit en seglivad ”sanning” för vaccinskeptiker. Trots överväldigande vetenskaplig evidens emot, lever myten kvar genom ständiga repetitioner i vissa forum och av vissa aktivister, kombinerat med anekdoter och känslomässiga berättelser som förstärker mottagarens minne av budskapet.

Desinformationens psykologi har därför blivit ett kritiskt forskningsfält. Hur kan man bryta effekten

av en lögn som upprepas? Enligt forskning från bland annat McGuire (1964), van der Linden et al., (2020), Ecker et al., (2014), Lewandowsky et al., (2012) är förhandsgranskning (prebunking) och faktagranskning (debunking) två strategier som används:

- Prebunking innebär att man förbereder människor på att de kan komma att möta viss felaktig information, och lär dem att känna igen retoriska knep eller felaktigheter innan de exponeras i stor skala. Genom att ”vaccinera” befolkningen kognitivt (inoculation theory) kan man minska effekten av senare upprepade lögner.
- Debunking innebär korrigering i efterhand: man bemöter en felaktighet som fått spridning genom att presentera motbevis och förklaring till varför originalpåståendet är fel. Effektiv debunking kräver enligt forskning att man inte enbart bestrider den falska utsagan, utan ersätter den i minnet med en alternativ förklaring. Viktigt är som sagt att inte förstärka myten genom överdriven upprepning av den, utan hellre fokusera på sanningen.

Repetition kan alltså användas även i korrigeringssyfte. APA (2023) rekommenderar att motbevis och korrekt information bör spridas ofta och upprepade gånger för att neutralisera misstag. Upprepning kan således beskrivas som ett tveeggat svärd. Samma kraft som etablerar lögnen kan användas för att etablera sanningen, förutsatt att budskapet når fram och accepteras. Men hur når man fram om människor har bekräftelsebias och redan misstror information från ”etablissemanget”? Här uppstår behovet av förtroende och nya angreppssätt, och en möjlig spelare som kan påverka utvecklingen är AI.

Risker i det digitala ekosystemet

Den tekniska utvecklingen har gett oss verktyg som kan sprida information med en hastighet och precision som tidigare generationer inte kunnat föreställa sig. Idag kan människor i många fall inte längre avgöra om en text är skriven av en AI eller en människa, och ofta bedöms de artificiellt skrivna texterna som bättre (Porter & Machery, 2024, Zhang & Gosline 2023). AI genomsyrar numera många aspekter av informationsflödet, från de algoritmer som styr våra sociala medieflöden till avancerade språkmodeller som kan producera texter, bilder samt ljud- och videoklipp. Dessa system kan, beroende på hur de används, antingen förstärka eller försvaga effekterna av de kognitiva snedvridningar vi har diskuterat.

På risksidan har flera studier pekat på hur AI kan förstärka otillbörlig informationspåverkan och propaganda. Goldstein et al., (2023) visade att stora språkmodeller (GPT-3) kan generera övertygande propagandatexter som får läsare att instämma med budskapet i lika hög grad som texter skrivna av verkliga propagandister. Bara genom att justera frågeställningen (prompten) och göra mindre efterredigering kunde forskarna få AI-texter att vara lika effektiva i påverkanssyfte som mänskliga motsvarigheter. Detta sänker tröskeln för antagonistiska aktörer att massproducera övertygande påverkan till en låg kostnad och ansträngning. Liknande resultat har även rapporterats av Spitale et al., (2023) medan Danry et al., (2024) visade att språkmodeller som skriver trovärdiga, men falska förklaringar, kraftigt kan förstärka desinformation. Vykopal et al., (2023) visade att språkmodellers artighet, struktur och saklighet kan förstärka trovärdigheten hos falska påståenden, även när innehållet är felaktigt. Detta tyder på att det inte bara är upprepning eller innehåll som påverkar, utan även tonläge och stil samt att det är betydligt enklare att korrigera detta för en språkmodell än för en mänsklig debattör.

