Tech giants and democracy: Artificial intelligence, misinformation, and digital platforms

Mads Fuglsang Hove, University of Southern Denmark
Rebecca Adler-Nissen, University of Copenhagen
Anja Bechmann, Aarhus University
Claes H. de Vreese, University of Southern Denmark
Frederik Hjorth, University of Copenhagen
Yevgeniy Golovchenko, University of Copenhagen

Acknowledgements: Thanks to Sofia Tang from the University of Copenhagen for helping generate images for the experiment. Thanks to August Larsen for helping with editing.

This report has been translated from the original Danish report by the Ministry of Digitalisation

Preface

The end of 2024 and early 2025 have been tumultuous, with many changes in the relationship between tech giants and democracy, and in the world more generally. This report was submitted to the Ministry of Digitalisation in September 2024.

Since then, significant developments have occurred that are relevant to this report and its findings. Meta has changed its content moderation policies, increased the exposure to political content on its platforms and set in motion the abolishment of fact-checkers, starting in the United States.

The owner of X, Elon Musk, has played a key role in the presidential campaign for Donald Trump and now serves as a high-level advisor, blurring the lines between acts of public good and private interest. European – including Ukrainian – leaders raise questions about the digital services supplied by US-based companies and if critical infrastructure can be used as leverage against allies.

EU digital regulation has been criticised at the highest political level in the United States, with reporting suggesting that legislation necessary for researchers and the public to study the impact of tech giants might be used to pressure the EU in negotiations over tariffs or other matters.

While this report does not relate these developments to the scientific literature, it is evident that the research – including public dissemination of research – on the societal consequences of social media and tech giants is more critical than ever before.

Executive summary

This report explores the impact of tech giants on Danish democracy, public discourse, and social cohesion. It examines the phenomenon of misinformation and the effects of social media on political polarisation and well-being in Denmark. Overall, the report demonstrates that tech giants significantly impact democratic discourse in Denmark, potentially exacerbated by the development of generative artificial intelligence. However, there is still a lack of systematic research on tech giants, democracy, and cohesion – especially in Denmark. Prepared as part of the Media Agreement for 2023-2026, the report is the first in a series of annual reports.

Key findings

- Danes are not substantially better than chance at discerning whether images are real or created using generative AI. However, there are differences in discernment abilities among Danes, with younger people, women, those with a more analytical mindset and those who have tried generative AI before being better at assessing the authenticity of an image. Our study has implications for understanding Danes' ability to identify misinformation created by generative artificial intelligence and which social groups are most vulnerable.
- When Danes are told that misinformation is a major societal problem, they are more concerned about whether they can trust what they see online and whether misinformation affects their political choices. However, our survey also suggests that Danes can handle nuanced information without negative democratic repercussions.
- Danes are generally in favour of social media moderation and deletion of offensive, illegal, or dangerous content. Danes overwhelmingly support the existing guidelines for content moderation. However, there are indications of different cultural perceptions of what Danes believe should be deleted or preserved, and the (typically American) social media guidelines for this, for example, concerning toplessness.
- There is much evidence to suggest that the amount of misinformation on social media is more limited than we often hear. Meanwhile, there is greater uncertainty about the impact of misinformation on society more generally. This is due to the fact that misinformation can originate from a range of more or less powerful actors, be deliberetly designed to engage, and be challenging to recognise, either due to the use of sophisticated techniques or its integration into otherwise accurate information.
- We still lack knowledge about social media behaviour, especially among Danes. This applies to the extent to which Danes are exposed to and interact with misinformation and politically polarising content, as well as the amount and type of content that is problematic for the well-being of Danes, especially children and young people.

Important implications

- Lack of access to data from tech giants, including data on user behaviour, is the biggest obstacle to studying the impact of tech giants on society. The vast majority of academic disagreements and unanswered questions arise because the data needed to address the complex issues that society is most interested in is rarely available. Data access is, therefore, essential to establishing an accurate picture of the challenges of misinformation that neither overestimates nor underestimates the societal scale of the problem.

- Danish authorities and legislators should work to ensure that better opportunities for data access from e.g. the EU Digital Services Act can be used by researchers, governments, and civil society. This requires persistence and resources to help ensure access to sufficient data as well as legal support on a case-by-case basis so that economics and power imbalances do not deter actors from seeking and working with such data.
- While anyone can be manipulated by misinformation created with generative AI, certain groups in society are more vulnerable. Improving education and technology literacy among these vulnerable groups could help reduce the number of Danes struggling to navigate a world increasingly shaped by AI-generated information.
- Key players, including politicians, authorities, and the media, must adopt a nuanced approach to misinformation, avoiding both minimising and exaggerating the extent of the problem. Evidence suggests that Danes are capable of navigating a nuanced and transparent information environment, which can help prevent negative impacts on, for example, support for democracy as a form of government.

Table of contents

Foreword

Optimism about the internet has shifted significantly over time. From its inception, innovative services and communities like eBay and MySpace saw a lot of interest. Later, optimism peaked with the Arab Spring, the Occupy Wall Street movement, and the election of Barack Obama as the president of the United States. These movements highlighted the potential for democratisation, the inclusion of voices often excluded from the public debate, and fostering a more direct relationship between politicians and citizens. This created the perception that tech giants, including social media, were central to society and democracy. We all had to get on Facebook, speak our minds and expand our networks.

However, if you run fast, you risk breaking things - and tech platforms have all been running fast. Tech giants such as Facebook, Google, Instagram, Twitter, YouTube, TikTok and others have faced scandals like Cambridge Analytica, Russian interference in democratic elections, conspiracy networks and most recently, concerns about the effects on the well-being of children, young people, and adults. As a result, public sentiment about tech giants has now shifted into a valley of pessimism. These scandals have naturally prompted a societal focus on developing a solid, systematic, and scientific foundation for understanding the impact of tech giants on society. This foundation is essential to determine whether there is cause for concern and, if so, how best to address it.

This report is the first in a series produced as part of the Media Agreement for 2023-2026, offering insights into key topics concerning the influence of tech giants on Danish democracy, social cohesion, and well-being. The headline is broad, but the insights are specific. This report focuses on misinformation on social media, examining what the research reveals about it, and how Danes engage with it. We review both what is known and unknown in the research, and present data from collected from a representative sample of the Danish population through a questionnaire and two experiments, providing new insights into how Danes relate to information on social media. Additionally, the report includes brief sections on political polarisation and well-being, pointing forward to future reports.

The report contains four parts. First, it outlines some of the larger structural issues related to tech giants and the challenge againstexisting social structures. Second, it reviews the current research literature, with a particular focus on misinformation and the impact of social media on political polarisation and well-being. Third, it highlights key questions that research has yet to adequately address, along with suggestions for how these questions could be explored.. Finally, the report surveys Danes on several key issues, including their attitudes toward content moderation, their ability to identify AI-generated content, and how they are affected by the way the risks of misinformation are discussed.

The report is the product of a collaboration between researchers from three interdisciplinary research environments: Digital Democracy Centre (University of Southern Denmark), Copenhagen Centre for Social Data Science (University of Copenhagen) and DATALAB - Centre for Digital Social Research (Aarhus University).

The tech giants' impact on society

The influence of tech giants on society spans a huge area, which cannot be fully covered here. Therefore, we focus on three aspects related to why, how, and to whom misinformation is spread on social media. Specifically, we will examine trends and challenges related to the business models of tech giants, the ways social media shapes public discourse, and the impact of user numbers and ad revenue.

Tech giants is a collective term for companies that, to varying degrees, base their business model on collecting huge amounts of data about users. The tech giants use this data and pass it on to third parties, who use it for targeted advertising to optimise their own business and to retain users (Danish Ministry of Business, 2024). While the term usually includes platforms such as Amazon and Uber, in this report, we primarily use the term for platforms that play a widespread role in shaping the public debate, such as Facebook and Google.

Therefore, other important issues must be addressed elsewhere, such as democratic and political participation, algorithmic biases that reinforce existing social inequalities, the high energy consumption of technology that contradicts efforts to combat climate change, and labour conditions where vulnerable social groups risk being exploited.

The business models of tech giants

Social media has become a crucial pillar of public debate. Stories *are broken* and debates take place on social media, with entire TV programmes built around, for example, Donald Trump's latest *tweets*. Many citizens also rely on social media as a news source. For instance, roughly 50 per cent of young American TikTok users report using the platform to stay informed about politics (Pew, 2024). And while most people primarily engage with the entertainment elements of social media, they are still exposed to news and political content, that in turn increase their political knowledge and engagement (Nanz & Matthes, 2022).

But nothing comes for free. While platforms like Google and Facebook may seem free, the companies behind them generate significant revenue. In most cases, this revenue comes from the data these platforms collect about their users. Advertisers use this data to target ads to specific segments. Instead of paying a monthly fee, as with a Netflix subscription, we ‘pay’ for access to social media and search engines with the digital traces we leave behind.

The business model also craves more and more. The longer a user spends on the platform, the more ad revenue and data the platform can collect. This data is made available to advertisers, allowing them to show more targeted content to the individual user, who then spends more time on the platform because the content retains the user.

The loop described is also known as *the hype machine* and is often where blame is placed when negative outcomes are discussed (Aral, 2021). If we are only served the content we like on social media, many point out the risk of echo chambers, filter bubbles, and polarisation

(Sunstein, 2001; Pariser, 2011). The effectiveness of Facebook and other tech giants in using our data,, whether it benefits advertisers, and whether it contributes to echo chambers, are beyond the scope of this report. Later sections will examine what research says about the impact of social media on misinformation, polarisation, and well-being.

However, the business models and design of the platforms cannot be viewed in isolation as just another product we can buy online. The tech giants have embedded themselves at the heart of society, influencing public institutions, economic transactions, and social and cultural practices (van Dijck et al., 2018). Over time, they have integral to existing societal institutions, forcing governments worldwide to adjust the foundational legal and democratic structures of society.

One way tech giants infiltrate society is by promoting their services as public goods (van Dijck et al., 2018). The strategy is familiar, but it proves far more powerful for tech giants. Self-diagnosis, alternative treatments, private education, and learning resources are just a few of the many services society has to deal with. A notable example in Denmark is the so-called Chromebook case, where the Danish Data Protection Agency ruled that there was no legal basis for disclosing personal data to Google. As a result, Danish schools faced uncertainty about how to proceed with teaching until KL reached an agreement with Google in the summer of 2024, expected to meet the requirements of the Danish Data Protection Agency. Services from Google, Microsoft, Amazon, and others have become essential for the functioning of the public sector.

However, steps are being taken toward establishing a legal framework. EU legislation, especially the Digital Services Act (DSA) and the Digital Markets Act (DMA), protects minors from being targeted by personalised ads and prohibits targeting based on sensitive data such as sexual preferences and religious beliefs. Similarly, the AI Act regulates the use of artificial intelligence, a key area of focus for tech giants. The Act prohibits using artificial intelligence in cases where it poses an unacceptably high risk, such as collecting facial images from the internet to create facial recognition databases.

The challenge, however, is that regulations designed to secure one benefit sometimes undermines another. For instance, data transparency is essential for authorities to track down crime and terrorist networks, yet it often conflicts with user privacy. Such conflicts also bring deeper challenges to the surface, including differences between the approach of (primarily American) companies public discourse and Danish traditions. One such challenge is content moderation, which is the subject of the next section.

While this section has discussed tech giants broadly, we now shift focus to social media platforms specifically. Platforms like Facebook, Instagram, X (formerly Twitter), TikTok and Snapchat provide users with extensive opportunities for interaction.

The framework for public dialogue on social media

The way we access our news has changed dramatically since the rise of social media. Today, social media is one of the most popular and frequently used ways to access news, especially among young people (Baptista & Gradim, 2020). In Denmark, Facebook remains the most

commonly used social media platform for news (32%), even among young people (34%), while Instagram (19%) and TikTok (15%) are also frequently used (Newman et al., 2024: 76). And there's plenty of news to go around. Every day, over 100 billion messages are sent through Facebook products alone, and more than 500 hours of video are uploaded to YouTube every minute (Morrow et al., 2021).

One of the challenges that has emerged is the perception of social media content – and the internet as a whole – as a kind of Wild West. A stark example is the onset of the Rohingya genocide in Myanmar in 2017. Facebook has acknowledged that in the months and years leading up to the atrocities against the Rohingya in Myanmar, its algorithms amplified hateful content targeting this population. Incited by Facebook posts, thousands of Rohingya were killed, tortured, raped, and displaced as part of Myanmar security forces' campaign of ethnic cleansing (Amnesty International, 2022; Stevenson, 2018).

However, social media platforms are not always interested in maintaining a Wild West either. Although platforms have generally adopted a very liberal stance on extreme content, including incitement to violence, factors such as the threat of regulation, the difficulty of resisting lawmakers and the fear of advertiser flight provide incentives for platforms to engage in content moderation. For example, Twitter (now X) lost over 40 per cent of its advertising revenue in the months following Elon Musk's controversial takeover and the increase of anti-Semitic tweets on the platform (Weiss, 2024). A clear sign that the advertisers who pay for the whole ordeal won't put up with anything.

However, self-regulation doesn't silence the critics. There are two main reasons for this. On the one hand, there is the criticism of *too much* content moderation. Some feel their political opinions are being censored, while others feel they are subject to American cultural imperialism, where female nipples are strictly prohibited. On the other hand, social media is criticised for *too little* content moderation. Russian and Chinese influence campaigns and targeted misinformation, violence and cross-border material targeting minors (Børns Vilkår, 2024) and hate speech against ethnic minorities and women (Zuleta & Burkal, 2017) are some of the issues that have received attention in Denmark.

The Digital Services Act (DSA) provides authorities, citizens, and researchers with several tools to tackle illegal and harmful activities and work towards a fairer and more open online environment. The EU has already begun utilizing the legislation and started investigations into whether X is doing enough to prevent illegal content on its platform. If the platform does not comply with the legislation, it can be fined up to 6 per cent of its global turnover and ultimately banned from the European market.

However, the new rules, including the requirement to give researchers and others access to data, are being implemented at a slow pace. There are few success stories, and it is still unclear how accurate and effective a tool the DSA will be for providing data access. There is a need for authorities, researchers, and platforms to learn exactly what the rules entail and how they are enforced. Nonetheless, with the DSA, future studies can hopefully go deeper into descriptions of what content is actually being moderated on the platforms.

The Digital Services Act (DSA) introduces a wide range of new rules.
In essence, the regulation implies:

- Clearer rules for removing illegal content
- Rights to appeal content moderation decisions
- More transparency on recommender systems and adverts
- Restricting who can be targeted with ads
- Requirement for annual reporting of content moderation efforts,
- among other things
- Requirements to share data with researchers and others
- Better sanctions available to the EU and Member States

Active social media users and social media ad revenue

Social media platforms are very reluctant to share information about who is active on their platforms and how much ad revenue is generated in each country. The biggest challenge when it comes to such calculations is data availability, where no calculation can perfectly determine the correct amount. However, it will become easier in the coming years to estimate the ad revenue of all major platforms, as the European Digital Services Act (DSA) requires the creation of so-called ad libraries where information about adverts on the platforms can be found.

In addition to the general challenge of data access, there is also great variation between platforms in terms of what data can be accessed. While Twitter was the platform with the best available data for researchers and civil society for many years, Meta (the company behind, among others, Facebook and Instagram) has – somewhat surprisingly – become the platform with the most openness about data. As a result, in this section, we focus on Facebook. Not because it's ideal but because it's what's currently available. Likewise, this approach can be used in the future to investigate platform activity and advertising expenditure.

The data behind the figures comes from Meta's Marketing API and Meta's Ad Library API. Both data sources make data available to authorised developers. While the former data source shows information about who the users on Meta's platforms are, the latter data source shows details of which adverts can be found on Meta's platforms.

Figure 1 shows a breakdown of daily active users[1] on Facebook across all EU countries. To facilitate comparisons between EU countries, the figures are compared to each country's population. Highlighted in orange, Denmark is at the top of the European countries in terms of who has the most active users on Facebook, with approximately 75 per cent of the Danish population logged in daily. For the 13+ year olds – the age limit for creating a user on Facebook - the percentage is 81 per cent. Notably, there is a significant variation in Facebook usage across Europe, with just 40 per cent of Greeks and Poles logging into Facebook daily.

[1] Meta reports the number of active users and not people, which is why the number can be seen as an upper limit, as it is possible for a person to have more than one user. Likewise, activity should be understood in a minimal sense so that a user is considered active simply by having opened, for example, Facebook's app.

