Studio Europa Maastricht

Policy Brief

collection

Digitalisation

EU Digital Services Act

STUDIO EUROPA MAASTRICHT

Maastricht University

Table of contents

Introduction to the policy brief collection on digitalisation

Dr Philippe Verduyn

Digitalisation has transformed society. The current digital landscape consists of a vast number of online platforms that are integrated into people's daily lives. For example, more than five billion people use social media to connect with others (1) and spend an average of 2.5 hours on these platforms each day (2). Moreover, many people turn to online marketplaces to treat others (or themselves) to a gift; a phenomenon reflected by Amazon's net sales revenue having reached 575 billion US dollars in 2023 (3). One of the latest illustrations of how digitalisation is changing society is the growing use of generative artificial intelligence, which allows the creation of digital content including text, images, videos or music.

Digital services have made life remarkably easier for many, but there are public concerns that digitalisation also has negative societal consequences. For example, social media can increase the prevalence of disinformation, hate speech, illegal goods, cyberbullying, digital addictions and mental health problems. Additionally, there are concerns about advertising aimed at minors, data leaks resulting in privacy violations and unfair competitive advantages for major online platforms. To address these concerns, enforceable legislation is of key importance.

The Digital Services Act (DSA) entered into full force in the European Union on 17 February 2024 (4). The main goal of the DSA is to prevent illegal and harmful online activities and the spread of disinformation. Furthermore, it aims to ensure user safety, the protection of fundamental rights and the creation of a fair and open online platform environment. The consequences of the DSA are most significant for very large online platforms that have more than 45 million users in the EU, as they are thought to pose the highest risks for society.

The aim of this collection is to provide critical reflection on the advantages and disadvantages of digitalisation and the DSA's role in protecting democracy and the rights and well-being of European citizens. The contributions to the collection are situated within the interplay of online platforms, their users, AI technologies and regulation (see Figure 1). All contributions are relevant for each of these four components, but their foci vary. In each contribution, the role of the DSA is discussed and key policy recommendations are provided.

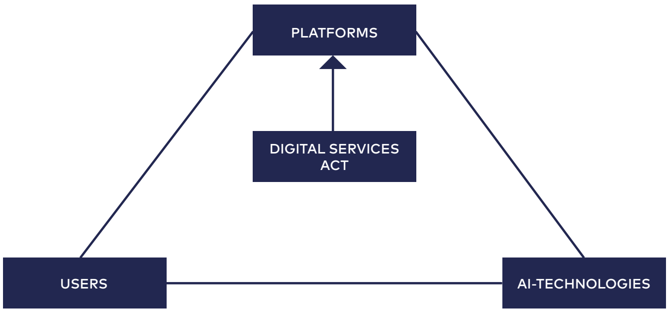


Figure 1. *The impact of the Digital Services Act on online platforms and the associated implications for platform users and integrated AI-technologies.*

The first contribution by *Wyatt* provides an analysis of the **Digital Services Act** itself. She highlights the need for the DSA to take into account the social context in which usage of digital services takes place, as users of these services are not only individual consumers but also members of larger collectives, including as citizens or patients. Wyatt further stresses the need for more inclusive policies as a significant part of the EU population is a non-user of digital services either by choice or by lack of digital skills. She also describes the textual features of the DSA and clarifies some of its key terms.

Two contributions primarily address the **characteristics of online platforms**. *Swierstra* discusses the degree to which the DSA succeeds in its objective to protect democracy from populism. He highlights three key features of online platforms that can foster populism and undermine democracy: the increased speed at which news spreads, that emotionally charged news travels faster than nuanced information, and the absence of gatekeepers and editorial filters. Swierstra illustrates the difficulty for the DSA to deal with these platform features as each one cannot only undermine but also enhance democracy. In their brief, *Gabriels* and *Prebreza* provide a detailed analysis of a very large online platform that is highly popular amongst underage users: Snapchat. They critically discuss the design choices of Snapchat that maximise user engagement but can also negatively affect the privacy and well-being of their users. Furthermore, they describe how policymakers - but also parents and teachers - can help adolescents use Snapchat and other social media platforms safely.

Three contributions focus on **users of online platforms**. *Van Prooijen* argues that the spread of misinformation and conspiracy theories on online platforms is not primarily the result of algorithmic recommendations. Instead, this is chiefly caused by users sharing such content to acquire or maintain social connections with others. Van Prooijen discusses the consequences of this insight for interventions that aim to battle online misinformation and conspiracy theories. *Alleva* explains how usage of social media can influence body image. She describes controlled experiments, but also provides case studies that vividly portray how social media users' body image is negatively impacted by online posts depicting unrealistic appearance ideals. However, Alleva also highlights how exposure to body positivity content on social media can improve body image and discusses the role of the DSA in the protection and enhancement of a positive body image. *Urovi* discusses the right of users of online platforms to know how these platforms handle their personal health data. She provides a case study on health apps which demonstrates that data leaks are highly prevalent, and that users of these apps are often unaware of how their data is collected, processed and shared. Urovi argues for standardised transparency reporting for health data within the DSA framework and proposes key elements that should be present in such transparency reports.