Ett kontroversiellt experiment från forskare vid Zürichs universitet (2025) visar att AI kan vara mer övertygande än människor i digitala miljöer under verkliga förhållanden. I studien publicerade forskarna kommentarer skrivna av språkmodeller (GPT-4o, Claude 3.5 Sonnet och Llama 3.1) som svar på inlägg på forumet Reddit, utan att användarna visste att texterna var AI-genererade. Resultatet visade att AI-kommentarerna påverkade användarnas åsikter tre till sex gånger mer än den genomsnittliga mänskliga motsvarigheten. Likaså undersökte Bashardoust et al., (2024) hur människor reagerar på falska AI-genererade nyheter jämfört med falska nyheter skrivna av människor. Deras experiment visade att även om deltagarna ansåg att AI-genererade falska nyheter var något mindre trovärdiga, var de ändå lika villiga att dela dem vidare. Med andra ord, texter med sämre kvalité stoppade inte spridningen. Deltagarna kunde tänka sig att sprida en AI-skapad falsk nyhet på sociala medier i lika stor utsträckning som en falsk nyhet skriven av människor. AI-lögner kan således spridas viralt nästan på samma sätt som vanliga lögner, även om folk är medvetna om AI:s inblandning.

En tidigare begränsning har varit att det krävs tid, språkkunskap och skicklighet att skriva övertygande propaganda. AI har nu tagit bort den barriären. Studien från Goldstein et al., (2023) utgår också från en äldre språkmodell. Dagens modeller är betydligt mer potenta och kan skriva texter som är avsevärt mycket mer polariserande och övertygande. Hotaktörer som Ryssland och Kina har redan börjat använda sig av stora språkmodeller för att massproducera, skala upp sina verksamheter och sprida material för otillbörlig informationspåverkan på såväl sociala medier som falska nyhetskällor (OpenAI 2025, Meta 2024 och Alphabet 2025).

Intressant nog visar forskning att även språkmodeller själva kan påverkas av övertygande falska narrativ i en dialog med en mänsklig användare. I en studie av Xu et al., (2024) kunde språkmodeller (GPT-4) övertygas om felaktiga svar genom en strukturerad samtalssituation, vilket tyder på att AI inte enbart är sårbart för påverkan genom sin träningsdata, utan även kan manipuleras i realtid genom dialog.

Ytterligare en dimension av AI är möjligheten till målgruppsanpassning, ibland ner på individnivå. Med stora datamängder om individers beteenden och preferenser kan AI skräddarsy budskap för bäst påverkan för enskilda personer. En studie av Salvi et al., (2024) demonstrerade i en kontrollerad miljö att en AI-debattör (baserad på GPT-4) som fick tillgång till persondata om motparten kunde övertyga deltagare mer framgångsrikt än mänskliga debattörer. Deltagare som debatterade mot en AI (som kände till deras bakgrund) hade 81,7 % högre sannolikhet att ändra sin inställning i språkmodellens riktning jämfört med om de debatterade mot en människa. Även utan personanpassning presterade språkmodellen bättre än människor (om än ej statistiskt signifikant i det fallet). Detta visar att AI inte bara kan sprida repetitiva budskap brett, utan också kan anpassa och finjustera dem för att bäst kunna påverka och bryta ned individers kognitiva försvar. I värsta fall skulle en antagonistisk aktör kunna använda en sådan teknik för att bryta ned motstånd genom att förse varje person med just den variant av lögnen de är mest benägna att tro på. Kombinationen av repetition och personanpassad övertalning är ett potentiellt kraftfullt och farligt verktyg. Hackenburg och Margetts (2024) visade att språkmodeller som GPT-4 är såpass övertygande i sin grundform att de inte nödvändigtvis behöver persondata för att påverka åsikter. Den generella formuleringen i sig räckte i vissa fall för att öka stödet för ett politiskt budskap med upp till tolv procentenheter.

Sammanfattningsvis finns betydande risker att AI-teknik förstärker kognitiva bias:

- Hög volym av upprepade budskap: AI kan massgenerera innehåll vilket underblåser illusorisk sanningseffekt genom ständig exponering.
- Automatisk selektion: AI-drivna flöden kan låsa in användare i en loop av bekräftande information (bekräftelsebias på steroider).
- Ökad trovärdighet: Paradoxalt nog uppfattar en del AI-genererat innehåll som objektivt; vissa deltagare kan tro att text från, en vad de upplever, ”neutral AI” är mer pålitlig än från en människa, vilket kan göra dem mindre kritiska till budskapet. Detta trots att AI-systemet faktiskt är påverkad av sin träningsdata och användarkontexten.
- Målgruppsanpassning: AI kan anpassa informationspåverkan för att exploatera just de svagheter och fördomar en viss grupp eller individ har, vilket överlistar generella motargument.