Figure 1: Percentage of people who are active on Facebook daily

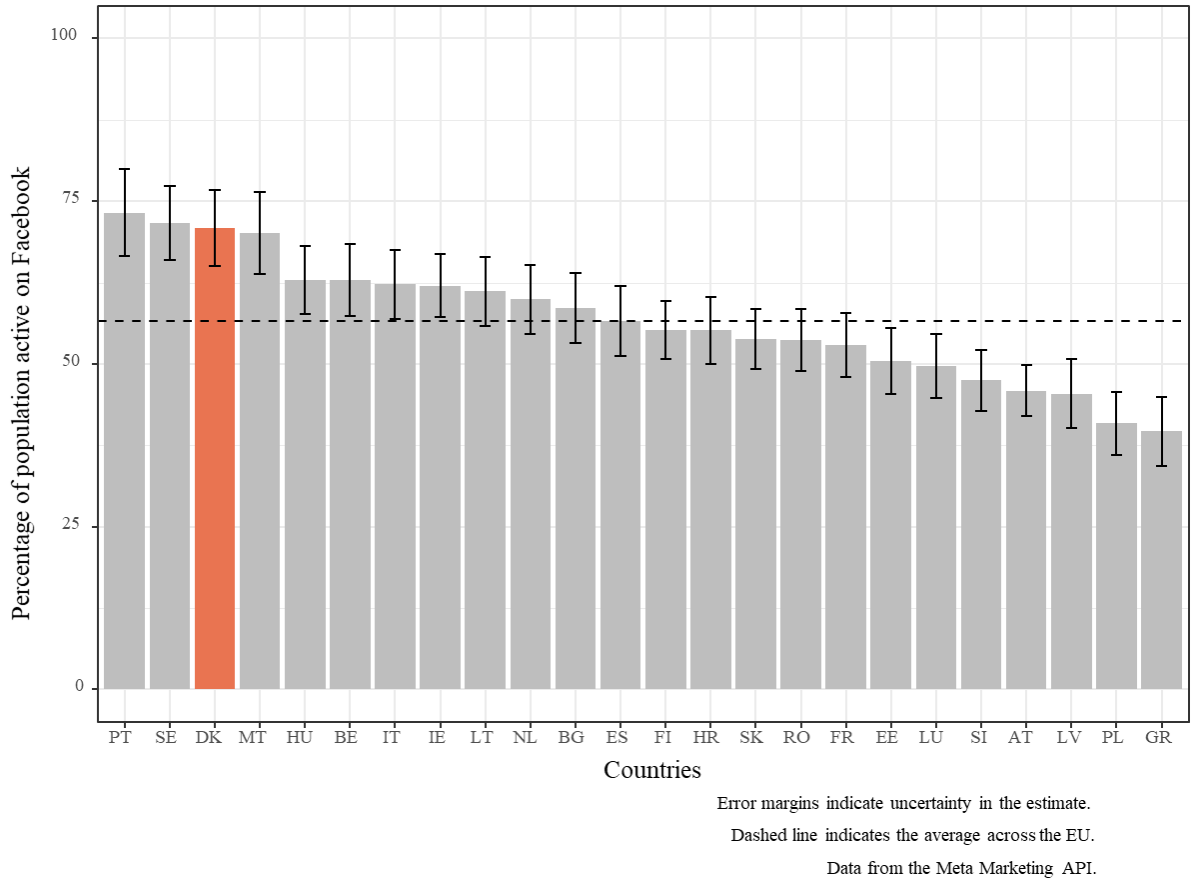
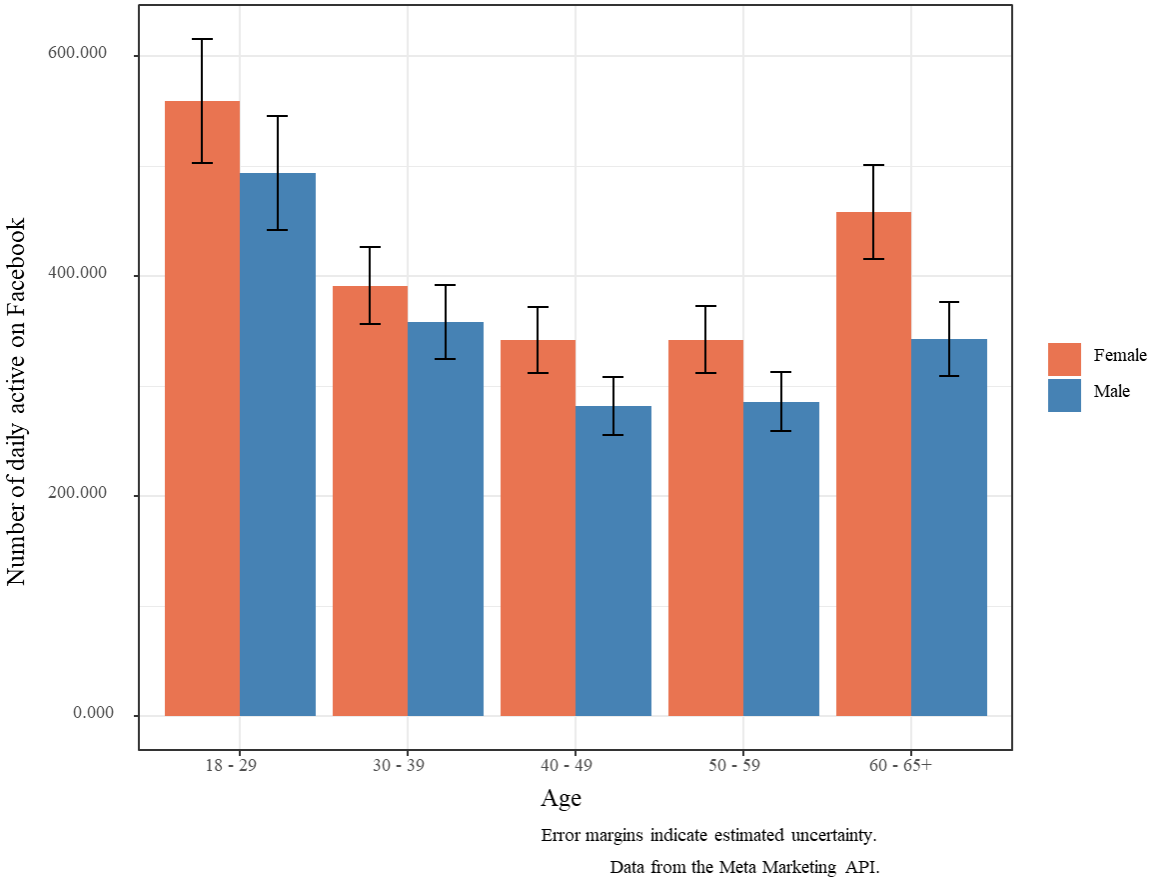


Figure 2 zooms in on which population groups are most present on Facebook, broken down by age and gender. As shown, there is higher proportion of active Facebook users among younger and older women. Additionally, women generally outnumber men in daily Facebook usage.

Figure 2: Young people, older people and women are more active on Facebook than middle-aged people and men



When it comes to ad revenue, it has long been difficult to directly observe how much revenue social media generates in each country. Generally, attempts to estimate advertising revenue have pointed out that an increasing share is shifting online and, more specifically, toward tech giants such as Google, Facebook and LinkedIn (Ministry of Culture, 2021). At the same time, data from Danske Medier Research reveals a steady decline in advertising revenue for daily newspapers in recent years (Danske Medier Research, 2024).

With the EU Digital Services Act, tech giants are required to create so-called ad libraries in which all adverts on the platforms must be published. While these libraries are still in the initial stages and access to their APIs is slow (Darius, 2024), it is not possible to systematically examine cross-platform ad revenue trends in this report. However, future reports are expected to use the ad libraries and methodology from the report here[2] to estimate platform ad revenue in a relatively accurate and systematic way.

We focus here on what can be observed directly, particularly the part of Meta's ad revenue related to social issues, where there has been transparency[3] about ad revenue for some time (see, e.g. Hove et al., 2024). Such ads are about politics broadly; for example, ads advocating a particular political viewpoint or mentioning a party are also included (Meta, 2024).

[2] See Appendix 1A for procedure.
[3] For reflections on the data quality of ad libraries see, e.g. Leerssen (2023) and Dommett (2023)

Ads about social issues on social media have been debated for many years, with concerns that they may manipulate voters and influence democratic elections. While research downplays this fear for several reasons (see, e.g. Hove, 2024), an overview of how much activity and money is spent on ads can help identify when they are placed and how much money ads related to public social debates represent in the advertising market.

Figure 3 shows the development in[4] how many millions of Danish kroner were spent in Denmark on adverts about social issues on Facebook and Instagram from January 2020 to June 2024. The figure clearly shows how the level from the end of 2021 to the end of 2022 was higher compared to 2020 and 2023 and that the level increased again in the spring of 2024. The fluctuations are highly correlated with the timing of the election, which is marked with dashed lines[5] in the figure. It is, therefore, clear that most adverts on social issues are placed when Danes are going to the polls. However, there is still between three and five million kroner worth of revenue outside election seasons.

Figure 3: Ad spending on social issues is highest during elections

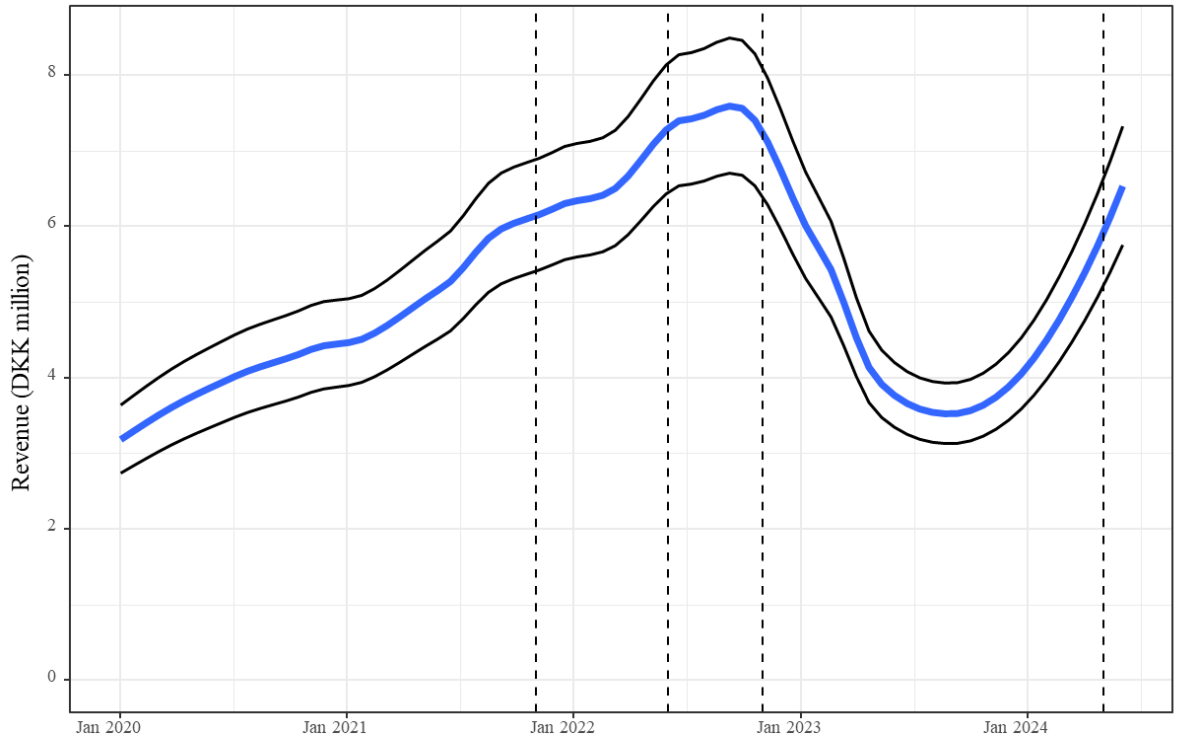


Figure 3 also highlights another interesting point: the order of magnitude for social ads is very low compared to estimates for Facebook's total ad revenue. Thus, social ads from January

[4] The development is smoothed with local regression to avoid large seasonal fluctuations such as summer holidays.

[5] Local and regional council elections in November 2021, referendum on the abolition of the defence opt-out in June 2022, general elections in November 2022 and European Parliament elections in June 2024.

2020 to June 2024 reached a revenue of almost DKK 300 million, while the estimate for Facebook's total ad revenue in Denmark in 2020 alone is DKK 1,387 million (Ministry of Culture, 2021). As a result, social advertising is not only highly volatile, but also represents a small part of the overall advertising market.

Reviews

Making judgments about the impact of tech giants on society requires actively engaging with a definition of democracy. What some consider healthy and necessary content moderation for democracy, others consider censorship. Our analysis and assessments are based on an understanding of democracy that ensures everyone has access to accurate information about news and politics, and where individuals can participate in public conversations without fear of intimidation and discrimination.

Based on the first part of the report, we make the following assessments.

A pluralistic set of digital service providers, including media, is crucial. A diverse ecosystem benefits both the market economy and democracy, for example, by ensuring that the media can fulfil its role as the public's watchdog. This aligns with the recently adopted European Media Freedom Act (EMFA), which requires Member States to ensure a diverse media environment and independent journalism.

Establish workflows that allow existing legislation to be effectively enforced and applied by authorities, researchers, and civil society. Legislation such as DSA and DMA is essential for analysing the societal impact of tech giants. However, it requires political persistence to realise the potential of this legislation for the benefit of governments, researchers, and civil society. This includes investing in effective information on legislation, providing resources to enforce the new rules, and facilitating access to the data enabled by the legislation for researchers, authorities, and other stakeholders.

Lack of data access remains the primary obstacle, particularly regarding knowledge of Danish conditions. Following the assessment of the need to ensure that existing legislation is utilised, it is crucial to recognise the fact that the legislation and programmes implemented are not only a pan-European responsibility but also a Danish one. The next phase will be critical for launching initiatives and programmes that ensure the knowledge produced is also rooted in Denmark. This will strengthen our evidence base for discussing how tech giants impact Danish society and identifying appropriate initiatives to address any challenges.

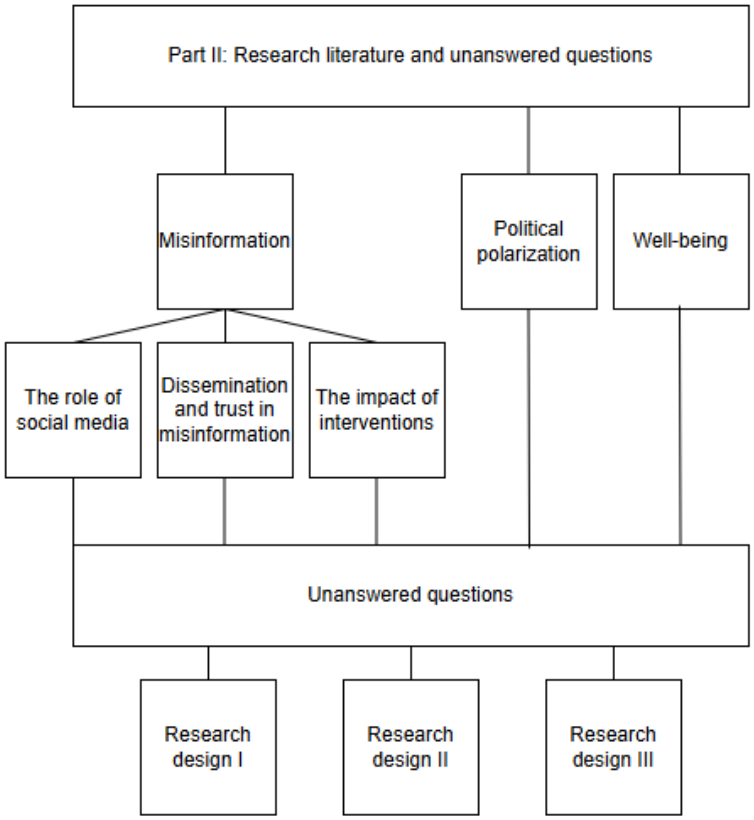
What do we know about the impact of social media on misinformation, political polarisation, and well-being?

Providing an overview of what we know and don't know from research on how social media affects our democracy, well-being or cohesion requires an unusually level head. Although talented researchers worldwide spend day in and day out researching the impact of social media, the task remains almost insurmountable. It has been – and continue to be – difficult to access the necessary data, including from the platforms themselves. Researchers are often left relying on questionnaires, observations, and other methods that can provide indications but are difficult to translate directly into the reality of what is happening. The problem is that we don't have a parallel world without social media to compare to. To think you can cut social media out of the equation and isolate the effect is naïve.

It is, therefore, easy to misinterpret (otherwise good and well-documented) research results. For instance, if a group of people logs off social media for three weeks and reports being in a better mood than those who stayed online, it doesn't necessarily mean that social media causes sadness. However, that doesn’t rule out the possibility either. It simply shows that in a world where we use social media every day, and the rest of our friends and acquaintances stay online, logging off for three weeks cane make us feel better. We don't have a parallel world without social media to compare to.

This section of the report provides an overview of research on the impact of social media on misinformation, political polarisation, and well-being. With this knowledge, we can identify important questions that remain unanswered and explore ways to address them.

The research literature is vast, offering many insights. Given the report's primary focus on misinformation, this section takes up the most space with several sub-sections. In addition, in line with the Media Agreement’s objectives, two shorter sections adress political polarisation and the impact of social media on well-being - areas that are also obvious focus areas for future reports. All three parts are summarised in a series of unresolved questions at the end of the literature review, leading to three concrete research designs that can be addressed in future reports. The figure below summarises the structure of the research review and the subsequent research questions, while Appendix 2a describes the methodology used for collecting and systematising the literature.



Misinformation on social media

The spread of misinformation is not a new phenomenon. In 17th-century France, citizens could buy a so-called "canard" on the streets of Paris - a newspaper that satiated the publics appetite for sensational and often false stories(Baptista & Gradim, 2020).

Today, misinformation is a concept familiar to most of us. It has been amplified by figures like former US President Donald Trump, who frequently referred to news media he disagreed with as *fake news*. Or the attempts by Russian trolls to influence the election with fabricated stories such as claims that Pope Francis urged Americans to vote for Trump or that Hillary Clinton was selling weapons to the Islamic State.

Experts and populations worldwide fear that misinformation can have negative effects, including influencing democratic elections, weakening societal cohesion, and eroding trust in the media and politicians. Going into the big election year of 2024, as many as 87% of people in countries voting in parliamentary elections were concerned that misinformation would influence the election (Quétier-Parent et al., 2023). The World Economic Forum, based on a survey of experts, identified disinformation as the biggest challenge facing society in the short term (World Economic Forum, 2024).

Wrong information is referred to in many different ways, often referred to as *fake news*, misinformation, and disinformation. While the terms capture important nuances – such as whether the information is intentionally false – for the sake of clarity, we will only refer to

misinformation, broadly defined as false claims presented as correct (Allcott & Gentzkow, 2017: 213).

The role of social media in spreading misinformation

The line between professionals and amateurs is even blurrier on social media

The content you encounter on social media is created by a more diverse group of so-called "content creators" than you would typically see in more traditional media. This blurs the line between who is considered a professional and who is an amateur. Combined with the ability to often remain anonymous, social media invites actors who may be interested in sharing misinformation – content that would have been harder to publish in non-digital formats (Kim et al., 2021; Shahzad et al., 2021).

Algorithmic curation of news risks increasing the spread of misinformation

However, it's not just who produces content on social media that increases the risk of misinformation spreading. Social media differs significantly from traditional media in terms of what users are exposed to. Unlike traditional media, social media uses algorithmic *feeds* that affect what is shown to individual users. As a result, users' demand for and attraction to sensational news can pave the way for the spread of misinformation (Akram et al., 2022; Shahzad et al., 2021). This is also the conclusion of one of the classic (and controversial) studies in the field, where researchers found that verified fake news spreads significantly further, faster, deeper, and broader than verified true news on Twitter (Vosoughi et al., 2018).

Algorithmic *feeds* show users content based on what they have previously clicked on, who they are friends with and what others are clicking on. In this way, posts and images that are highly attractive may go viral, with millions of people occasionally being shown the same post.

However, the extent to which algorithmic *feeds* influence what users see and shape their views remains uncertain. Many challenge the idea that algorithms drive the misinformation being spread and seen on social media. For instance, a US study shows that Americans who watch content from extremist channels on YouTube already exhibit sexist attitudes and high levels of racial antagonism. Similarly, researchers have observed that users blocked from interacting with exessive misinformation migrate to other platforms to seek out the same content (Budak et al., 2024). Comparable patterns are evident in Denmark, where a Danish study shows that individuals who are hostile on social media also display hostility in real life (Bor & Petersen, 2022). Consequently, yt is difficult to directly separate what is caused by social media algorithms and what is caused by user demand in the spread of misinformation.

Social media business models disincentivise the fight against misinformation

Another key explanation for why misinformation can be allowed to spread on social media is social media business models. To varying degrees, tech giants use the data they collect about users to optimise their business and resell to companies that want to advertise to specific segments (Aral, 2021). The attractiveness of misinformation thus risks becoming a good that generates clicks and attention and ultimately more revenue (Kaushik, 2024; Sanders & Jones,

2018). The incentive for the platforms is, therefore, not clear, as it is both difficult and potentially costly to reduce the amount of misinformation on the platforms.

Dissemination and trust in misinformation

Economics, ideology, and entertainment drive those who create misinformation

Those who create misinformation content are often driven by financial, ideological, and entertainment motives (Wu et al., 2024; Kim et al., 2021; Baptista & Gradim, 2020). The economic motive is often about luring people to a website by creating sensational *clickbait* headlines, with ad impressions generating income for the misinformation actors. Another example was high on the media agenda in Denmark earlier this year, where fake celebrity ads on Facebook attempted to scam people with promises of quick financial gains (Nisgaard, 2024). And it's not entirely unsuccessful. NewsGuard, which assesses the credibility of websites, estimates that the economic market for misinformation amounts to approximately DKK 18 billion annually in advertising revenue alone (Skibinski, 2021). Helped along by mainstream companies unknowingly advertising on websites known for spreading misinformation (Ahmad et al., 2024).

The ideological motive, on the other hand, is what most of us would probably associate with misinformation. Here, misinformation is spread to benefit favoured politicians and smear political opponents. It could be fake news sites like Breitbart News trying to help Republicans by spreading fake news about Democrats, or it could be foreign states like Russia trying to influence sentiment and election results in their favoured direction (Golovchenko et al., 2020).

Finally, there are those who spread misinformation and *troll* purely for their own and others' entertainment.