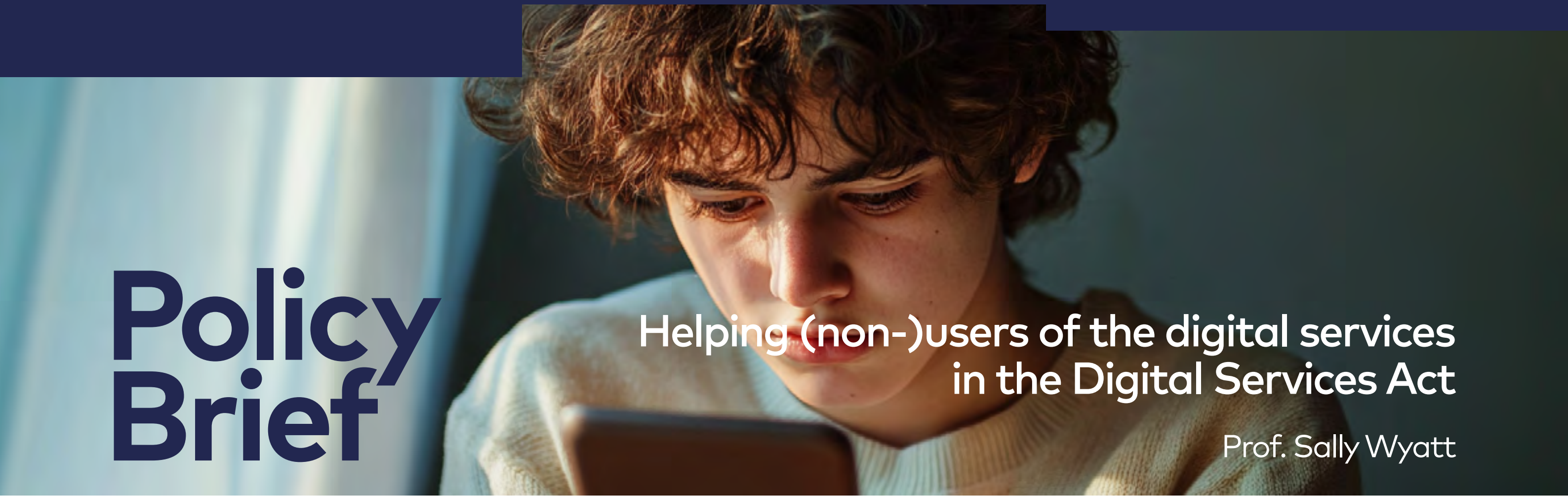
Three contributions examine the role of **AI technologies** and technologies relying on AI in today's society. *Frissen* explains the operating mechanisms behind generative AI technologies that allow the creation of synthetic media (e.g., deepfake videos). Moreover, he describes how synthetic media are being used to create malignant but also benign engineered realities. Synthetic media contribute to the ongoing epistemic crisis, and Frissen offers a number of guidelines that are relevant for current and future EU regulations including the DSA. *Zarkogianni* clarifies the nature of AI technologies that have become an integral part of very large online platforms and search engines; these include generative AI technologies but also recommender systems and information retrieval systems. She describes how these technologies can enhance platform engagement and user satisfaction but also explains the risks of these technologies. To better protect individuals and society from these risks, she argues for a human-centred approach and highlights the need to bridge the DSA with the AI Act that governs emerging AI technologies. In their brief, *Mahr, Heller and Hilken* discuss Extended Reality (XR), which encompasses Augmented Reality and Virtual Reality technologies that are facilitated by AI. They provide an overview of beneficial applications of XR technologies in domains such as education, entertainment and health, but also highlight the possible risks of these technologies. The DSA, but also the associated Digital Markets Act, are of key importance for the future of XR and the authors illustrate how both acts might support or hinder XR innovation in organisations developing and using XR solutions.

Together, the contributions to this collection provide a multidisciplinary perspective on the opportunities and risks of digitalisation for individuals, organisations and society. Furthermore, they explain how the DSA can be utilised to make sure that the opportunities outweigh the risks.

Author information:
Philippe Verduyn is an Associate Professor at the Faculty of Psychology and Neuroscience at Maastricht University. His research focuses on the impact of social media on well-being. Philippe has taken on the role of editor for this policy brief collection.

References

1. Statista. Social media - statistics & facts 2024
[Available from: https://www.statista.com/topics/1164/social-networks/#topicOverview.
2. Statista. Average daily time spent on social media worldwide 2012-2024 2024
[Available from: https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/.
3. Statista. Amazon - statistics & facts 2024
[Available from: https://www.statista.com/topics/846/amazon/#topicOverview.
4. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending directive 2000/31/EC (Digital Services Act).