Dessa faktorer innebär att AI, utan etisk styrning (men även med etisk styrning i exempelvis diktaturer), kan bli en katalysator för spridning av falsk eller vinklad information. Kognitiva snedvridningar som illusorisk sanning och bekräftelsebias kan då utnyttjas systematiskt och i stor skala. Det är därför avgörande att förstå inte bara psykologin utan också tekniken, för att kunna möta dessa utmaningar.

Möjligheter och utmaningar

Lyckligtvis är inte AI enbart en del av problemet; det kan även vara en del av lösningen. Samma system som kan användas för att sprida informationspåverkan kan också omprogrammeras för att identifiera och motverka dem. Forskning visar också på att språkmodeller kan användas proaktivt för att bryta illusions- och bias-cykler med framgångsrika interventioner. Därutöver kan vi på samma sätt förstå egna sårbarheter och anpassa metoder för att mildra effekterna av informationspåverkan. Ett sätt är att utnyttja generativa språkmodeller som verktyg för debunking och dialog.
En studie av Costello et al., (2024) visade att personliga dialoger med en språkmodell (GPT-4) kunde minska tron på konspirationsteorier hos personer som redan var övertygade konspirationsteoretiker. I experimentet engagerades deltagare som trodde på olika konspirationsteorier i en strukturerad dialog med en språkmodell: de fick presentera sina argument, och språkmodellen svarade med motargument och fakta i tre dialogrundor. I genomsnitt reducerades tilltron i stort till konspirationsteorier med cirka 17–20% efter denna relativt korta interaktion. Över 50% av deltagarna upplevde också att de fick en något minskad tilltro till konspirationsteorier. Effekten var stor även hos inbitna troende (så kallade ”true believers”) som annars brukar vara svårast att påverka. Detta är unikt, då tidigare försök att korrigera tron på konspirationsteorier oftast har misslyckats eller gett mycket små effekter (O´Mahony et al., 2023). En uppföljande studie (Costello et al., 2025) visar dessutom att effekten kvarstod vid uppföljning två månader senare, samt att deltagarnas tro även minskade något på andra konspirationsteorier än den som diskuterats. Det tyder på en viss generalisering. Dialogen levererade inte bara specifika fakta utan den påverkade även deltagarnas sätt att värdera källor och resonera generellt. Rani et al., (2025) visade också att dialog med en språkmodell kan förbättra människors förmåga att avgöra om visuella nyheter är sanna eller falska, men effekten försvinner när det digitala stödet tas bort. Resultaten speglar hur starkt vi påverkas av

illusorisk sanningseffekt och bekräftelsebias. När AI guidar oss i realtid kan vi tillfälligt tänka klarare, men våra grundläggande kognitiva snedvridningar kvarstår. Långsiktig motståndskraft mot desinformation kräver därför inte bara tekniska hjälpmedel utan också träning, utbildning och övning i kritiskt tänkande. Slutligen bör nämnas att AI kan assistera forskare i att förstå kognitiva bias bättre. Genom att simulera mänskligt beteende eller analysera stora mängder data om spridningsmönster, kan AI hjälpa till att kartlägga hur en lögn föds, sprids, och antingen dör ut eller blir ”sanning” i offentligheten. Denna meta-nivå av förståelse kan i sin tur leda till bättre strategier för utbildning och policy.