Misinformation makes up a small proportion of news but is also difficult to measure

The research literature does not completely agree on how big a challenge misinformation is. One challenge in analysing the prevalence of misinformation on social media is that it is often difficult to agree on how to understand misinformation and to measure what is misinformation and what is not. Therefore, depending on how misinformation is defined and measured, there is a risk of both underestimating the true amount of misinformation (Pennycook & Rand, 2021) and overestimating it (Budak et al., 2024). Another challenge is that our knowledge is limited by the fact that most research focuses on the US, which is a special *case* with very high levels of polarisation. Therefore, care must be taken when trying to translate American results to Denmark.

Previous research estimates that fake news sites account for between 0.7% and 6% of news stories *linked* to on social media on average (Altay et al., 2022). To put this into absolute terms, in the three months leading up to the 2016 US presidential election, there were at least 38 million Facebook shares of fake news (Allcott & Gentzkow, 2017), and posts by Russian internet trolls had a reach of up to 126 million US citizens on Facebook (Budak et al., 2024) and they sent 109,000 tweets (Golovchenko et al., 2020).

Whether those numbers are large typically receives two types of objections in the research literature. Firstly, there is the importance of selection. The type of people who share and

interact with misinformation is often a defined group (Budak et al., 2024). New research shows how 80 per cent of all shares of fake news on Twitter during the 2020 US presidential election were shared by 2,107 registered voters, corresponding to 3 per thousand of US Twitter users (Baribi-Bartov et al., 2024). Similarly, fake news is only a small part of what the average person on social media likes, shares, or clicks on (Pennycook & Rand, 2021; Baptista & Gradim, 2020: 11).

The second objection is based on the fact that people are generally exposed to a lot of information on the internet. If you take the 126 million Americans who may[6] have seen posts from Russian internet trolls, it only accounts for 0.004% of the content in their Facebook *feed* (Budak et al., 2024). Taking estimates of all online misinformation, not just on Facebook or by Russian trolls, the estimate is that misinformation makes up 0.15% of what Americans and 0.16% of what French people see online (Altay et al., 2023).

This suggests that exposure to misinformation is a small proportion of the information people receive on social media and that it is particularly concentrated in a small group.

Although this may sound positive, it does not mean the spread of misinformation can be downplayed. Firstly, misinformation thrives in times of crisis or uncertainty when reviewing and checking information is harder (Akram et al., 2022). Secondly, misinformation is also spread from otherwise trustworthy actors and in a way where the misinformation is subtle. A recent research article clearly lays out the challenge. Specifically, the researchers show how identified misinformation about COVID-19 received 8.7 million views on Facebook in the first three months of 2021, while factually correct but implicitly misleading vaccine scepticism was viewed hundreds of millions of times with a significantly higher persuasiveness (Allen et al., 2024). Thirdly, another recent research paper shows that Russian propaganda on geopolitics is regularly disseminated by a broad swath of US media on both ideological wings (Yang et al., 2024). Therefore, it's important to remember that just because something is difficult to measure doesn't mean that there is no effect.

Inattention and mental shortcuts affect whether people believe misinformation

When it comes to why people believe and accept misinformation, research typically points to two explanations. The first explanation is that people share misinformation because it aligns with their political worldview. In contrast, the second suggests that people share misinformation because they do not pay enough attention when they see and share information on social media. While research suggests that both factors are at play, evidence suggests that awareness is the stronger of the two factors (Pennycook & Rand, 2021; Kim et al., 2021).

In addition, research suggests that when information is fast-paced on social media, people often fall for misinformation because of *heuristics* – mental shortcuts the brain uses to reach a conclusion faster. Three of these shortcuts are mentioned in particular. Firstly, people are more likely to believe (mis)information if it is recognisable (Wu et al., 2022; Pennycook & Rand, 2021; Greenspan & Loftus, 2020). For example, if you've heard a rumour before, you're more likely to believe it to be true. Secondly, people often respond to social *feedback*. If you see

[6] It is not possible to determine whether a person has actually seen the post, only that it has been presented in their *feed*.

that a social media post has many *likes* or shares, you are more likely to find the post credible as many other people seem to interact with the content (Wu et al., 2021; Pennycook & Rand, 2021; Bryanov & Vziatysheva, 2021). Thirdly, a person's emotions and personality are part of the equation. For example, research suggests that people who report high levels of positivity or negativity are more likely to believe fake news (Wu et al., 2021; Pennycook & Rand, 2021; Bryanov & Vziatysheva, 2021).

Unclear how well people can distinguish between true and false on social media

Closely related to the question of whether people believe misinformation is whether they can distinguish what is true from what is false. However, it's difficult to simultaneously capture something real and avoid simply testing people's ability to remember what they've heard on the news. Therefore, the results are also mixed regarding the final conclusion about the ability to distinguish true from false. Some studies find that people do no better than chance, while others show that people are very much able to distinguish (Bryanov & Vziatyshea, 2021).

This may be due to attention span and ability to think analytically. Thus, people who are more analytically minded and attentive are more able to separate true from false information, and they are better at assessing which news is politically *biased* than the less analytically minded and attentive (Pennycook & Rand, 2021; Bryanov & Vziatysheva, 2021; Baptista & Gradim, 2020). Additionally, people are better at separating true from false when it comes to politics than when it comes to topics where most people generally have less knowledge, such as health and science (Bryanov & Vziatysheva, 2021). And while people are more likely to believe information that is politically congruent with their views, they are also better at identifying misinformation in politically congruent information. Overall, this indicates that belief in misinformation comes from inattention and not because they have been politically *hijacked* into believing something (Pennycook & Rand, 2021).

Besides attention and analytical thinking skills, the research literature also points to other differences. Those less able to distinguish often have lower digital skills, less education, strong ideological beliefs and distrust of media, and older people and men are more likely to believe misinformation (Baptista & Gradim, 2020).

The impact of interventions against misinformation

This section focuses on possible interventions that can be used against misinformation. Specifically, we look at what the research says about the effects of different interventions. Therefore, we do not address the part of the research dealing with identifying misinformation online.

Fact-checking is an effective intervention but not very scalable

One of the most researched methods to combat misinformation is fact-checking, often marked with warning signs. Warning signs can include a yellow warning triangle or red cross and a description that professional fact-checkers have rated the content as false. Research findings show that such interventions can reduce trust in, sharing, and interaction with misinformation (Martel & Rand, 2023; Pennycook & Rand, 2021). Three things in particular affect whether and how well the intervention works. Warning signs are effective when they are prominently placed, clearly mark content as false and include an explanation of what is wrong with the

information shown (Martel & Rand, 2023; Morrow et al., 2021; Greenspan & Loftus, 2020; Walter & Tukachinsky, 2020). Furthermore, the effect is small when the correction comes from other citizens compared to when it comes from news organisations and experts (Walter et al., 2021; Walter & Tukachinsky, 2020). At the same time, the effect depends on the topic in question, where it is easier to intervene against misinformation about crime and health and, to a lesser extent, against *marketing* and politics (Walter & Murphy, 2018).

Previously, there were concerns in research that fact-checking could have the opposite effect, known as *backfiring,* where people become even more convinced of their false beliefs. However, recent studies indicate that such *backfire* rarely occurs (Morrow et al., 2022; Bryanov & Vziatysheva, 2021). However, research suggests that fact-checking does not entirely remove trust in misinformation, as not everyone sees it (Martel & Rand, 2023). People tend to remember (mis)information more easily than fact-checking (Baptista & Gradim, 2020; Greenspan & Loftus, 2020), and warning signs have a hard time with repeated misinformation, as a new warning sign does not appear every time (Martel & Rand, 2023; Morrow et al., 2022).

While fact-checking can be effective, there are several challenges associated with the method. Fact-checking is not easily scalable. It requires people to review and check content, making it challenging to roll out widely. Another challenge is the risk of creating "implicit truth", where the absence of warning signs can lead people to believe that the information is true when it is not necessarily so (Martel & Rand, 2023). A third challenge is that the effectiveness of fact-checking depends on the platform and medium, which is why newer challenges such as *deepfakes* and generative AI risk making fact-checking more difficult (Martel & Rand, 2023; Morrow et al., 2021).

'Vaccination' against misinformation as an innovative intervention

Another approach to tackling misinformation that has become particularly popular in recent years is *prebunking*, also known as *inoculation* techniques. These techniques teach people to recognise the strategies often present in misinformation, such as polarising language (Pennycook & Rand, 2021). The aim is to "vaccinate" the population against falling for fake news by increasing their critical thinking and *digital literacy*.

Research has shown that such techniques can significantly improve people's ability to identify fake news. This works better than the fact-checking that occurs after a person has been exposed to misinformation (Greenspan & Loftus, 2020). For example, participants in the US and India became 26% and 19% better at identifying fake news after receiving training on recognising misinformation (Bryanov & Vziatysheva, 2021).

While *inoculation* has proven effective, it doesn't necessarily solve the fundamental problem of people often sharing information without assessing its accuracy. In addition, the solution also has scalability challenges, as it would be difficult to get people worldwide to participate in such a training programme.

***Accuracy prompts* and *crowdsourcing* are more scalable but not without challenges**

More scalable approaches to combating misinformation include the interventions *accuracy prompts and crowdsourcing*.

Accuracy prompts involve asking people to consider whether the information they are about to share is true before they are allowed to click send (Pennycook & Rand, 2021; Greenspan & Loftus, 2020). This method can be effective as it makes people stop and reflect on the credibility of the content, reducing the spread of misinformation (Morrow et al., 2021). Specifically, *accuracy prompts* lower intentions to share fake news by approximately 10 per cent, and those who received such a *prompt* are 72% better at identifying fake news than those who did not receive one (Pennycook & Rand, 2022). There is no difference in the effect of *accuracy prompts* across gender, ethnicity, ideology, education, or desire for accuracy. However, the effect is greater in older people, people scoring higher on a cognitive reflection test, and those generally more alert (Pennycook & Rand, 2022).

Crowdsourcing is where ordinary users flag content they believe is misleading (Pennycook & Rand, 2021). This way of labelling misinformation is particularly appealing because it is significantly more scalable and continues to effectively lower trust in misinformation, albeit not as much as fact-checked content (Martel & Rand, 2023). While there might be an expectation that political divides would influence judgments, people generally agree on what is low and high-quality news across political viewpoints (Pennycook & Rand, 2021). However, this system requires, among other things, that users are aware of which media outlets are known to spread misinformation, which is not always the case, especially regarding more niche-orientated media.

There are several challenges associated with these methods. Firstly, the challenge of "implicit truth" is not overcome as it will still not realistically be possible to cover all platform content. Secondly, there is a greater risk of *tainted truth*, where misplaced warnings on true information reduce trust in correct information in general (Martel & Rand, 2023).

The effectiveness of warning signs also varies across platforms and types of media. For example, *memes* and *deepfakes* are particularly challenging to label correctly. Different types of warning signs also have varying degrees of success, with detailed warnings generally being more effective than general warnings. In addition, it is crucial that warning systems are transparent and fair to avoid accusations of censorship and maintain public trust.

***Deplatforming* can potentially be effective but is a balancing act with perceived censorship**
Another often-discussed option, which is primarily in the hands of the platforms themselves, is the possibility of *deplatforming*. This includes the possibility of blocking and suspending users who spread misinformation (Nasery et al., 2023). A new study examines the impact of social media platform Twitter *deplatforming* a large number of users who shared misinformation, including then US President Donald Trump (McCabe et al., 2024). The study shows how the intervention reduced the amount of misinformation circulating on the platform. Several other users who were prone to spreading misinformation but had not been *deplatformed* left Twitter shortly afterwards.

However, there are only a few empirical studies and few strong causal estimates (Golovchenko, 2022; King et al., 2013). At the same time, other research shows that although Facebook removed anti-vaccine content during the COVID-19 pandemic, it did not reduce the number of interactions with anti-vaccine content (Broniatowski et al., 2023). When it comes to *deplatforming* as a strategy, it is still unclear whether it is an effective strategy.

A challenge with *deplatforming* and other content moderation is the risk that users may perceive it as censorship (Nasery et al., 2023). In general, social media users express both a desire for platforms to do more to remove misinformation and, at the same time, a desire for less censorship (Morrow et al., 2021). We see the same pattern in Denmark, where Danes largely agree on the need for content moderation on online platforms but are more divided on whether it harms freedom of expression (Centre for Social Media, Tech and Democracy, 2023).

Having reviewed key parts of the literature on misinformation, we now move on to two shorter sections that outline the overall findings and discussions within the research literature on the impact of social media on well-being and political polarisation, respectively.

Political polarisation

One of the most debated topics regarding the effects of social media is political polarisation. The main argument is that social media and its algorithms facilitate echo chambers that change the information people are exposed to, thus dividing politicians and the population into ideological groups (Iyengar et al., 2019).

However, research has been divided on how much blame social media carries for political polarisation. Research examining people's attitudes using questionnaires finds a correlation between having polarised opinions and receiving news from media you agree with politically (Garrett et al., 2014; Lu & Lee, 2019). Similarly, several experimental studies show how, for example, exposure to ideologically aligned information sources can have a polarising effect (Levendusky, 2013).

However, the challenge is that asking people questions in a questionnaire or testing maps and simulated content in an artificial situation doesn't get to the heart of the problem. For example, it's hard to know how easily such findings generalise and how long the effects last. Therefore, getting to the heart of the matter requires opening the lid on social media algorithms and giving researchers access to run experiments directly on the platforms.

A first step is a large-scale collaboration between Meta, the company behind Facebook and Instagram, and a large number of researchers who have been allowed to run experiments on the platforms during the 2020 US presidential election campaign. The results of Facebook and Instagram's ability to highlight political polarisation among their users during the presidential election campaign are clear. For example, the researchers blocked one group of voters' access to Facebook and Instagram during the election campaign; for another group, they adjusted the algorithm according to the amount of content from news sites that the voters were ideologically aligned with; for a third group they removed the ability to view (*reshares*); and for a fourth group, they removed the algorithmic curation of *newsfeed* content. None of the experiments showed any evidence that the changes impacted users' level of political polarisation (Guess et al., 2023a; Guess et al., 2023b; Nyhan et al., 2023; Allcott et al., 2024).

The studies from the collaboration with Meta are not unproblematic, nor do they overcome the problem that we can't find a world that isn't awash in social media, which is why the studies

alone indicate that content on Facebook and Instagram does not seem to polarise now that we all have social media and have been using it for many years. Therefore, while the conclusion is that research is struggling to find that social media is polarising users at this point, more research is needed to answer the concern conclusively.

The impact of social media on well-being

Social media is currently a hot topic in both politics and research. Globally, as well as in Denmark, there is a considerable focus on the amount of time people, especially young people, spend on social media. The fear is that social media is leaving a trail of ills in its wake, especially the unhappiness that this section focuses on.

Although the question has been analysed and discussed in the research world for years (see, e.g. Orben & Przybylski, 2019 and Twenge et al., 2020), the issue reached boiling point in the spring of 2024 when renowned American social psychologist Jonathan Haidt published his latest book *The Anxious Generation*. Haidt argues that the fear of social media is justified: mobile phones and social media have replaced physical play and socialisation and caused significant mental health problems as young girls are constantly exposed to unhealthy beauty ideals, and young boys are exposed to extreme content such as murder and porn. Specifically, Haidt shows how the rise in mental health problems comes at a time when more people are increasingly using digital technologies.

Haidt's arguments have been widely criticised (Odgers, 2024; Thorp, 2024) referring to many research results that cannot find clear evidence for the critical portrayal of social media when looking at the correlation over time (Heffer et al., 2019; Odgers & Jensen, 2020; Orben, 2020; Valkenburg et al., 2022). Directly opposed to Haidt's argument, some believe that the correlation is reversed, with young people who already have mental health problems using social media more and perhaps in different ways than young people without mental health problems (Heffer et al., 2019). Similarly, studies that experimentally test the thesis by having a group abstain from social media for some time find that it is not possible to find a non-zero effect of social media on well-being on average (Ferguson, 2024; Radtke et al., 2022).

However, the research results are a matter of great debate. Several studies point in the opposite direction and thus point to the negative impact of social media on the mental health of children and young people in particular (Weigle & Shafi, 2024; Blanchard et al., 2023; Khalaf et al., 2023; Ergün et al., 2023). Similarly, some studies have attempted research designs that utilise the natural variation in, for example, the gradual roll-out of social media, Facebook. As Facebook was gradually rolled out at US universities, researchers were able to compare students' well-being before and after Facebook was rolled out on their campus (Braghieri et al., 2022). The researchers found that Facebook had a negative impact on mental health, arguing that Facebook provides an unhealthy opportunity to compare oneself with fellow students.

However, this study is also criticised (Eckles, 2023), and at a higher level of abstraction, it can be difficult to always know exactly which leg to stand on. Firstly, the disagreements – even those that are not strictly research-based – can be about how well-being is defined and that we know very little about what content people actually see on their screens or how social

media is used differently by different people (Reeves et al., 2020). Secondly, it's difficult to isolate the impact of social media when it's such a widespread and integral part of most people's daily lives, which is why most studies can only show a subset of the overall narrative. Therefore, it is crucial to keep the research on track and obtain more and better data to get an accurate picture of the extent of the problem.

Still unanswered questions and suggestions for research designs

While the previous part of the report focused on what we know from research about misinformation, political polarisation, and well-being, we now focus on what questions remain unanswered and how it will be possible to answer some of them in future reports. The following is, therefore, first a review of a number of unanswered questions and then a review of three research designs that could answer some of those questions.

Questions still unanswered

The unanswered questions in this section are formulated as a series of research questions, some of which are addressed in the following section to define research designs that can be used in future reports.

We also formulate several relevant research questions related to misinformation in light of **generative AI.** While the literature review downplays the fear of misinformation at some points, there is a danger that generative AI could change that. But there is also the potential for positive side effects. We address this with questions that capture potential shifts in the spread and belief of misinformation due to generative AI.

Generative AI (generative artificial intelligence) is a technology that can create new content (e.g. text, images, audio or video) based on existing data and instructions. The most well-known example of generative artificial intelligence is the chatbot ChatGPT

Misinformation

Danes' behaviour towards misinformation

The literature review revealed two striking features. Firstly, what we know from research about misinformation is predominantly American. However, the US is very different from Denmark in terms of language, polarisation and (political) culture – to name just a few key differences. It can, therefore, be difficult to translate the results from the US to Denmark directly, so it is necessary to know more about misinformation in a Danish context. Secondly, most existing research is based on surveys, which can only identify people's attitudes and perceptions of misinformation and rarely their behaviour.

Therefore, the first research questions we pose in this section emphasise the need to capture Danes' behaviour towards misinformation.

To what extent are Danes exposed to and interact with misinformation?
To what extent does Danes' interaction with misinformation affect their overall information diet on social media?

Production and distribution of (mis)information

One of the key effects of generative AI is not only creating unintentional misinformation through “hallucinations” but also the ability to quickly produce a lot of realistic information. Therefore, one of the dangers is that generative AI can be used to create misinformation deliberately in large quantities.

This is especially important given that the fabric of social media is (to varying degrees) built around virality: a post, image, or video that breaks through the ceiling and gets spread to millions of people. While the process isn't random, the formula isn't necessarily clear to anyone other than the algorithm running the show. With generative AI, it becomes easier for malicious actors to create more content, all of which is entered into the great virality lottery, increasing the risk of misinformation going viral.

At the same time, generative AI can, of course, also be used to both create and circulate true information, so the balance between how many tickets good and bad actors have in the lottery does not necessarily change.

To what extent does generative AI make producing and distributing misinformation easier?