Diskussion

Repetitionens makt över människans sätt att ta till sig information synes vara odiskutabel. Syftet med föreliggande rapport är att beskriva hur repetition kan påverka inlärning och attityder, samt att koppla de inledningsvis nämnda talesätten till illusorisk sanningseffekt, bekräftelsebias och relaterade kognitiva fällor. Vilken roll kan AI och framförallt generativa språkmodeller spela i att förstärka eller motverka dessa fenomen i pedagogiska sammanhang? Denna rapport har belyst dess dubbla ansikte: å ena sidan som en förutsättning för lärande och kunskapsbildning – repetition är kunskapens moder – och å andra sidan som en källa till vilseledning när felaktig information kan existera obesvarad – den upprepade lögnen som blir ”sann”. Genom att knyta an till psykologisk forskning om illusorisk sanningseffekt, bekräftelsebias och andra kognitiva snedvridningar har vi sett att människans kognition är benägen att förväxla det bekanta med det sanna, och att vi aktivt söker bekräftelse på det vi redan tror, vilket förstärker effekterna av upprepad information. Dessa insikter sätter fingret på varför myter, propaganda och konspirationsteorier kan få fäste även hos personer som i grunden är intelligenta och bildade. Det är inte brist på kunskap per se, utan hur hjärnan bearbetar information under begränsningar och bias som ligger bakom. Det är viktigt att inse att dessa kognitiva snedvridningar inte innebär att människor är hopplöst irrationella eller oförmögna att lära sig sanningen. Snarare belyser de var utmaningarna ligger i att vara en kritisk konsument av information. Det krävs en medveten ansträngning för att kontrollera sig själv så att man inte automatiskt håller med bara för att något är bekant (mota illusorisk sanningseffekt). Att aktivt söka upp alternativ information och motargument för att undvika ensidighet (överkomma bekräftelsebias). Att våga ta ett steg tillbaka och fråga ”vet jag redan något som motsäger detta?” vid nya påståenden (undvika kunskapsförsumlighet). Slutligen att vara uppmärksam på sina egna känslomässiga reaktioner och motiv när man bedömer information (minska motiverat tänkande).

Rapporten har också belyst AI:s snabbt växande roll i denna dynamik. AI kan förstärka kognitiva snedvridningar genom att förstärka ekokammare via algoritmer, och automatisera målgruppsanpassad påverkan samt upprepa vinklade budskap via språkmodeller. Generativa språkmodeller kan producera övertygande texter i stor skala, och forskning visar att dessa texter kan vara lika övertygande som de som skrivits av människor. Det innebär att AI-teknik bokstavligen kan förstärka effekten av repetitiva budskap genom skalbar och automatiserad spridning, och med den följer både pedagogiska möjligheter och propagandistiska risker. Samtidigt framträder också AI:s potential som en del av lösningen. Vi har sett exempel där språkmodeller i strukturerad dialog med konspirationsteoretiker kan minska tron på falska idéer, även hos så kallade true believers. Detta visar att repetitionens makt kan vändas till något gott, om den kombineras med fakta, dialog och respekt. AI-system kan användas för att skala upp den sortens debunking som tidigare varit beroende av mänsklig tid och emotionell arbetskapacitet. Exempelvis i skolan blir denna dubbelhet tydlig. AI kan där fungera som stöd för individualiserat lärande, rättvisa bedömningar och tillgänglig kunskapsförmedling – men också som källa till passivisering, automatiserade stereotyper eller ytligt kunskapssökande. Teknikens pedagogiska potential är stor, men den måste användas medvetet. Verktyg som adaptive learning och AI-bedömningar kan minska bias som halo-effekten eller recency-effekten, men om träningsdatan är sned kan nya, maskinellt förstärkta snedvridningar uppstå (VanLehn, 2011 och Holstein et al., 2019).

Vi kan därför se AI som ett janusansikte, vilket kan reproducera våra fördomar, men också hjälpa oss att genomskåda dem. Det kan upprepa lögner, men också upprepa sanningen tills den vinner gehör. Om AI integreras i undervisning, samhällsdebatt och informationsmiljöer utan kritisk reflektion riskerar det att förstärka de värsta sidorna av våra kognitiva begränsningar. Men om det implementeras med etisk medvetenhet, transparens och tvärvetenskaplig styrning, kan det bli ett verktyg för demokrati, förståelse och intellektuell motståndskraft.

I slutändan påminner resultaten oss om att människans benägenhet för upprepningens fällor är något vi kan hantera med ökad medvetenhet och teknikens hjälp, men aldrig helt eliminera. För det psykologiska försvaret blir dessa insikter värdefulla för hur man bäst kan bygga motståndskraft. Utbildning av allmänheten i källkritik är fortsatt fundamentalt. AI är ett komplement, ingen ersättning, för en vaksam och rationell motståndskraft. De initiala talesätten som presenterades i början av rapporten innehåller båda en del av sanningen. Repetition kan vara kunskapens moder, om kunskapen är äkta och repetitionen sker med eftertanke. Men repetition kan också göra lögnen starkare, om vi passivt låter den fortplanta sig utan ifrågasättande. Vår uppgift som samhälle blir att skapa förutsättningar där det förra dominerar över det senare. Där sanna budskap upprepas högt och falska budskap kvävs i sin linda. AI kommer att vara en faktor i denna ekvation. Huruvida den kraften främst gynnar sanningen eller lögnen beror på de val vi gör idag, beväpnade med insikten om hur våra sinnen fungerar.

AI:s effektivitet i påverkan bygger på dess förmåga att

framstå som objektiv, att argumentera faktabaserat och att aldrig tröttna. Så länge det finns beräkningskapacitet och energiförsörjning kan AI driva långvariga, lågintensiva påverkanskampanjer utan att drabbas av mänskliga svaghetstecken som emotionell utmattning eller uppmärksamhetsbrist.

Repetition sker nu i realtid, globalt, och algoritmer förstärker våra mänskliga avsikter antingen mot sanning eller falskhet. AI kan fungera som en superladdning på både problemet och lösningen. Om den lämnas otyglad i händerna på antagonistiska aktörer, kan AI dränka oss i trovärdigt formulerade lögner, upprepa dem i oändlig variation, och rikta dem precis dit de är som mest lockande. Det vore en dystopi där kognitiva snedvridningar systematiskt exploateras med maskinell effektivitet.

Referenser

American Psychological Association. (2023). Using psychological science to understand and fight health misinformation: An APA consensus statement. https://www.apa.org/pubs/reports/misinformation-consensus-statement.pdf

Bashardoust, A., Feuerriegel, S., & Shrestha, Y. R. (2024). Comparing the willingness to share for human-generated vs. AI-generated fake news [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2402.07395

Beard, M., North, J., & Price, S. (1998). Religions of Rome: Volume 1, A History. Cambridge University Press.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. I M. A. Gernsbacher, R. W. Pew, L. M. Hough & J. R. Pomerantz (Red.), Psychology and the real world: Essays illustrating fundamental contributions to society (s. 56–64). Worth Publishers.

Center for Countering Digital Hate. (2021). The disinformation dozen: Why platforms must act on twelve leading online anti-vaxxers. https://counterhate.com/research/the-disinformation-dozen/

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological Bulletin, 132(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. Science, 385(6714), eadq1814. https://doi.org/10.1126/science.adq1814

Costello, T. H., Pennycook, G., & Rand, D. G. (2025). Just the facts: How dialogues with AI reduce conspiracy beliefs [Preprint]. https://doi.org/10.31234/osf.io/h7n8u_v2

Danry, V., Pataranutaporn, P., Groh, M., Epstein, Z., & Maes, P. (2024). Deceptive AI systems that give explanations are more convincing than honest AI systems and can amplify belief in misinformation. arXiv. https://doi.org/10.48550/arXiv.2408.00024

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. Proceedings of the National Academy of Sciences, 113(3), 554–559. https://doi.org/10.1038/srep37825

DeVerna, M. R., Aiyappa, R., Pacheco, D., Bryden, J., & Menczer, F. (2024). Identifying and characterizing superspreaders of low-credibility content on Twitter. PLOS ONE, 19(5), e0302201. https://doi.org/10.1371/journal.pone.0302201

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public interest, 14(1), 4-58.

Ebbinghaus, H. (1913). Memory: A contribution to experimental psychology (H. A. Ruger & C. E. Bussenius, Övers.) Teachers College, Columbia University. (Originalarbete publicerat 1885)

Ecker, U. K. H., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. Journal of Experimental Psychology: Applied, 20(4), 323–335. https://doi.org/10.1037/xap0000028

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. Journal of Verbal Learning and Verbal Behavior, 20(5), 540–551. https://doi.org/10.1016/S0022-5371(81)90165-1

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. Journal of Experimental Psychology: General, 144(5), 993–1002. https://doi.org/10.1037/xge0000098

Fazio, L. K., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. Psychonomic Bulletin & Review, 26(5), 1705–1710.

Festinger, L. (1957). A theory of cognitive dissonance. Stanford University Press.