The effects of (mis)information

One of the situations where there is a high demand for information is in crises. However, these cases are characterised by little information being available because the media, politicians, and others are still investigating them.

Public awareness of the existence of generative AI and its risk of being used to spread misinformation risks lowering trust in the news we see and are exposed to in crises.

Conversely, the public's awareness of this can also increase the demand for authentic knowledge by being more critical and reflective of the news they are presented with.

How does generative AI affect public trust in news and information in crises?

In recent years, one of the most debated cases of generative AI has been so-called *deepfakes*. *Deepfakes* are characterised by replacing one person's expression with another in a convincing way. It could be creating a video where a politician expresses opinions and statements they have never had or said. An example of this was in the spring of 2024, when the Danish People's Party came under fire for producing and sharing a *deepfake* with Mette Frederiksen stating that Danes work too little and that all public holidays should be abolished (Uldall, 2024).

Besides being problematic in itself, such phenomena risk leading to what is known as *tainted truth*. Situations where people question whether they can trust otherwise true information with the fear that they cannot distinguish between what is true and what is false.

On the other hand, discussing misinformation and generative AI may increase the population's critical thinking. Similar to the idea of being able to "vaccinate" against misinformation, such a discussion will help people better identify when information is trustworthy and when it is not.

To what extent does generative AI affect people's ability and belief that they can distinguish between true and false information?

The section on the prevalence and belief in misinformation touched on which groups of the population research has identified as most at risk when it comes to misinformation. One of the arguments for why some people are more prone to falling for misinformation is that older people, for example, are not as internet savvy as the younger population.

Generative AI makes it easier to produce high-quality content at a scale where it's less necessary to aim the elephant gun at individual groups. Instead, it can be easier to aim to convince all population groups with a scattergun approach. On the positive side, it can also make it easier to reach people with credible information about what information to watch out for.

To what extent are differences between those who are good and those who are less good at identifying misinformation blurred due to generative AI?

Interventions against misinformation

Interventions such as warning signs and fact-checking are important in the fight against misinformation. However, such interventions are at risk of being affected by generative AI, which can flood *crowdsourced* interventions with false fact-checks, among other things. Such flooding can cause *warning fatigue* (Morrow et al., 2022), where the sheer volume of warnings makes people tired of having warnings and fact-checks on everything. Similarly, fact-checking becomes harder to perform because there is more uncertainty about whether an image or video is AI-generated or not. This opens the door for malicious actors to skirt criticism because the information about them is false and AI-generated.

Although this report hasn't focused on the part of the research dealing with identifying misinformation, this is where generative AI presents positive opportunities. Generative AI can potentially be used to scale otherwise hard-to-scale interventions and make it easier to identify misinformation.

To what extent does generative AI influence the effectiveness of interventions against misinformation?

Political polarisation

Research still can't tell us much about what political polarisation on social media looks like outside of the highly elite and affectively polarised US. Therefore, extending the focus to multi-party systems like Denmark alone will increase society's awareness of the potentially polarising life on social media. Against this backdrop, we find it relevant to focus on the following questions:

To what extent are Danes presented with different content compared to each other on social media?
To what extent does content become more polarised over time?
To what extent do Danes interact with polarising content?

Well-being

While it's not easy to isolate the effect of social media on user well-being, there are still important and more accessible questions to explore, especially in Denmark. There is still a lack of knowledge about how much problematic content children and young people are exposed to and the differences in the content that thriving and non-thriving young people see on social media. Therefore, it makes sense to test the hypotheses often put forward in documentaries, political debates and the like, especially by examining the following questions:

To what extent are young people exposed to problematic content on social media?
What characterises differences in interaction and exposure patterns between adolescents that vary in self-reported well-being?

The questions formulated in this section have highlighted some of the areas where we still lack knowledge about misinformation, well-being, and political polarisation in relation to social media. These are, therefore, apparent questions for future research and studies to address. In the next section, we identify some of them and investigate how they could be examined.

Suggestions for future topics and research designs

The main challenge when researching social media and its impact on society is the lack of data and good research designs. This section sets out three research designs based on some of the uncovered research questions in the previous section. The three research designs explain how to investigate the different questions and the specific data needs for each study. These research designs can be used as benchmarks for future studies.

To stay focused on the Media Agreement's priorities for the impartial study, we set up three research designs that focus on Danes' online behaviour towards misinformation, political polarisation on social media in Denmark, and the impact of social media on the well-being of Danish children and young people.

Research Design I: Danes' behaviour towards misinformation online

The literature review focused primarily on people's perceptions of misinformation. This is because people's behaviour regarding misinformation is data that is not easily accessible. However, it is an important question, and this research design outlines what a study of Danes' behaviour towards misinformation on social media could look like. The research questions are as follows:

- To what extent are Danes exposed to and interact with misinformation?
- To what extent does Danes' interaction with misinformation affect their overall information diet on social media?

The research design requires data on what content Danes are exposed to and interact with on different platforms, as well as knowledge about their media and information diet, ideological attitudes, predispositions, and demographic characteristics. Since the research design has all

Danes in mind, Facebook, Instagram, and YouTube are the ideal platforms to focus on, as these platforms are the most widely used in Denmark and have been for a long time. Specifically, data for this design will come from two sources:

1) **A questionnaire survey of a representative sample of Danes** included questions on ideological self-placement, belief in conspiracy theories, predispositions, information diet, and some demographic variables.
2) **Observational behavioural data on content exposure through DSA.** Data should include the content that the representative sample of Danes has been exposed to and interacted with on Facebook, Instagram, YouTube, and Snapchat. In addition, the data should include information about whether the content has been reported as misinformation and whether the platform has subsequently taken down the post.

The data can be used for a number of analyses. Firstly, **descriptive statistics on how much and what type of misinformation Danes are exposed to and interact with**. This applies to both the amount of misinformation and its proportion of the total information diet. It can also provide an overview of which sources the misinformation most often comes from. Secondly, **whether Danes are more exposed to misinformation after interacting with it**. This can be investigated with a *staggered difference-in-differences* design, for example, which examines whether interaction with misinformation increases future exposure to misinformation. Thirdly, **correlations between Danes' background characteristics and exposure to misinformation**. This could be whether some Danes are more exposed to misinformation across ideological self-placement, belief in conspiracy theories or political trust.

Research Design II: Political polarisation on social media in Denmark

The vast majority of existing research on the issue of political polarisation has been conducted in the US. However, as mentioned, the US is very different from Denmark in several key areas, including the level of elite polarisation. It is, therefore, difficult to transfer such knowledge directly to Denmark. Specifically, the research design here is, therefore, set up to answer the following research questions:

- To what extent are Danes presented with different content compared to each other on social media?
- To what extent does content become more polarised over time?
- To what extent do Danes interact with polarising content?

To answer the two research questions, observational data on what content Danes see and interact with on social media and how this content evolves over time are needed. For this research design, Facebook and Instagram are the most obvious social media platforms from which to analyse data, as they are used by the largest proportion of Danes. Specifically, data for this design will come from three sources:

1) **A questionnaire survey was conducted among a representative sample of Danes**. The questionnaire contained questions on ideological self-placement, affective polarisation, and several demographic variables.
2) **Data donations from the same representative sample of Danes**, containing which pages they have interacted with on Facebook and Instagram and which interests the platforms have inferred based on their behaviour.

3) **Observational data on content creation through *Meta Content Library***. Here, you can retrieve all posts from public pages (e.g. media and politicians) on Facebook since 2009.

This data can be used for a number of analyses. Firstly, **network analysis can be used to investigate the distance between Danes** in terms of which sites they follow and interact with. This will help describe the distance between different population groups in terms of what they see and interact with on social media. Secondly, **correlations between what Danes interact with and political beliefs** can identify how much political and content segregation there is on social media. Thirdly, **a time series analysis of the degree of polarisation in content from 2009 to today**. This could be investigated by analysing the sharing of news on social media by politicians, opinion leaders and interest groups (Eady et al., forthcoming), and it would also be possible to examine the development of negative content over time with sentiment analysis and analysis of Danes' type of reactions to the content.

Research Design III: The impact of social media on the well-being of Danish children and young people

As the literature review described, the extent to which young people are exposed to problematic social media content and how it affects them (positively and negatively) is still debated. As we still lack clear theoretical expectations about, among other things, differences between young people in their exposure to problematic content and its effects, the research design here will focus on the following two more fundamental research questions:

- To what extent are young people exposed to problematic content on social media?
- What characterises differences in interaction and exposure patterns between adolescents that vary in self-reported well-being?

To answer the two research questions, observational data on the content young people are exposed to on social media and self-reported data on their well-being are needed. To get the most substantial possible basis, data from Instagram, TikTok and Snapchat, which are particularly prevalent among young people in Denmark, is analysed. Data in this case will come from two sources:

1) **A questionnaire survey among a representative sample of young people**, including questions measuring self-reported well-being, as well as social and demographic factors such as interaction with friends, gender, age, and parents' educational status.
2) **Observational behavioural data on content exposure through DSA**. For each of the representative young people, the platforms are asked to provide access to data on what content the young people have been exposed to and interacted with over the past two years and information on what interests the platforms have inferred the young person to have based on their behaviour.

This data can form the basis for a number of analyses. Firstly, **a descriptive presentation of problematic content** based on content coding that considers the total amount of content presented. Secondly, **a correlation between well-being and exposure to problematic content**. Thirdly, **a network analysis of young people's interaction patterns** provides insight into which interests and interactions correlate with problematic content.

Limitations and resource requirements for the three research designs

The research designs described are designed to provide a more systematic descriptive analysis than has generally been possible in the research literature to date. However, it is relevant to note that the research designs cannot determine causal relationships. For example, the research design on the relationship between problematic social media content and young people's well-being cannot distinguish what causes young people with varying well-being to seek out different content and what causes them to be unhappy because they are exposed to problematic content. Such a design is challenging to set up without experimental manipulation, which is not possible with existing data sources and, in most cases, would not be ethical. Therefore, the three research designs will primarily reflect descriptive inference combined with some quasi-experimental designs.

All three research designs are resource and time-intensive. Data obtained through either the Digital Services Act (DSA) or *Meta Content Library* options requires lengthy and uncertain data request processes. In addition, data donations require a relatively large amount of effort from participants. They have to request data from the platform, wait to receive it, and then *upload* it to the research group, all of which increases the cost of obtaining data. Similarly, surveys among children and young people are more expensive as we are dealing with a group that is harder to reach for this type of survey. All three research designs, therefore, require both time and significant resources if they are to be conducted in the format described.

New survey on Danes' views and abilities in relation to misinformation and generative AI

In this final part of the report, we present new insights into Danes' views and abilities in relation to misinformation and generative AI. We do this based on a survey conducted in June 2024, in which we asked 2,091 Danes, a representative sample of the population, questions about three different topics.

As we saw in the first part of the report, content and content moderation on social media are important for online public conversation. However, there is still much criticism that too little is being done in this area. It's also an area where cultural schisms become apparent: the Danes who participate and discuss on social media are often subject to American cultural norms that sometimes clash with general Danish liberalism. The first theme we asked Danes to address was their *attitudes towards content moderation and regulation of social media*.

The literature review showed that it is uncertain how good people are at distinguishing between true and false information and that this difference could potentially increase with the advent of generative AI. However, there are no such studies in Denmark, and most existing studies lack the ability to distinguish between people's memories and the ability to identify misinformation. Therefore, the second theme we asked Danes to consider is their *ability to identify misinformation*.

The literature review also revealed negative indirect effects of misinformation, such as whether fear of misinformation also lowers trust in true news. It was also concluded that the proportion of misinformation online compared to other information is less than most people realise. Therefore, we are interested in knowing whether the awareness and publicity of misinformation have potentially negative knock-on effects that are important to consider. The third and final theme we asked Danes to consider is *how different interpretations of the dangers of misinformation*.

Attitudes towards content moderation and social media regulation

The first theme we address here is the question of Danes' attitudes towards content moderation and social media regulation. However, the literature review showed that attitudes are difficult to capture. People tend to have conflicting desires, such as wanting a high degree of content moderation and a high degree of freedom of expression simultaneously (Morrow et al., 2022).

Instead, our survey looks at people's attitudes towards actual content moderation policies on Facebook. We select different examples of content that is removed from the platform according to its policy and ask respondents how much they agree that this type of content should be removed from social media. These examples also vary in extremity, with the most extreme issue being whether threats of violence are outright illegal under Danish law (Section 266 of the Criminal Code). Together, the responses indicate Danes' overall satisfaction with existing content moderation policies and differences across different types of policies. All presented examples and corresponding statements from Facebook's content moderation

policy can be found in Appendix 3A. However, it should be noted that we have not investigated whether Facebook follows its content moderation policy and removes content accordingly.

Figure 4 shows the distribution of respondents' agreement with the content moderation examples presented. Each respondent was asked to rate six types of content, and their opinion on content moderation is summarised in an index. A low value indicates that the respondent does not agree with the content moderation policies, while a high value indicates strong agreement. In addition, the two types of content that the majority and least agree should be removed are marked with dashed lines at their respective averages, as is the content type closest to the distribution average.

Figure 4: Danes are generally – but not wholly – satisfied with existing content moderation policies

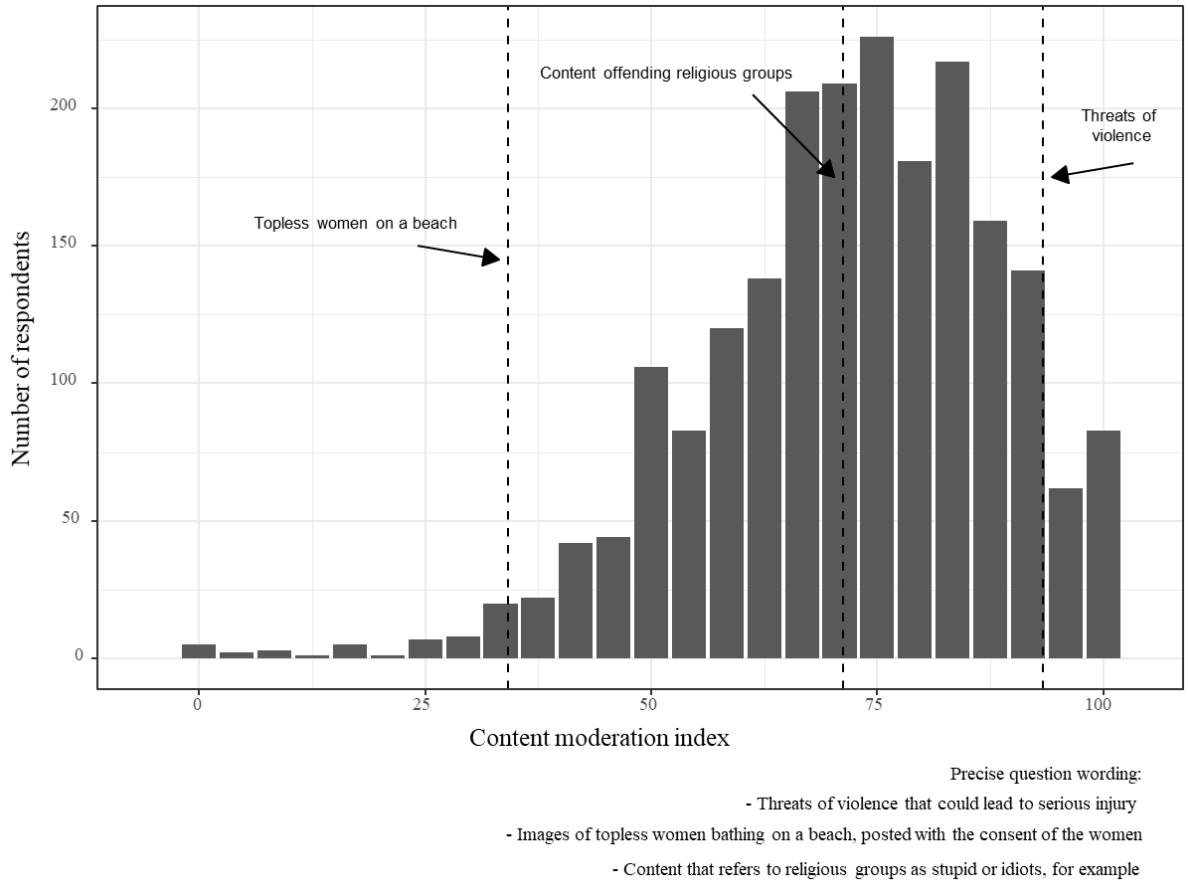


Figure 4 shows how **Danes generally agree with the content moderation they experience on Facebook, but not unconditionally**. Firstly, the spread in the distribution shows a big difference in how much Danes agree. But it's also worth noting that most Danes have moderately positive attitudes towards existing content moderation. Only four per cent of Danes say they "strongly agree" with all policies, and only eight per cent, on average, disagree more than they agree. Secondly, there is a big difference in how much Danes agree with the different types of content moderation. Almost all Danes strongly agree that threats of violence should be removed, while most disagree that images showing topless women on a beach posted with their consent should be removed. This is true for both men and women, although, on average,

women are slightly more positive about content moderation than men on all questions. The policies close to the distribution average are content moderation that offends religious groups and information that contradicts health authorities' recommendations.

In addition to content moderation, we asked respondents about their opinions on four regulatory recommendations from the government's appointed tech expert group (Ministry of Business, 2024). The expert group made several recommendations on how to ensure democratic control over the business models of tech giants. These recommendations, if implemented, would change the content we see on social media and how they are algorithmically structured.

Using the same structure as Figure 4, Figure 5 shows the distribution of respondents' agreement with the various recommendations, where a high value on the index indicates high agreement, while a low value indicates low agreement. The two recommendations with the most agree and disagree votes are marked with dashed lines.