Garrett, R. K. (2009). Echo chambers online? Politically motivated selective exposure among Internet news users. Journal of Computer-Mediated Communication, 14(2), 265–285. https://doi.org/10.1111/j.1083-6101.2009.01440.x

Goldstein JA, Chao J, Grossman S, Stamos A, Tomz M. How persuasive is AI-generated propaganda? (2024) https://doi.org/10.1093/pnasnexus/pgae034

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. Science, 363(6425), 374–378. https://doi.org/10.1126/science.aau2706

Hackenburg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. Proceedings of the National Academy of Sciences, 121(24), e2403116121. https://doi.org/10.1073/pnas.2403116121

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conferment of referential validity. Journal of Verbal Learning and Verbal Behavior, 16(1), 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1

Hassan, A., & Barber, S. J. (2021). The effects of repetition frequency on the illusory truth effect. Cognitive Research: Principles and Implications, 6, 38. https://doi.org/10.1186/s41235-021-00301-5

Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018). The psychological roots of anti-vaccination attitudes: A 24-nation investigation. Health Psychology, 37(4), 307–315. https://doi.org/10.1037/hea0000586

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. American Economic Review, 93(5), 1449–1475. https://doi.org/10.1257/000282803322655392

Kahneman, D., & Tversky, A. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. Psychological Science in the Public Interest, 13(3), 106–131. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., Gignac, G. E., & Oberauer, K. (2013). The role of conspiracist ideation and worldviews in predicting rejection of science. PLOS ONE, 8(10), e75637. https://doi.org/10.1371/journal.pone.0075637

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 37(11), 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

Lord, C.G. and Taylor, C.A. (2009), Biased Assimilation: Effects of Assumptions and Expectations on the Interpretation of New Evidence. Social and Personality Psychology Compass, 3: 827-841. https://doi.org/10.1111/j.1751-9004.2009.00203.x

Marsh, E. J., & Umanath, S. (2014). Knowledge neglect: Failures to notice contradictions with stored knowledge. I D. N. Rapp & J. L. Braasch (Red.), Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences (s. 161–180). MIT Press

McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. I L. Berkowitz (Red.), Advances in experimental social psychology (Vol. 1, s. 191–229). Academic Press. https://doi.org/10.1016/S0065-2601(08)60052-0

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220.

O'Mahony C, Brassil M, Murphy G, Linehan C (2023) The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. PLoS ONE 18(4): e0280902. https://doi.org/10.1371/journal.pone.0280902

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news headlines. Journal of Experimental Psychology: General, 147(12), 1865–1880. https://doi.org/10.1037/xge0000465

Peters, U. (2022). What is the function of confirmation bias?. Erkenntnis, 87(3), 1351-1376.

Porter, B., & Machery, E. (2024). AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. Scientific Reports, 14, 26133. https://doi.org/10.1038/s41598-024-76900-1

Rani, A., Danry, V., Lippman, A., & Maes, P. (2025). Can dialogues with AI systems help humans better discern visual misinformation? arXiv. https://doi.org/10.48550/arXiv.2504.06517

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2024). On the conversational persuasiveness of large language models: A randomized controlled trial [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2403.14380

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. Science Advances, 9(26), eadh1850. https://doi.org/10.1126/sciadv.adh1850

Universität Zürich. (2025). Large-Scale Field Experiment on AI-Driven Persuasion. https://regmedia.co.uk/2025/04/29/supplied_can_ai_change_your_view.pdf

van der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. Frontiers in Psychology, 11, 566790. https://doi.org/10.3389/fpsyg.2020.566790

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

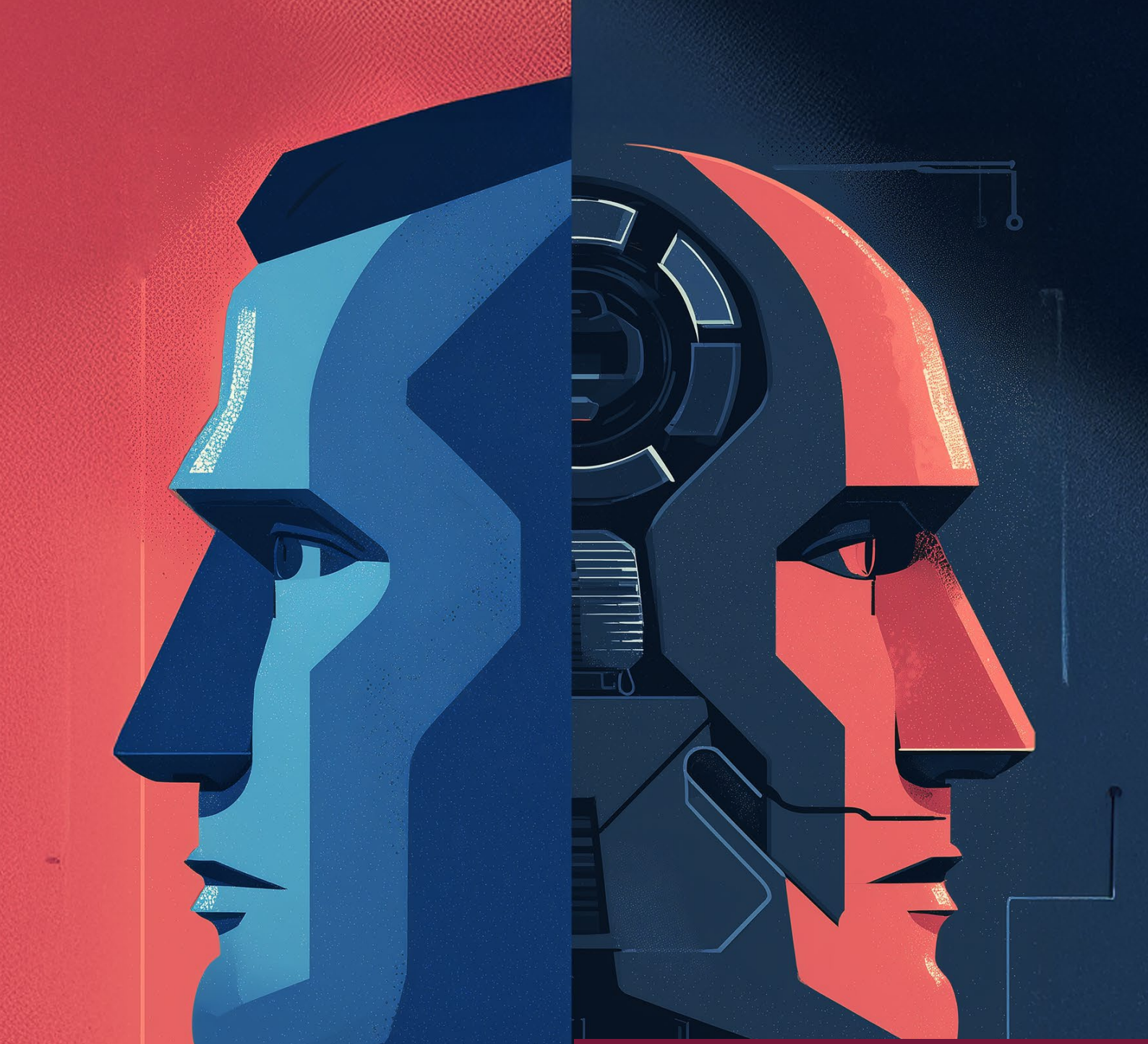
Vellani, V., Zheng, S., Ercelik, D., & Sharot, T. (2023). The illusory truth effect leads to the spread of misinformation. Cognition, 236, 105421.

Vykopal, I., Pikuliak, M., Srba, I., Moro, R., Macko, D., & Bielikova, M. (2024). Disinformation Capabilities of Large Language Models. I Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (s. 14830–14847). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.793

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology, 12(3), 129–140. https://doi.org/10.1080/17470216008416717

Xu, R., Lin, B. S., Yang, S., Zhang, T., Shi, W., Zhang, T., Fang, Z., Xu, W., & Qiu, H. (2023). The Earth is Flat because…: Investigating LLMs' Belief towards Misinformation via Persuasive Conversation. arXiv. https://doi.org/10.48550/arXiv.2312.09085

Zhang, Y., & Gosline, R. (2023). Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. Judgment and Decision Making, 18, e41. https://doi.org/10.1017/jdm.2023.37



Myndigheten för psykologiskt försvar

Myndigheten för psykologiskt försvar

Våxnäsgatan 10
653 40 Karlstad
E-post: registrator@mpf.se
Telefon: 010 - 183 70 00