Figure 5: Danes are generally very positive about regulatory proposals

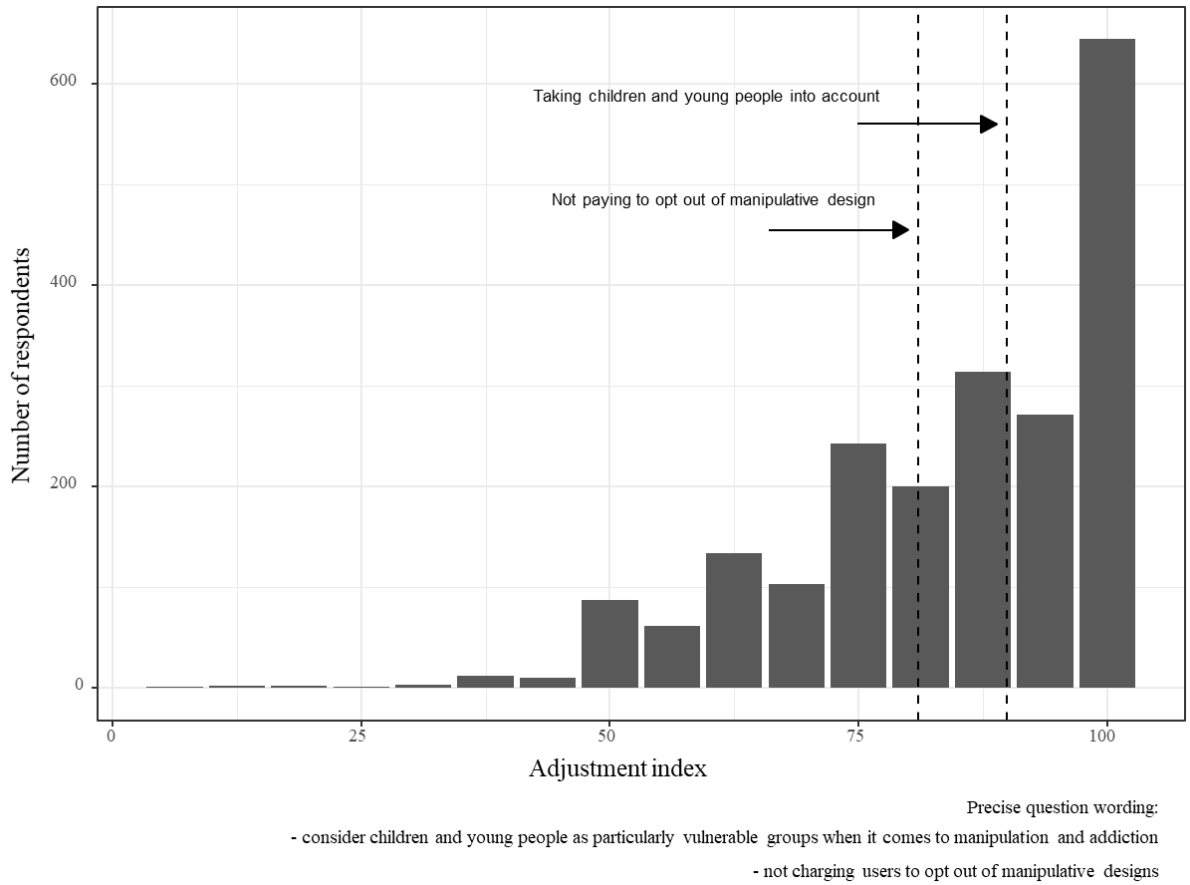


Figure 5 shows that **Danes are generally very positive about the recommendations for regulating tech giants**. Almost 1/3 of all respondents said they "strongly agree" with all four recommendations. In contrast to the assessment of content moderation, respondents' levels of agreement with each type of regulation were similar. Most agree that tech giants need to consider children and young people as particularly vulnerable groups when it comes to manipulation and addiction. Conversely, the question of whether it should cost money to opt

out of manipulative design is the one with the least support. However, both questions average high on the index, which indicates two things. Firstly, Danes generally seem favourably disposed towards regulation. Secondly, the differences are not about what type of regulation Danes want but rather whether or not they want more regulation.

Danes' ability to identify AI-generated content

The discussion about people's ability to identify AI-generated content has gained momentum after the introduction of ChatGPT, where the general public has gained experience and exposure to generative AI. One democratic threat that has been discussed is whether images, audio clips, and videos created with generative AI can help fuel the spread of misinformation and ultimately trick people with realistic but false content.

While the research described earlier has been interested in the question of whether people can identify misinformation, there are two significant gaps in our understanding of whether people can distinguish between true and false information. Firstly, it is questionable whether many research findings show people's ability to *identify* misinformation or, instead, people's ability to *remember* whether something has been written about in the media. Secondly, there is little research on whether people can distinguish between real and AI-generated images.

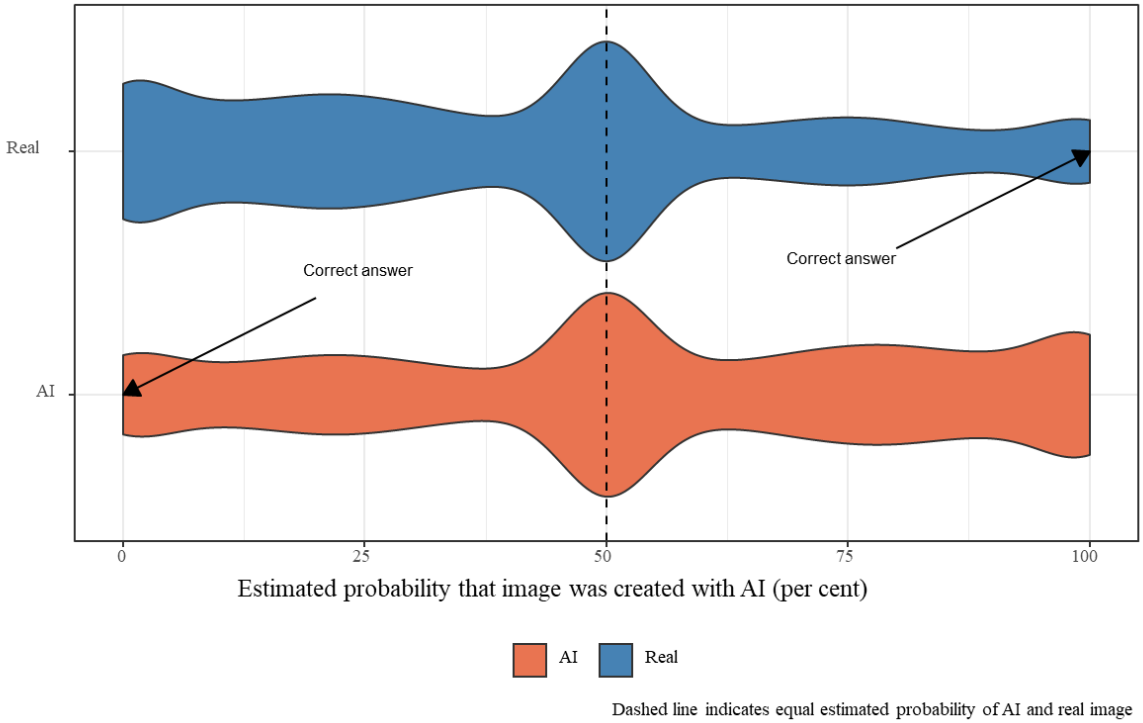
We designed an experiment to assess how good Danes are at judging the authenticity of the images they are exposed to. Each respondent was asked to rate seven images according to how confident they were that each image was created with generative AI on a scale from 0 to 100. All the images depicted actual events, which the respondents were also made aware of. However, whether a respondent was exposed to the actual or AI-generated version of the image is random. A detailed description of the design can be found in Appendix 3C.

Since generative AI will most often be used to create fake scenarios, it is a limitation of our study that the scenarios we present to respondents are actual events. However, this is a necessary prerequisite for real images to be compared to AI-generated images, which eliminates the effect of a respondent remembering the event.

How respondents rated the authenticity of the images they were presented with is shown in Figure 6. The x-axis of the figure shows the respondents' assessment of how likely they think it is that the image was created with AI. A value of 100 indicates that the respondent feels absolutely certain that the image was created with AI. The y-axis of the figure separates the two types of images that have been presented: real and AI-generated images, respectively. The thicker the distribution, the more respondents there are on the scale.

Firstly, Figure 6 shows that **most Danes often feel unsure whether an image was created with AI**. This is shown by the distribution being thickest around the value of 50, indicating that the respondent considers it just as likely that the image was created with AI as it is real.

Figure 6: Danes are often in doubt about the authenticity of images but mostly guess in the direction of the correct answer



Secondly, the figure shows that **most guesses are in the direction of the correct answer but that many guesses miss the mark**. The blue distribution is thicker towards the right, which correctly indicates that the image is created with AI. In comparison, the red distribution is thicker towards the left, which correctly indicates that the image is real. However, many guesses in the opposite direction indicate that many Danes not only often feel doubtful about the authenticity of the image but are also often deceived.

Thirdly, it appears that **Danes are slightly better at correctly identifying real images than AI-generated images.** This is shown by the thickness of the blue distribution on the left side being thicker than the red distribution on the right side, indicating the wrong answer to the presented image. This indicates – if not determines – that Danes are less likely to guess wrong because they are afraid that what they see is false (*tainted truth*) and more likely to guess wrong because AI-generated images look real.

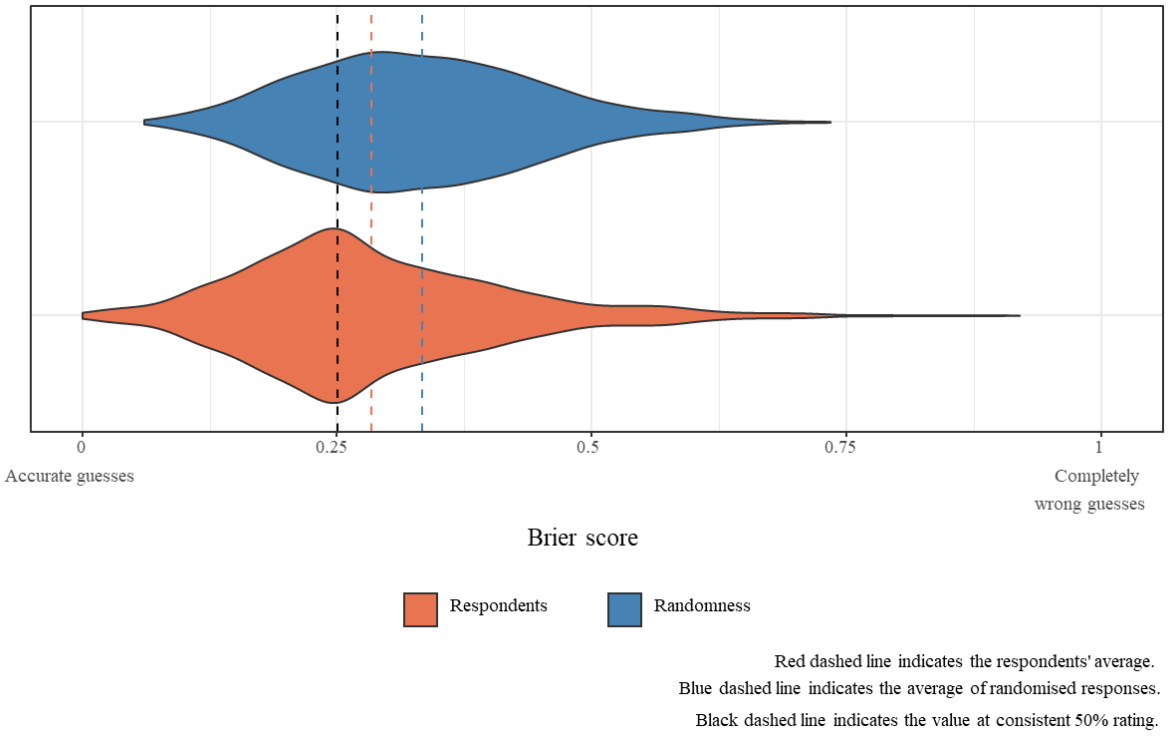
To evaluate whether Danes are generally able to assess the authenticity of images, we use so-called *Brier scores*. This calculation has the advantage of better assessing each respondent's overall ability to identify the authenticity of images. It is also a widely used measure to evaluate predictions and similar items. Brier scores calculate the relationship between the probability the respondent has given the image to be true or false and whether or not the image was actually created with AI. To understand the logic of the scores, imagine that it rains 50 per cent of all days. If the weather service states in its forecast every evening that it will rain with a 50 per cent probability the next day, the forecast will, on average, be completely correct – but not very useful. It would be much more informative if the weather service correctly predicted with certainty every evening whether it will rain the next day. Brier

scores take into account both whether guesses are correct on average and whether they are expressed with certainty, capturing this important distinction.

In this hypothetical example, if the weather service predicts a 70 per cent probability that it will rain tomorrow and it actually does rain, the weather service has made a good prediction and gets a Brier score of 0.09. If it doesn't rain, the weather service hasn't been as good, with a score of 0.49. A Brier score of 0 indicates that a respondent was completely confident in their answers and got it right every time. In contrast, a Brier score of 1 indicates a respondent who was utterly confident in their answers but got it wrong every time. When interpreting Brier scores, it's important to remember that *lower* Brier scores indicate *more accurate* guesses.

The respondents' ability to identify AI images can be seen in Figure 7. The figure reads that the thicker the distribution, the more respondents are placed there on the Brier score. The closer to 0 the distribution is, the better the respondents are at identifying the authenticity of the images.

Figure 7: Danes don't do significantly better than randomness in rating images



Firstly, Figure 7 shows that **Danes are not significantly better than random guessers when identifying the authenticity of images**. The red distribution shows how well the respondents in the questionnaire did, while the blue distribution shows what the result would have been if all images had been guessed entirely randomly. As you can see, the red distribution is slightly further to the left than the blue distribution, showing that respondents are doing somewhat better than if they had just guessed randomly. This can also be seen by the dashed lines, where the red dashed line indicates the respondents' average, and the blue dashed line indicates the random average. However, the average for respondents is slightly above the value 0.25 (marked with a black dashed line), which is the value a respondent would get if the probability was consistently assessed at 50 per cent. On average, respondents would have

done better if they had avoided guessing what they believed and instead opted for the conservative centre position.

It's important to remember that, if anything, we would expect the respondents in the questionnaire to perform *better* than they would in a real-life scenario, for example, when viewing social media. As part of the survey, respondents were informed that some of the images were created with AI. They spent more time looking at the images (median = 7 seconds) than they would have if they had seen them in their Instagram *feed*. Furthermore, the study's images were created relatively quickly using publicly available image-generation programmes. We would expect disinformation campaigns by professionals or state actors to be of even higher quality.

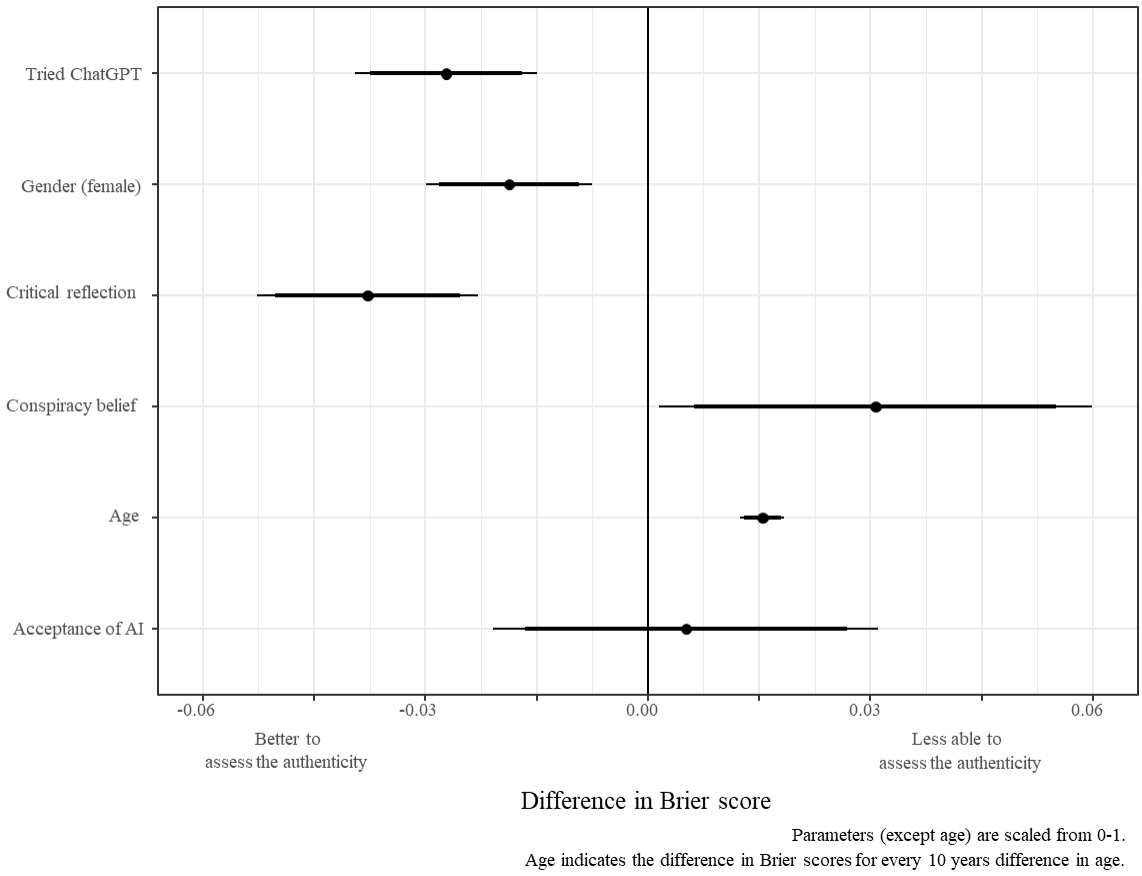
Secondly, Figure 7 shows **a big difference in how good Danes are at identifying the authenticity of images**. This can be seen by the fact that the red distribution stretches far. There is a smaller group that is close to the value 0 and is, therefore, very good at identifying the authenticity of the images. Next, the black dashed line marks a large group around the 50 per cent estimated probability level. Finally, a medium-sized group, including those to the right of the blue dashed line, performed worse than if they had answered randomly.

As Figure 7 shows, respondents' abilities to identify an image's authenticity vary greatly. Figure 8 shows the differences that characterise those who are good and those who are less good at assessing the authenticity of images. The figure is read so that the dot indicates the difference in Brier scores, while the lines indicate the confidence intervals of 90 per cent and 95 per cent, respectively. If the lines overlap 0, marked with a black line, there is no reasonable certainty that there is a difference in how good people are.

Figure 8 shows that **women who have tried ChatGPT and those with a higher level of critical reflection are better at assessing the authenticity of images**[7]**.** At the same time, the figure shows how **older people and those more likely to believe in conspiracy theories are less able to assess the authenticity of images**. However, there is no difference in how good Danes are at determining the authenticity of images depending on whether they think it is acceptable to use generative AI on social media.

[7] The coefficients in Figure 8 reflect different model specifications described in Appendix 3D from a corresponding *Directed Acyclic Graph* (DAG). All coefficients have the same sign and significance in bivariate specifications, except that "Acceptance of AI" becomes significant and negative in a bivariate model.

Figure 8: Differences in who is good at assessing the authenticity of images



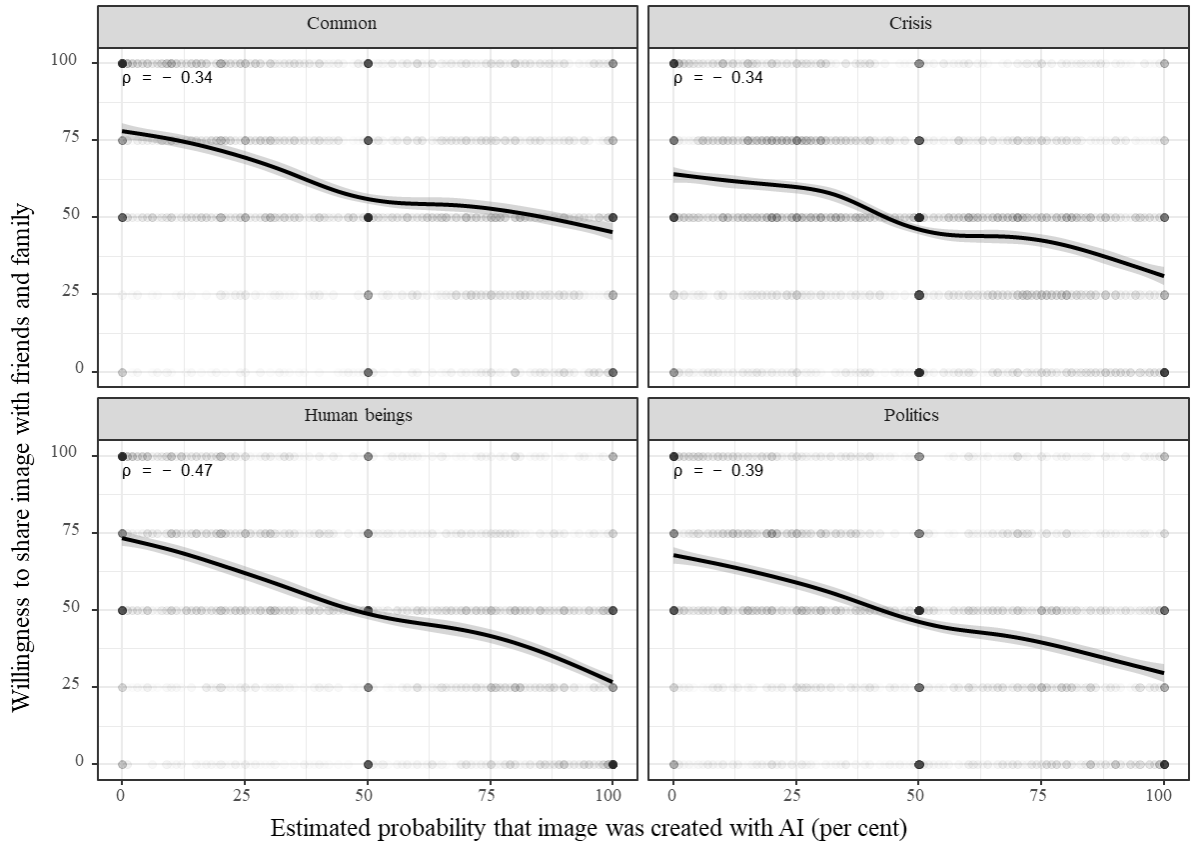
Overall, Figure 8 indicates that there are certain social groups we should be concerned about when it comes to the risk of being deceived by an AI-generated image. However, some of the factors might be used proactively: If a person has tried their hand at generative AI and thus has a sense of what such content looks like, it may be possible to help the rest of the population identify similar content. A parallel can be drawn to the literature review's focus on *inoculation* techniques, where exposure to, e.g. *deepfakes* in small doses and controlled settings may teach people to identify the techniques malicious actors are trying to use to deceive them.

However, the figure also indicates that education on what AI is (often labelled *digital literacy*) and what it can look like is no magic potion. Older citizens, perhaps due to less life experience with IT, may find it more challenging to identify issues, even if they have help. Similarly, people who are more likely to believe in conspiracy theories may find it difficult to let go of their dissident worldview when someone tries to convince them that images they believe are covering something up are not true. The results suggest that diffuse scepticism cannot replace specific abilities to distinguish between true and false information and may even do more harm than good.

In addition to exploring how well people identify the authenticity of images, we asked respondents to consider how comfortable they would be sharing each image. This way, we can investigate whether there is a correlation between how comfortable Danes are with sharing an image and whether they believe the image was created with AI.

Figure 9 shows how **Danes are less comfortable sharing images they believe are created with AI**. This is shown by the slope of the line being negative for all types of images. The correlation is particularly strong regarding images showing people, with respondents being quite comfortable sharing images they believe are real but very uncomfortable when they think the image is AI-created.

Figure 9: Danes are less comfortable sharing images they think are created with AI



The figure also shows that Danes are generally more concerned about sharing images depicting crises and politics, even if they believe the image is real. This may be because, with even a slight fear that the image could be AI-created, the consequences of sharing an AI-created image of a forest fire are perceived to be greater than those of an apple tree.

The opposite is true for images that show ordinary things, such as an apple tree or a cow grazing. Most people are neither uncomfortable nor comfortable sharing the image with friends and family, even if they think the image is AI-generated.

The importance of addressing misinformation

Whether the way we talk about misinformation has negative repercussions may seem like a very theoretical question. However, it's easy to imagine the negative consequences of people walking around worried about misinformation and whether it affects democratic elections. One of the consequences that has been formulated is the risk of a so-called *'liar's dividend'* (Chesney & Citron, 2019). The idea is that it is easier for dishonest people to dismiss true

information as false by claiming that the image, video, or audio clip was created with AI. In this way, the public pays the liar an interest as the price of keeping their ears open to misinformation.

As a thought experiment, imagine if it was 2024 that the damaging *Access Hollywood Tape* was released, which, shortly before the 2016 US presidential election, showed Donald Trump making statements about how a celebrity can easily make sexual advances towards women (Fahrenthold, 2016). With a population that has all heard of *robocalls* voiced by Joe Biden, Russian disinformation campaigns and a wealth of convincing AI images and videos, it's not inconceivable that the counter-argument about the authenticity of the tape would be easier to make in 2024 than in 2016.

Therefore, the third theme of the questionnaire tests whether we should be worried about the negative democratic consequences of speaking alarmist about misinformation. We follow the design of Jungherr & Rauchfleisch (2022), who conducted a similar experiment in the US. This involves exposing different groups to different statements about how big a problem misinformation is and then asking respondents about several statements, such as how worried they are about misinformation affecting democratic elections. Specifically, we randomly divide the respondents into three groups, with one group receiving a statement that misinformation is not really a societal problem ('dismissive'), another that whether misinformation is a societal problem is uncertain ('balanced'), and a third that misinformation is the biggest democratic challenge in the short term ('alarmist'). A detailed description of the experiment can be found in Appendix 3E.

Specifically, we examine the effect of presenting an alarmist awareness of misinformation on respondents' behaviour:

- Support for democracy as a form of government
- Fear of being accused of spreading misinformation
- Fear of being exposed to misinformation
- Fear that AI and misinformation have negative effects on democracy

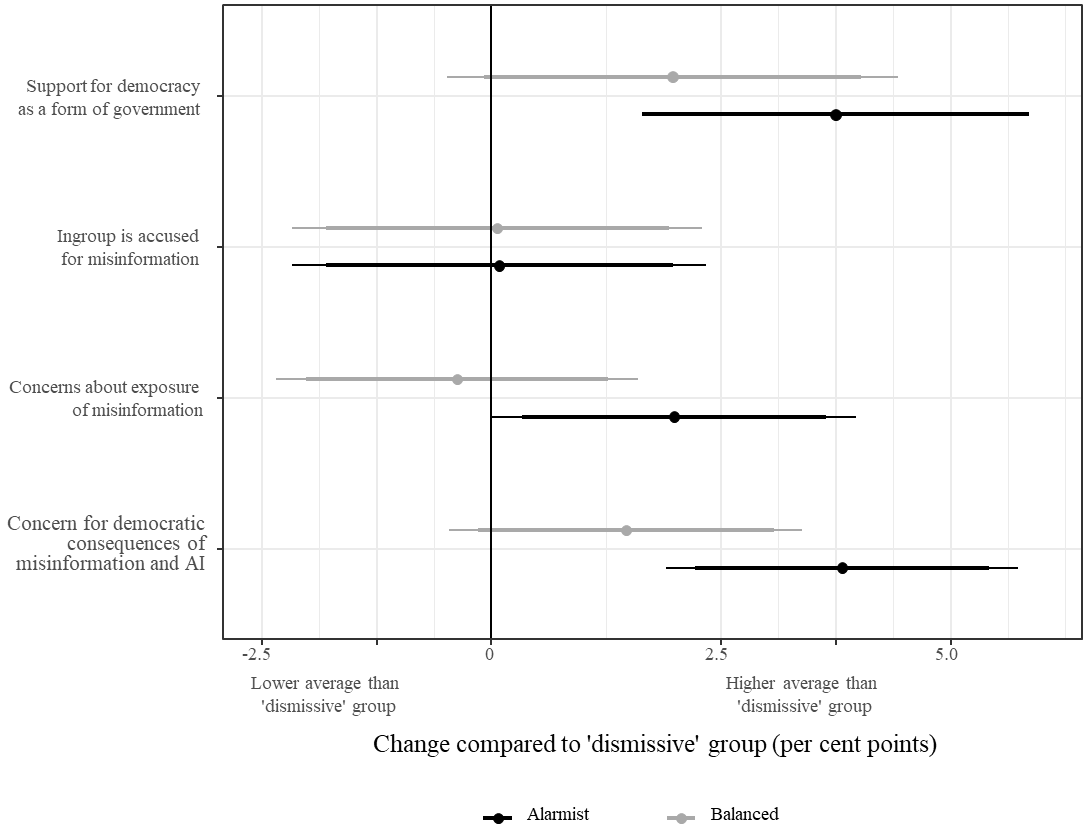
All four elements are constituted by two to three questions, all of which can be found in Appendix 3E.

Figure 10 is a so-called coefficient plot. This means that the figure shows whether there is a difference between what respondents in the different groups answer and whether we can say with reasonable certainty that this difference is not due to chance. The points in the figure are interpreted as differences from the group that has been told that misinformation is not a particular challenge (the 'dismissive' group). For example, let's look at the top row of the figure. The group that received the alarmist description (black dot and line) is approximately 3.5 percentage points more positive towards democracy as a form of government than those who received the dismissive description. Those given the balanced description (grey dot and line) are about 1.6 percentage points more positive towards democracy as a form of government. Still, as the line overlaps 0 (black vertical line), we can't say with reasonable certainty that it's not just chance playing tricks on us. From this, we can deduce that an alarmist description seems to make people more favourable towards democracy as a form of government than a dismissive description. However, the opposite is not true with a balanced description.

Figure 10 shows how **an alarmist interpretation of the challenges of misinformation impacts how Danes perceive important issues for democracy**.

On the negative side, Danes are becoming **more concerned about being exposed to misinformation and the impact of misinformation and AI on democracy and democratic elections**. This is negative as it can increase the risk of dishonest people finding it easier to cheat and get out of scandals and people losing trust that what they see and hear is true. The fear that misinformation can affect democracy and democratic elections can seem particularly dangerous. At its extreme, this could mean easier conditions for politicians who won't accept election defeats, citing AI manipulation and misinformation. While such a scenario is admittedly less likely in a Danish context, such fears can fuel distrust in politicians, the media, and other societal actors.

Figure 10: An alarmist interpretation of the problems of misinformation increases fear of misinformation and its effects but also increases support for democracy



On the positive side, Danes express **greater support for democracy as a form of government**. It is worth noting that this goes against our theoretical expectation of finding the same pattern reported in the US by Jungherr & Rauchfleisch (2022), so extra care should be taken when interpreting this. The difference between our results and the US experiment may be that Danes trust societal institutions, politicians, and the media more. Danes may interpret an alarmist interpretation of misinformation as a problem as a response to a societal challenge. It increases people's confidence in democracy's ability to identify and address societal issues. The assumption is that before the experiment, Danes have been exposed to various descriptions of misinformation challenges from journalists and politicians in their daily lives,

making them sceptical when they are told that misinformation is not a problem and satisfied when they hear that someone is responding. However, further investigation will be required to determine such an explanation.

Another important implication from the experiment is that there **are no differences between the group exposed to a balanced description of misinformation and the group exposed to a dismissive description of misinformation**. This lack of difference can be seen as positive because the balanced description nuancedly outlines the challenges of misinformation without exaggerating or dismissing the problem. Therefore, the lack of impact means that a balanced way of communicating will give Danes accurate and nuanced information while avoiding negative repercussions.

Reviews

Based on the results of the survey, we have the following overall assessments:

Older people are particularly vulnerable to the risk of being scammed by content created with generative AI. While the debate often focuses on the risks to children and young people, our results also show the importance of focusing on older people and possibly initiating processes with NGOs that work with them to strengthen their ability to navigate online.

Education may strengthen Danes' ability to distinguish between real and AI-generated content. Those who are better at critical reflection and those who have tried generative AI are generally better at distinguishing between real and AI-generated content. This indicates that it may be possible to empower Danes through education on generative AI and what to be aware of as a citizen. Such efforts are in line with the part of the literature review that showed that people can be 'immunised' against misinformation by being exposed to small amounts in controlled situations, which can improve their ability to identify misinformation in the future. However, similar to vaccinations against diseases, it's crucial to think about how to ensure lasting effects. It is important to develop people's awareness and ability to critically reflect to prepare them for future developments. Generative AI is expected to improve over time, so a continuous and general focus on understanding technology is crucial when discussing training initiatives.

Diffuse scepticism is not beneficial and is potentially counterproductive. Related to the assessment of the importance of education is the issue of scepticism. Here, we see how experience and critical thinking positively correlate with the ability to identify the authenticity of content. Conversely, those with higher levels of belief in conspiracies are generally poorer. High levels of belief in conspiracy theories indicate a general distrust, but such distrust can potentially impair Danes' ability to assess the authenticity of images. Therefore, communication, initiatives and training must generate experience and reflection rather than diffuse scepticism.

It's important not to use alarmist language and to remember that Danes can relate to nuanced information. Our results clearly show how Danes react when exposed to alarmist messages about the dangers of misinformation. However, such reactions can be negative to the extent that malicious actors are allowed to cheat, citing people's general fear of

misinformation and artificial intelligence. However, talking about the dangers of misinformation in a balanced and nuanced way has no adverse effects. We also know from studies during the COVID-19 pandemic that transparent information from the authorities increases Danes' trust in them and lowers their belief in conspiracy theories (Petersen et al., 2021).

Danes have a generally high appetite for content moderation and regulation, but significant differences exist between existing and desired content moderation. When asked to rate their satisfaction with existing content moderation policies, Danes are generally satisfied, with some exceptions. Most obviously, there is the question of whether nudity on social media is acceptable. The responses give the impression that Danes generally do not enjoy being subject to an American moral code of nudity. However, based on existing data, we cannot say anything about how important it is to Danes. In terms of regulation, all proposals presented were generally well supported.

Survey mission and organisation

The Media Agreement for 2023-2026 includes an agreement on an annual independent report on tech giants' impact on democracy, well-being, and societal cohesion in Denmark. This report is the first in the series.

The report's purpose is to form the first foundation for collecting, structuring, and producing knowledge about tech giants' influence on society, thereby strengthening democratic control over them. Over the years, the reports can focus on different and changing themes, depending on technological developments, the latest knowledge and research, and the current political focus. This first report focuses on the overall effects of tech giants on democratic discourse.

The report is prepared by researchers from Digital Democracy Centre (University of Southern Denmark), SODAS (University of Copenhagen) and DATALAB (Aarhus University).

Literature list

Ahmad, W., Sen, A., Eesley, C., & Brynjolfsson, E. (2024). Companies inadvertently fund online misinformation despite consumer backlash. *Nature*, *630*(8015), 123-131.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211-236.

Allcott, H., Gentzkow, M., Mason, W., Wilkins, A., Barberá, P., Brown, T., ... & Tucker, J. A. (2024). The effects of Facebook and Instagram on the 2020 election: A deactivation experiment. *Proceedings of the National Academy of Sciences*, *121*(21), e2321584121.

Allen, J., Watts, D. J., & Rand, D. G. (2024). Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science*, *384*(6699), eadk3451.

Akram, M., Nasar, A., & Arshad-Ayaz, A. (2022). A bibliometric analysis of disinformation through social media. *Online Journal of Communication and Media Technologies*, *12*(4), e202242.

Amnesty International (2022). Myanmar: Facebook's systems promoted violence against rohingya; Meta owes reparations. *Amnesty.org*. https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

Aral, S. (2021). *The hype machine: How social media disrupts our elections, our economy, and our health--and how we must adapt*. *Crown Currency.*

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social media+ Society*, *9(1)*, 20563051221150412.

Altay, S., Nielsen, R. K., & Fletcher, R. (2022). Quantifying the "infodemic": People turned to trustworthy news outlets during the 2020 coronavirus pandemic. *Journal of Quantitative Description: Digital Media*, *2*.

Baptista, J. P., & Gradim, A. (2020). Understanding fake news consumption: A review. *Social Sciences*, *9*(10), 185.

Blanchard, L., Conway-Moore, K., Aguiar, A., Önal, F., Rutter, H., Helleve, A., ... & Knai, C. (2023). Associations between social media, adolescent mental health, and diet: A systematic review. *Obesity Reviews*, *24*, e13631.

Bor, A., & Petersen, M. B. (2022). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review*, *116*(1), 1-18.

Bryanov, K., & Vziatysheva, V. (2021). Determinants of individuals' belief in fake news: A scoping review determinants of belief in fake news. *PLoS one*, *16*(6), e0253717.

Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature*, *630*(8015), 45-53.

Børns Vilkår (2024). Børns liv med sociale medier: Hvordan forholder børn sig til videoindhold, influencere og AI-chatbots. *Børns Vilkår*. https://bornsvilkar.dk/wp-content/uploads/2024/05/Boern-og-unges-liv-paa-sociale-medier_enkeltsidet.pdf

Center for Sociale Medier, Tech og Demokrati (2023). *Danskernes holdning til den demokratiske samtale på online platforme*. Center for Sociale Medier, Tech of Demokrati. https://slks.dk/fileadmin/user_upload/SLKS/Omraader/Medier/Tech-center/Undersoegelse_-_Danskernes_holdning_til_den_demokratiske_samtale_paa_online_platforme.pdf

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, *107*, 1753.

Danmarks Statistik (2023). *Danmark bruger sociale medier mest i EU*. Danmarks Statistik. https://www.dst.dk/da/Statistik/nyheder-analyser-publ/nyt/NytHtml?cid=46771

Danske Medier Research (2024). *Annoncemarked*. Danske Medier. https://danskemedier.dk/branchetal-statistik/marked/

Darius, P. (2024). Researcher Data Access Under the DSA: Lessons from TikTok's API Issues During the 2024 European Elections. *Tech Policy.press*. https://www.techpolicy.press/-researcher-data-access-under-the-dsa-lessons-from-tiktoks-api-issues-during-the-2024-european-elections/

Dommett, K. (2023). The inter-institutional impact of digital platform companies on democracy: A case study of the UK media's digital campaigning coverage. *new media & society*, *25*(10), 2763-2780.

DR Analyse (2024). *Medieudviklingen 2023*. DR https://www.dr.dk/om-dr/fakta-om-dr/medieforskning/medieudviklingen/2023

Duffy, Clare (2023). *Elon Musk's X is encouraging users to follow conspiracy theorist Alex Jones after reinstating his account*. CNN https://edition.cnn.com/2023/12/11/tech/elon-musk-x-promoting-alex-jones-after-reinstating-his-account/index.html

Eady, G., Bonneau, R., Tucker, J. A., & Nagler, J. (under udgivelse). News sharing on social media: Mapping the ideology of news media content, citizens, and politicians. *Political Analysis*.

Ecker, U., Roozenbeek, J., van der Linden, S., Tay, L. Q., Cook, J., Oreskes, N., & Lewandowsky, S. (2024). Misinformation poses a bigger threat to democracy than you might think. *Nature*, *630*(8015), 29-32.

Eckles (2023). *thefacebook and mental health trends: Harvard and Suffolk County Community College* https://statmodeling.stat.columbia.edu/2023/08/22/thefacebook-and-mental-health-trends-harvard-and-suffolk-community-college/

Ergün, N., Özkan, Z., & Griffiths, M. D. (2023). Social media addiction and poor mental health: examining the mediating roles of internet addiction and phubbing. *Psychological reports*, 00332941231166609.

Erhvervsministeriet (2024). Grænser for tech-giganternes udvikling og anvendelse af kunstig intelligens: Delafrapportering 2 fra regeringens ekspertgruppe om tech-giganter. *Erhvervsministeriet*. https://www.em.dk/Media/638460136860682710/web_Tech-ekspertgruppens%20afrapportering%20vedr.%20AI%20MF_v01_140324.pdf

Fahrenthold, D. (2016). Trump recorded having extremely lewd conversation about women in 2005. *Washington Post*. https://www.washingtonpost.com/politics/trump-recorded-having-extremely-lewd-conversation-about-women-in-2005/2016/10/07/3b9ce776-8cb4-11e6-bf8a-3d26847eeed4_story.html

Ferguson, C. J. (2024). Do social media experiments prove a link with mental health: A methodological and meta-analytic review. *Psychology of Popular Media*.

Garrett, R. K., & Stroud, N. J. (2014). Partisan paths to exposure diversity: Differences in pro-and counterattitudinal news consumption. *Journal of Communication*, *64*(4), 680-701.

Golovchenko, Y., Buntain, C., Eady, G., Brown, M. A., & Tucker, J. A. (2020). Cross-platform state propaganda: Russian trolls on twitter and YouTube during the 2016 US Presidential Election. *The International Journal of Press/Politics*, *25*(3), 357-389.

Golovchenko, Y. (2022). Fighting propaganda with censorship: A study of the Ukrainian ban on Russian social media. *The Journal of Politics*, *84*(2), 639-654.

Guess, A., Nyhan, B., Lyons, B., & Reifler, J. (2018). Avoiding the echo chamber about echo chambers. *Knight Foundation*, *2*(1), 1-25.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... & Tucker, J. A. (2023a). Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, *381*(6656), 404-408.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... & Tucker, J. A. (2023b). How do social media feed algorithms affect attitudes and behavior in an election campaign?. *Science*, *381*(6656), 398-404.

Greenspan, R. L., & Loftus, E. F. (2021). Pandemics and infodemics: Research on the effects of misinformation on memory. *Human Behavior and Emerging Technologies*, *3*(1), 8-12.

Haidt, J. (2024). *The anxious generation: How the great rewiring of childhood is causing an epidemic of mental illness*. Random House.

Hove, M. F. (2024). Hvem målretter dig og virker det? Politikernes brug af annoncer på sociale medier. *Magtudredningen 2.0*. https://ps.au.dk/fileadmin/Statskundskab/Billeder/Forskning/Forskningsprojekter/Magtudredning/Essays/Tema15/Mads_Fuglsang_Hove.pdf

Hove, M. F., Hobolt, S. B., van Dalen, A., & de Vreese, C. H. (2024). Partiernes brug af online politisk microtargeting. In *Partiledernes kamp om midten: Folketingsvalg 2022*. Djøf Forlag.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*(1), 129-146.

Jungherr, A., & Rauchfleisch, A. (2022). Negative downstream effects of disinformation discourse: evidence from the US. *Preprint at SocArXiv https://doi. org/10.31235/osf. io/a3rzm*.

Khalaf, A. M., Alubied, A. A., Khalaf, A. M., & Rifaey, A. A. (2023). The impact of social media on the mental health of adolescents and young adults: a systematic review. *Cureus*, *15*(8).

Kim, B., Xiong, A., Lee, D., & Han, K. (2021). A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions. *PloS one*, *16*(12),

King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American political science Review*, *107*(2), 326-343.

Kulturministeriet (2021). Annonceomsætning 2021. *Kulturministeriet, Mediernes Udvikling*. https://kum.dk/kulturomraader/medier/mediernes-udvikling/publikationer/annonceomsaetning-2021

Leerssen, P., Dobber, T., Helberger, N., & de Vreese, C. (2023). News from the ad archive: How journalists use the Facebook Ad Library to hold online advertising accountable. *Information, communication & society*, *26*(7), 1381-1400.

Lu, Y., & Lee, J. K. (2019). Partisan information sources and affective polarization: Panel analysis of the mediating role of anger and fear. *Journalism & Mass Communication Quarterly*, *96*(3), 767-783.

Nisgaard, Allan (2024). Ekspert om falske kendis annoncer på Facebook: Alle kan blive ramt af den. DR https://www.dr.dk/nyheder/viden/teknologi/ekspert-om-falske-kendis-annoncer-paa-facebook-alle-kan-blive-ramt-af-den

Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., ... & Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, *620*(7972), 137-144.

Hancock, J., Liu, S. X., Luo, M., & Mieczkowski, H. (2022). Psychological well-being and social media use: A meta-analysis of associations between social media use and depression, anxiety, loneliness, eudaimonic, hedonic and social well-being. *Anxiety, Loneliness, Eudaimonic, Hedonic and Social Well-Being (March 9, 2022)*.

Hasell, A., & Weeks, B. E. (2016). Partisan provocation: The role of partisan news use and emotional responses in political information sharing in social media. *Human Communication Research*, *42*(4), 641-661.

Kaushik, D. (2024). Policy Responses To Fake News On Social Media Platforms: A Law And Economics Analysis. *Statute Law Review*, *45*(1), hmae013.

Levendusky, M. S. (2013). Why do partisan media polarize viewers?. *American Journal of Political Science*, *57*(3), 611-623.

Martel, C., & Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, 101710.

McCabe, S. D., Ferrari, D., Green, J., Lazer, D. M., & Esterling, K. M. (2024). Post-January 6th deplatforming reduced the reach of misinformation on Twitter. *Nature*, *630*(8015), 132-140.

Meta (2024). About social issues. *Meta Business Help Center.* https://www.facebook.com/business/help/214754279118974?id=288762101909005

Miller, J., Mills, K. L., Vuorre, M., Orben, A., & Przybylski, A. K. (2023). Impact of digital screen media activity on functional brain organization in late childhood: Evidence from the ABCD study. *cortex*, *169*, 290-308.

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

Murray, Conor (2021). TikTok algorithm error sparks allegations of racial bias. *NBC News*. https://www.nbcnews.com/news/us-news/tiktok-algorithm-prevents-user-declaring-support-black-lives-matter-n1273413

Nanz, A., & Matthes, J. (2022). Democratic consequences of incidental exposure to political information: A meta-analysis. *Journal of Communication*, *72*(3), 345-373.

Nasery, M., Turel, O., & Yuan, Y. (2023). Combating fake news on social media: a framework, review, and future opportunities. *Communications of the Association for Information Systems*, *53*(1), 833-876.

Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., ... & Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, *620*(7972), 137-144.

Newman, N., Fletcher, R., Robertson, C., Arguedas, A. & Nielsen, R. (2024). Reuters Institute, Digital News Report 2024 https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2024-06/DNR%202024%20Final%20lo-res-compressed.pdf

Odgers, C. L., & Jensen, M. R. (2020). Annual research review: Adolescent mental health in the digital age: Facts, fears, and future directions. *Journal of Child Psychology and Psychiatry*, *61*(3), 336-348.

Odgers, C. L. (2024). The great rewiring: is social media really behind an epidemic of teenage mental illness?. *Nature*, 628(8006), 29-30.

Orben, A. (2020). Teenagers, screens and social media: a narrative review of reviews and key studies. *Social Psychiatry and Psychiatric Epidemiology*, *55*(4), 407-414.

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature human behaviour*, *3*(2), 173-182.

Oversight Board (2023). *Oversight Board Overturns Meta's Original Decisions in the "Gender Identity and Nudity" Cases.* Meta's Oversight Board. https://www.oversightboard.com/news/1214820616135890-oversight-board-overturns-meta-s-original-decisions-in-the-gender-identity-and-nudity-cases/

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, *25*(5), 388-402.

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature communications*, *13*(1), 2333.

Petersen, M. B., Bor, A., Jørgensen, F., & Lindholt, M. F. (2021). Transparent communication about negative features of COVID-19 vaccines decreases acceptance but increases trust. *Proceedings of the National Academy of Sciences*, *118*(29), e2024597118.

Pew Research Center (2024). *About half of TikTok users under 30 say they use it to keep up with politics, news.* Pew Research Center. https://www.pewresearch.org/short-reads/2024/08/20/about-half-of-tiktok-users-under-30-say-they-use-it-to-keep-up-with-politics-news/

Quétier-Parent, S., Lamotte, D., Gallard, M. (2023). *Elections & social media: the battle against disinformation and trust issues.* Ipsos. https://www.ipsos.com/en/elections-social-media-battle-against-disinformation-and-trust-issues

Radtke, T., Apel, T., Schenkel, K., Keller, J., & von Lindern, E. (2022). Digital detox: An effective solution in the smartphone era? A systematic literature review. *Mobile Media & Communication*, *10*(2), 190-215.

Reeves, B., Robinson, T., & Ram, N. (2020). Time for the human screenome project. *Nature*, *577*(7790), 314-317.

Sanders, A. K., & Jones, R. L. (2018). Clicks at Any Cost: Why Regulation Won't Upend the Economics of Fake News. *Bus. Entrepreneurship & Tax L. Rev.*, *2*, 339.

Shahzad, K., Khan, S. A., Iqbal, A., Shabbir, O., & Latif, M. (2023). Determinants of fake news diffusion on social media: A systematic literature review. *Global Knowledge, Memory and Communication*.

Stevenson, A. (2018). Facebook Admits It Was Used to Incite Violence in Myanmar. *The New York Times.* https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html

Sunstein, C. R. (2001). *Republic. com*. Princeton University Press.

Thadani, T. (2024). Elon Musk's X accused of bias after pro-Harris accounts labeled as 'spam'. *The Washington Post* https://www.washingtonpost.com/technology/2024/08/07/musk-x-harris-bias/

Thorp, H. H. (2024). Unsettled science on social media. *Science*.

Twenge, J. M., Haidt, J., Joiner, T. E., & Campbell, W. K. (2020). Underestimating digital media harm. *Nature Human Behaviour*, *4*(4), 346-348.

Uldall, Rosa (2024). DF deler deepfake video af statsministeren: tendensen kan være undergravende. *DR.* https://www.dr.dk/nyheder/indland/df-deler-deepfake-video-af-statsministeren-tendensen-kan-vaere-undergravende

Valkenburg, P. M., Meier, A., & Beyens, I. (2022). Social media use and its impact on adolescent mental health: An umbrella review of the evidence. *Current Opinion in Psychology*, *44*, 58-68.

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.

Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it?. *Communication research*, *47*(2), 155-177.

Walter, N., Brooks, J. J., Saucier, C. J., & Suresh, S. (2021). Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis. *Health communication*, *36*(13), 1776-1784.

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication monographs*, *85*(3), 423-441.

Weigle, P. E., & Shafi, R. M. (2024). Social media and youth mental health. *Current psychiatry reports*, *26*(1), 1-8.

Weiss, G. (2024). We now know just how much money Elon Musk's X made after his takeover — and it's a lot less than before his purchase, *Business Insider* https://www.businessinsider.com/x-revenues-plunged-months-after-elon-musk-took-over-report-2024-6

World Economic Forum (2024). Global Risks 2024: Disinformation Tops Global Risks 2024 as Envrionmental Threats Intensify. *World Economic Forum* https://www.weforum.org/press/2024/01/global-risks-report-2024-press-release/

Wu, Y., Ngai, E. W., Wu, P., & Wu, C. (2022). Fake news on the internet: a literature review, synthesis and directions for future research. *Internet Research*, *32*(5), 1662-1699.

Yang, Y., McCabe, S., & Hindman, M. (2024). Does Russian Propaganda Lead or Follow? Topic Coverage, User Engagement, and RT and Sputnik's Agenda Influence on US Media. *The International Journal of Press/Politics*, 19401612241271074.

Zuleta, L. & Burkal, R. (2017). Hadefulde ytringer i den offentlige online debat. *Institut for Menneskerettigheder.*

Appendix

Appendix 1A: Collecting activity and ad data from Facebook

Data regarding the number of daily active users is collected from the Facebook Marketing API. The API is created and maintained by Facebook and gives authorised developers access to data about how many users are active on Facebook within different categories. Thus, we made API calls asking for information about how many active users were on Facebook during different days of the week in September 2024, of which we took an average to avoid the effects of different activities during the week. The information returned includes an estimate of the number of active users, including a lower and upper estimate, which is also visualised in Figures 1 and 2.

For Figure 1, we collected data on the number of active users for each country, while Figure 2 only includes users in Denmark. In Figure 2, we have made API calls for a number of age groups, as well as the number of male and female users within each age group.

Ad revenue data for Facebook and Instagram ads is collected from the Meta Ad Library API (ad library). The data collected only includes adverts about social issues, as these are the only type of adverts that can be accessed more than 12 months back, and they are currently the only adverts for which there is data on the cost in dollars for each advert.

Specifically, we have retrieved information on all social media ads from 1 January 2020 to 1 July 2024 and estimated the monthly revenue by calculating how much of a given ad's budget is placed each month, given that the ad spend is constant. For example, an ad placed from 1 January 2020 to 15 February 2020 with a total spend of DKK 150 will be calculated as costing DKK 100 in January and DKK 50 in February. As estimates of the total ad price for a given day or month are unavailable, it is not possible to consider that some days or months are more expensive than others. However, the challenges associated with this are reduced as we report the development in Figure 3 with local regression. In addition, the development in Figure 3 is also visualised with uncertainty estimates based on the consumption for a given ad being reported in a range, e.g. between DKK 100 and 200. Therefore, the development is calculated based on the average of the spread, while the visualised uncertainty indicates the lower and upper estimates, respectively.

In this report, we have chosen to focus exclusively on Facebook and Instagram and social media adverts, as these platforms and adverts have been open for the longest time and are expected to have the highest data quality. However, the approach can eventually be expanded to include more platforms and other types of adverts. As part of the EU Digital Services Act (DSA), all large online platforms (VLOPs) are required to create publicly accessible and libraries containing all adverts that are active or have been active on their platforms in the past year. For these ad libraries, researchers must also be able to access an API so that data can be pulled directly from the platform and structured. However, the accesses are still new and are being criticised for the quality of the data (Darius, 2024). Therefore, it will – hopefully – eventually be possible to use such access to estimate the total ad revenue on the large online platforms.

Appendix 2A: Identifying and systematising research literature

The identification of research literature was done in three phases. Firstly, we have searched for *review* articles and meta-studies on *Web of Science*[8]. Secondly, we have included empirical articles in major journals that have recently been published and were, therefore, not included in the identified *review* articles. Thirdly, we have shared the individual parts of the literature review (misinformation, polarisation, and well-being) with other researchers at Danish universities conducting research in these fields for input to ensure that the key articles in the field are included.

The literature search focused on capturing key debates and empirical findings within the research literature on misinformation, political polarisation, and well-being in relation to social media. Therefore, a necessary condition for studies to be included in the literature review has been the relationship between social media and the research field in question. This means, among other things, that articles that primarily focus on misinformation related to COVID-19 and only secondarily on social media's impact are not included. Similarly, for delimitation reasons, we choose not to include the part of the literature that specifically focuses on how statistical models and the like can be used to identify misinformation, as this is more about computational methods than the relationship between social media and misinformation. The table below shows an overview of the different steps taken in the first phase of identifying misinformation literature and how many articles have been identified.

Steps	**Number of articles**
Search for articles with the keywords (in title or abstract): AI OR generative AI OR generative OR big tech OR platforms OR social media OR Facebook OR Instagram OR YouTube OR Google OR Twitter OR Pinterest OR LinkedIn OR Snapchat OR TikTok OR Bing AND misinformation OR disinformation OR conspiracies OR conspiracy theory OR conspiracy OR fake news OR false information OR hoax OR fraud OR information disorder OR malinformation OR rumors OR propaganda OR deepfake OR infodemic OR information warfare	14,834
Application of filters + Language (English) and year (2004 - 2024) + Research field (communication science and political science) + Document type (review)	 14,834 4,052 158
Abstract *screening*	**Relevant: 36**

[8] Due to the report's focus on misinformation, this first phase of literature identification has only been conducted for the research literature on misinformation.

	Not the right research field: 32 Not empirical or not reviews: 20 Technical (*fake news detection*): 39 Not primary focus on misinformation: 16 Studies primarily focused on COVID-19: 11 Limited to specific countries: 4
Review articles included after reading	17

Appendix 3A: Attitudes towards content moderation on Facebook

Precise wording in the questionnaire and *community standards* for moderation battery. Wording in *community standards* is marked in italics.

Threats of violence that could lead to serious human or material damage
Threats of violence that could lead to serious injury (less severe violence). We remove such threats against public figures and groups not based on protected characteristics if they are credible, and we remove them against all other targets (including groups based on protected characteristics) regardless of credibility

Images that show very thin people in a way that is associated with eating disorders
Content that depicts ribs, collarbones, mid-thighs, hips, a curved abdomen, or prominent spine or shoulder blades when shared with terms associated with eating disorders.

Images of topless women bathing on a beach, posted with the consent of the women
Images of real naked adults if it depicts: Uncovered female nipples except in the context of breastfeeding, childbirth and the moments following childbirth, medical or health context (e.g. following breast removal, breast cancer awareness or sex reassignment surgery) or as an act of protest.

Content that refers to religious groups as stupid or idiots, for example
Content that targets an individual or group of individuals based on their protected characteristics (race, ethnicity, national origin, religious affiliation, caste, sexual orientation, gender, gender identity and serious medical condition), with: Mental deficiencies are defined as relating to: intellectual abilities including, but not limited to, stupid, stupid, idiots. Education, including but not limited to illiterate, ignorant. Mental health, including, but not limited to, mentally ill, retarded, crazy, insane.

Information that contradicts medical evidence, such as statements like: "An increasing number of vaccinations explains why so many children have autism today"
Incorrect information about vaccines. We remove incorrect information, primarily about vaccines, if public health authorities determine the information is false and likely to contribute

directly to vaccine refusal. They include Vaccines cause autism (e.g. "An increasing number of vaccinations explains why so many children have autism today").

Content that infringes copyrights
Upon receipt of a notification from a rights holder or an authorised representative, we remove or restrict content that: Infringes copyrights. Violates trademark rights.

Principles for selecting themes in moderation battery
Meta distinguishes between "We remove" and "We remove when we have a little more context". We only use policies from the former category to avoid ambiguity.

Meta has six categories in its community standards that describe different types of moderated content. 1) Violence and criminal behavior, 2) Satefy, 3) Objectionable content, 4) Integrity and authenticity, 5) Respecting intellectual property, 6) Content-related requests and decisions. We've included examples for all but point 6, as this does not include content that is removed without further context.

There is a wide variation in the severity and intrusiveness of what is being moderated. We mix it up, so we get variety in the type of content, but avoid extremes we expect everyone to agree with, such as threats to individuals to kill them.

Appendix 3B: Attitudes towards regulatory recommendations

The regulations presented to respondents are taken from the government's expert group on tech giants, which has recommended them to the Ministry of Industry, Business, and Financial Affairs. These recommendations are divided into four themes, of which we have selected one from each of the themes on *the co-responsibility of tech giants for information credibility on their platforms* and *regulation of unauthorised use of copyrighted material by tech giants*. In addition, we have selected two questions from the theme of *protecting children and young people from harmful use and the development of artificial intelligence on the services of tech giants*. We did not select any from the theme of *tech giants' market dominance in AI development* as we considered the recommendations too general to capture opinions on this in a questionnaire. Respondents have been asked the following questions:

How much do you agree or disagree that social media should...
- Declare content generated by artificial intelligence
- Consider children and young people as particularly vulnerable groups when it comes to manipulation and addiction
- Document that they do not infringe copyright
- Not charging users to opt out of manipulative designs

Appendix 3C: Design of AI image recognition experiment

The experiment is designed to measure how good individuals are at distinguishing which images are real, and which are created with generative AI. To avoid respondents guessing correctly by accident, all respondents are exposed to seven images. Randomisation for whether the image is real or created with generative AI varies for each image.

Each respondent is exposed to an image that either shows a real or AI-generated image of the same situation. We use stimulus *sampling* to avoid unintended stimulus effects of a single image being easily recognisable or people being better at identifying AI in certain situations (e.g., humans). It is a logic where the stimulus is randomly selected from a pool of stimuli.

The pool of images consists of 52 images, half of which are AI-created and half real. Thus, for each situation, there is both a real and AI-created image. The images are further divided into four categories: ordinary, people, crisis, and politics.

"Common" images include images such as an apple tree and the Chinese *bullet train*. "People" includes images of e.g. Jeff Bezos and an elderly woman receiving a vaccination. "Crisis" includes images of crises, such as a plane crash in Mogadishu and a forest fire in California. "Politics" includes images that may bring to mind political situations and discussions, such as Brandenburg Gate illuminated in Ukrainian colours and the Za'atri refugee camp for Syrian refugees. It's worth noting that for some images, people appear in categories other than "people", but these are images where people are not the focus.

Real images are found on wikimedia.org, where only images without *copyright* have been selected. The AI images are created using the free and publicly available openart.ai. Each image took 5-15 minutes to create by typing in a few different *prompts*. The best generated images were then chosen to be included in the experiment. The AI images are created to mimic the situation where smaller players, such as those interested in attracting social media users to their website, create and share images. Therefore, the image quality would likely have been higher if it had been part of a larger government disinformation campaign. However, this only makes our test easier, which is why the result where respondents perform at about the level of randomness is expected to be an upper bound compared to larger and more sophisticated disinformation campaigns.

As an example of a real and AI-created image, the following images are meant to show people observing a solar eclipse on *The Mall* in Washington D.C. The image on the left is AI-created, and the image on the right is real.

Before the respondents are exposed and asked to consider whether they think the image they've seen was created with AI and how comfortable they would be sharing it, they are given an introductory text. The text reads as follows:

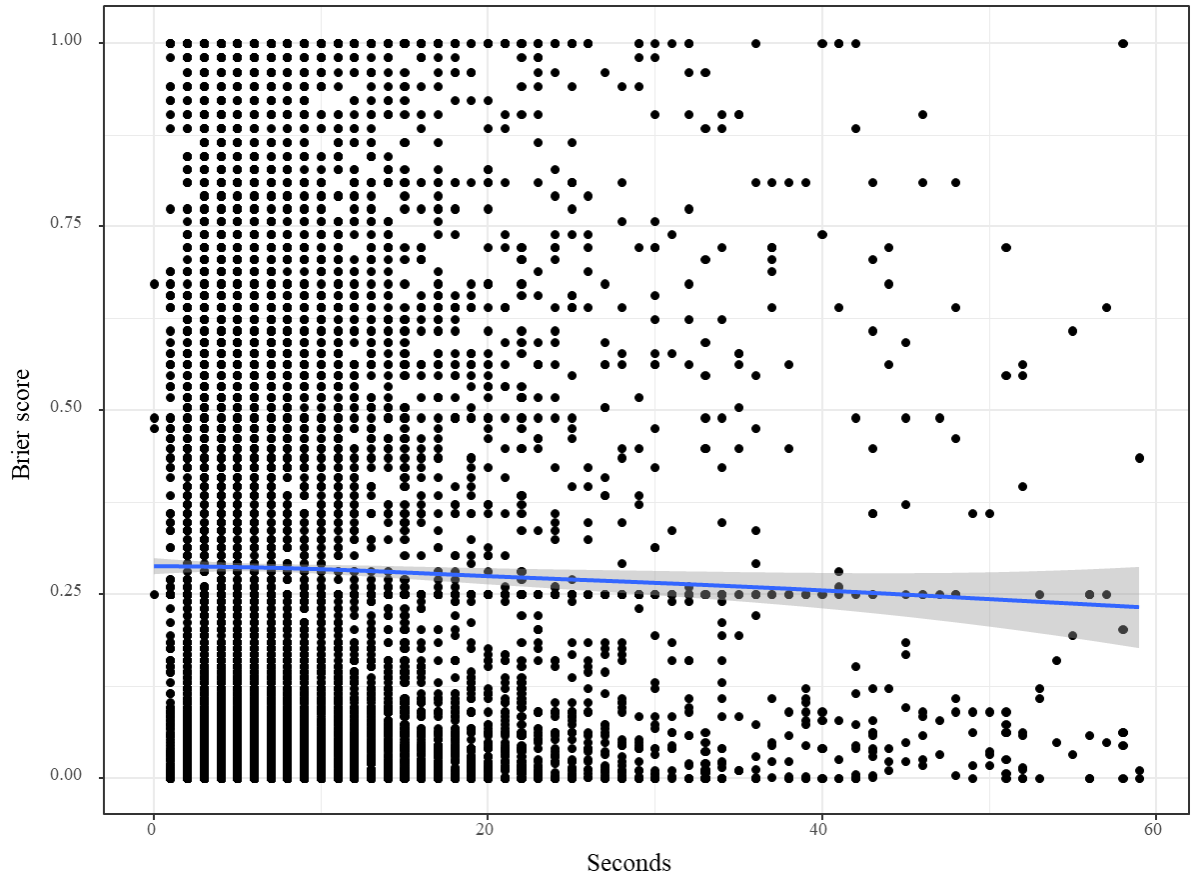
You will now be presented with a series of images showing events or situations that have happened in real life. Some images are actual images, while others are created with generative artificial intelligence ("AI"). You will be asked to give your initial impression of whether you think the image was created with artificial intelligence and how comfortable you would be sharing the image with family and friends.

As shown in the introduction test, we *prime* respondents by telling them that some of the images they encounter are created with AI. However, it's hard to imagine a design where respondents are asked to assess whether an image is created with AI without *priming* them one way or another. This design choice implies that respondents should be *better* at identifying the authenticity of images now that they are aware of it. Since the experiment's result was that respondents performed on average at about the level of randomness, we wouldn't expect them to have performed better had they *scrolled* through their Instagram *feed* without considering whether what they quickly scrolled past was created with AI.

Another hope of the introductory text was that respondents would not spend too much time looking at the images. This is because we are trying to mimic a situation where you are exposed to a series of images, like on social media, and not a quiz on whether you can identify AI images. To determine whether respondents followed the prompt to give their first impressions, we measured the number of seconds each spent on the image page of the questionnaire. Each respondent spent an average of 7 seconds (median) looking at the images. To test the robustness of the results, we ran the analysis only on image tasks in which the respondent spent 20 seconds or less on the page. We find no difference in the results and, therefore, keep all respondents in the main analysis.

However, this doesn't mean that respondents are equally bad at judging the authenticity of images regardless of how much time they spend looking at images. As Figure 3C1 shows, there is a slight negative correlation (0.01 higher Brier score every 10 seconds) between the number of seconds spent looking at an image and how good the image authenticity rating is. The correlation is based on a *pooled* regression. There is so little variation among respondents in terms of how many seconds they spend looking at the images that it is not possible to meaningfully use the variation within units.

Figure 3C1: Correlation between seconds spent on task and Brier score



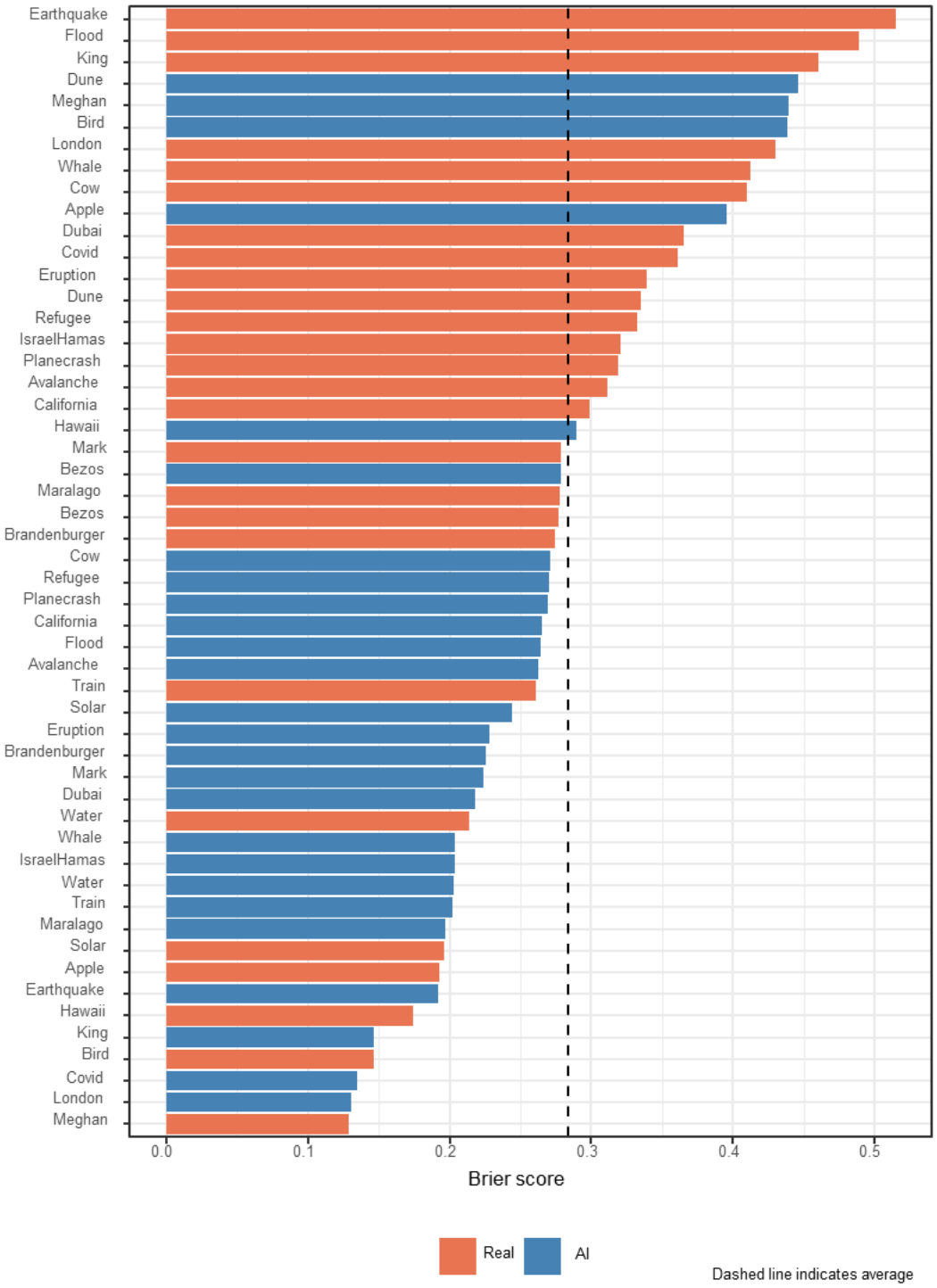
Since the respondents were randomly assigned different images, we can also check how much difference there is across images regarding how easy or difficult they were to assess their authenticity. It can also tell us whether respondents found it more challenging to assess the authenticity of real or AI-created images.

Figure 3C2 shows average Brier scores per image. The figure shows a big difference in how difficult it was for respondents to assess the images, with an AI-generated image of Meghan Markle being easy to assess. In contrast, it was challenging to assess the authenticity of an AI-generated image of a flood.

It also shows how AI-created images generally have high Brier scores, indicating that it is primarily AI-created content that respondents have had difficulty assessing the authenticity of, while they are better at assessing real images.

It is worth noting that despite respondents receiving different images that vary in difficulty, this has neither been possible to assess a priori nor should it give rise to any shifts in the distributions as it rather mimics reality where some AI images are good and others less good. Additionally, the variation does not introduce *bias* in our regressions. This is because differences in severity are random measurement errors on the dependent variable, which does not introduce *bias,* only inefficiency.

Figure 3C2: Brier scores per image

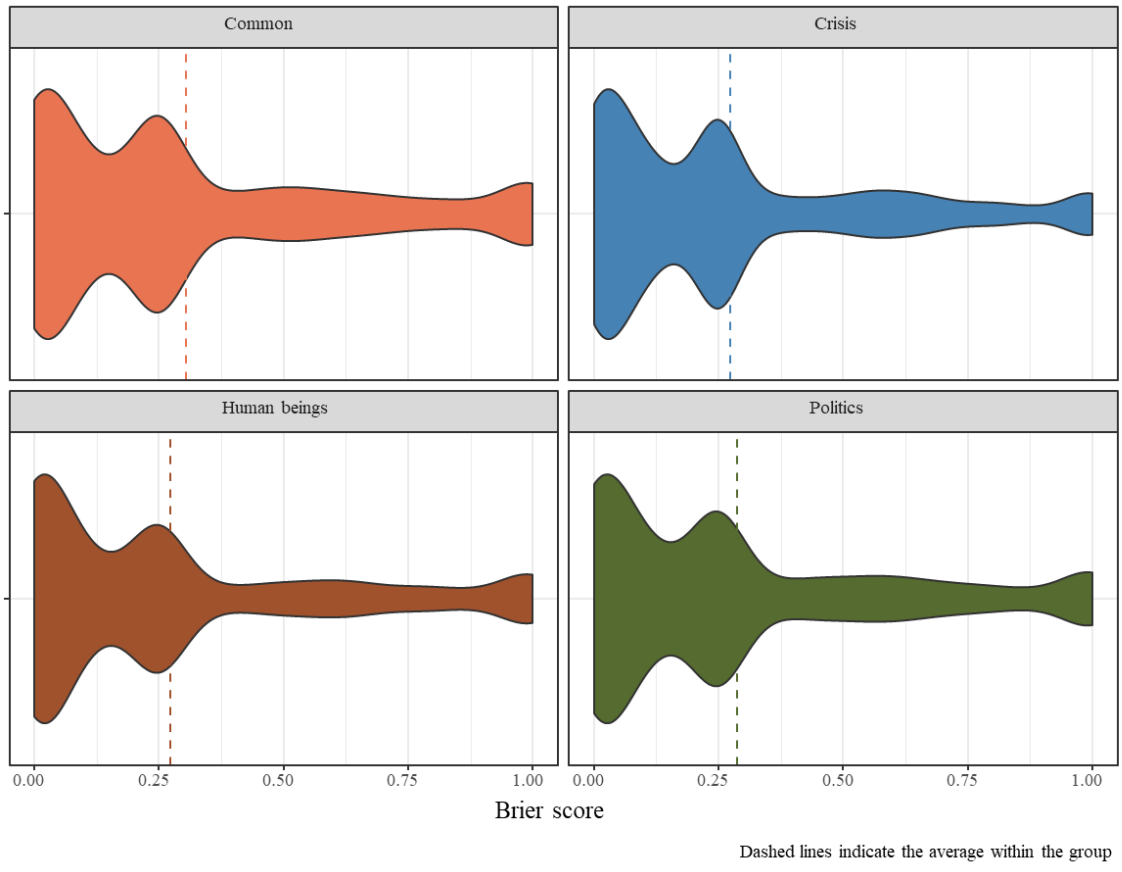


The respondents' ability to identify different types of images depending on their theme (crisis, political, people, or ordinary) is also visualised. This can be seen in Figure 3C3, which shows the distribution of Brier scores per category of images.

The figure clearly shows how Danes have equal difficulty identifying whether an image is genuine, regardless of image category. This is shown by the fact that the four distributions are very similar and that the average, marked with dashed lines, is close to each other. This may seem surprising as images showing people, for example, are also difficult for respondents to read, even though we interact with people daily and are expected to have a pretty good sense of human detail.

The figure thus potentially has another important political implication in that it indicates that it is not easy to create regulatory barriers against specific types of images, such as regulating what AI-generated content can depict due to how difficult it is for users to read. Identifying an AI image of a plane crash seems as difficult as identifying a cow grazing.

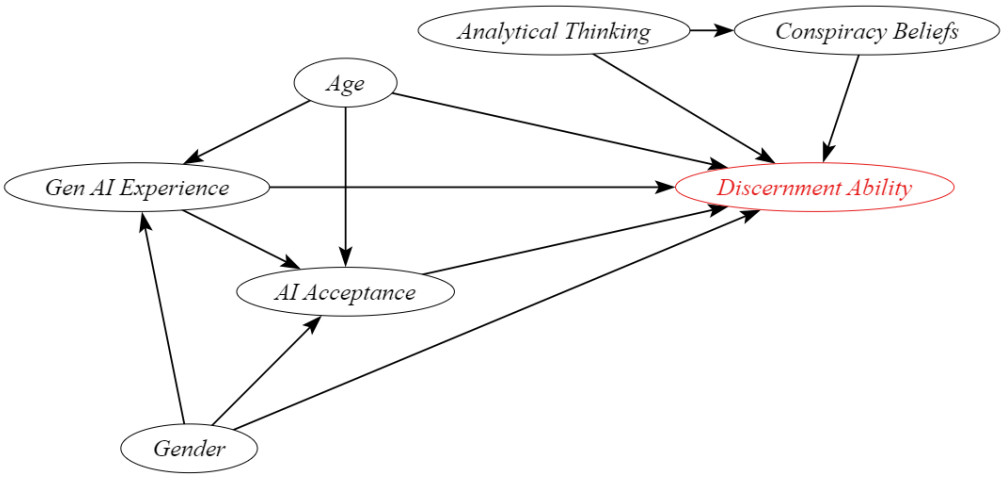
Figure 3C3: Distribution of Brier scores per category of images



Appendix 3D: Specification of regression models of AI image experiment

The differences between stimulus groups are calculated with OLS regression. The models are configured according to the causal diagram below.

This means that two of the models include statistical checks. This applies to the relationship between experience with AI and the ability to identify the authenticity of images, and the relationship with acceptance of AI as an independent variable. The first model controls for the age and gender of the respondent. The latter model controls for age, gender, and experience with AI.



Appendix 3E: Design of misinformation awareness experiment

The experiment is designed to identify the effect of different interpretations of the societal challenges of misinformation. Specifically, whether the level of alarmism affects respondents' fear of being exposed to misinformation, fear of misinformation negatively affecting democracy and democratic elections, fear of being accused of spreading misinformation, and support for democracy as a form of government.

The experiment has three stimulus groups. The following stimuli are assigned to the different groups:

Misinformation: *It's hard to say how significant a challenge misinformation is. Misinformation has proven to be less of a problem in some countries. Research from the US shows that the amount of misinformation is often less than feared. However, little is known about whether the same is true in Denmark or if that situation is changing now that generative artificial intelligence can be used to generate misinformation.*

Balanced information: *It's hard to say how significant a challenge misinformation is. Misinformation has proven to be less of a problem in some countries. Research from the US shows that the amount of misinformation is often less than feared. However, little is known about whether the same is true in Denmark or if that situation is changing now that generative artificial intelligence can be used to generate misinformation.*

Alarmist information: *Fears about the spread of misinformation have increased in recent years, especially with the rise of generative artificial intelligence ("AI") that can quickly and realistically generate misinformation. This is why the World Economic Forum has identified misinformation as the biggest global threat in the short term, especially because almost half of the world's population will vote in democratic elections in 2024.*

Next, all respondents are asked to answer the following questions (scale 1-5 from strongly agree to strongly disagree):

- Information I see on social media is generally trustworthy*
- Traditional news media doesn't share misinformation on social media
- Elected politicians don't share misinformation when posting on social media
- Generative artificial intelligence ("AI") is positive for Danish democracy*
- Misinformation affects who wins elections, for example. to the Danish Parliament*
- Democracies are indecisive and have too much political bickering*
- Having experts instead of politicians making decisions for what they think is best for the country will be a great way to run this country*
- Parties representing people like me are often falsely accused of spreading misinformation
- In political discussions with friends and family, I fear being accused of spreading misinformation

Questions 1-3 are part of an index about the respondent's fear of being exposed to misinformation. Questions 4-5 are part of an index on respondents' fear of negative democratic effects of misinformation and AI. Questions 6-7 are part of an index on the respondent's support for democracy as a form of government. Questions 8-9 are part of an index on the respondent's fear of being accused of spreading misinformation.

* indicates whether the difference between the alarmist and dismissive group is statistically significant at a 95% confidence interval.