Policy Brief

Helping (non-)users of the digital services in the Digital Services Act

Prof. Sally Wyatt

Summary

Digital services are becoming increasingly central to the economies of EU member states and to their administrative functioning. The Digital Services Act (DSA) (1) came into effect in early 2024 and aims to hold very large online platforms accountable for the content they post and share with millions of residents and citizens within the EU. The key recommendations do not yet officially cover smaller providers. This means that providers operating primarily in smaller language communities are exempt from the provisions of the DSA, though they are recommended to follow the guidelines.

This has consequences not only for people as consumers of services but also for people as citizens. Many terms are used in the DSA to describe people, including recipient, consumer, person, child, citizen and user. What is striking about this list is the emphasis on people as individuals, and their relationships to private business. With the exception of citizen and person, people using digital services in their roles as patients, passengers and audiences are absent. These collective interests and public values must also be considered.

The interests of non-users of digital services should also be taken into account. People might not use online platforms for a variety of reasons, including physical, cognitive and socioeconomic limits. Non-use might happen because of the fear of harms, especially for women who are subject to misogynistic abuse. Policy-makers need to take seriously their needs and protections.

The DSA is a long, complex document. The EU should provide summaries of its policy documents that are readable by large segments of the population. This is also important for ensuring democratic accountability and engagement.

What is the DSA?

The Digital Services Act (DSA) (1) was passed into European Union (EU) law in late 2022 and applies across all Member States since February 2024. It aims to hold to account 'providers of intermediary services' for the content they post and share via online platforms and search engines. For now, the DSA applies to 19 large providers, each with over 45 million users in the EU. This includes those companies that have become household names over the past twenty years, such as Google, Meta (owner of Facebook and Instagram), LinkedIn, Booking.com and Wikipedia. Smaller providers, which might be important for some EU citizens, especially those operating in and aimed at smaller language communities are expected to apply good practices highlighted by the DSA. However, they are exempt from implementing potentially costly regulations so as not to overburden small businesses.

The major goal of the DSA is to foster safer online environments. Online platforms must implement ways to prevent and remove posts containing illegal goods, services or content. It also bans targeted advertising based on a person's sexual orientation, religion, ethnicity or political beliefs and puts restrictions on advertising targeted at children. With the DSA, people now have the possibility to opt out of recommender systems and profiling, features which have been key to the advertising revenues and business models of large platforms.

The DSA is the latest in a series of efforts by the EU to regulate online content and codify self-regulatory efforts undertaken by online platforms themselves. As Van Hoboken and his colleagues (2) argue, with the DSA, the EU wants to create a safer online environment for all and set standards for the rest of the world about how to do so. They argue that the DSA 'introduces significant legal innovations: a tiered system of due diligence obligations for intermediary services, the regulation of content moderation through terms of service enforcement, systemic risk assessment obligations for the most widely used platforms and search engines, and access to data for researchers' (2, p. 5).

Outline

In this policy brief, I focus on how the DSA thinks about people. How are they described? What terms are used, and what do they suggest about the nature of the relationships between the EU, online platforms and those who use them? For those who are interested, I provide some details of the DSA as a written text (see Box 1), and define two key terms already mentioned above and used in the DSA, namely online platform (see Box 2) and intermediary service (see Box 3). The next section examines how the people who use the services offered by those platforms or providers of intermediary services are described. The section thereafter addresses non-users of digital services who may also be affected by the growing centrality of digital services in access to public and private services. This policy brief concludes with four policy recommendations.

Who are the users affected by the DSA?

In the DSA (1), many words are used to describe its intended beneficiaries. The following terms appear, listed in order of frequency (indicated in brackets after each term, singular and plural forms

Box 1: Textual features of the DSA

The DSA consists of 102 pages of dense bureaucratic text. It has 65,689 words. The so-called readability index is 14.932, suggesting one needs a university degree to be able to understand it. The average sentence length is a staggering 52 words (3). Most English-language newspapers and government agencies aim for 15-20 words per sentence in their articles and policy documents, as there is much evidence to suggest that sentences longer than that are very hard to follow.

Box 2: Definition of online platform

Article 4 of the DSA (1, p. 42-44) includes a list of 24 definitions of terms that are widely used in the text. The definition for online platform is a single sentence of 80 words. I will not reproduce it in full, but it starts as follows: '"online platform" means a hosting service that, at the request of a recipient of the service, stores and disseminates information to the public' (1, p. 43). More usefully, Poell et al. define platforms 'as (re-) programmable digital infrastructures that facilitate and shape personalised interactions among end-users..., organised through the systematic collection, algorithmic processing, monetisation, and circulation of data' (4, p. 3). The DSA definition begs the question of what a hosting service is, though it is included in the definition of our second term of interest, 'intermediary service' (see Box 3).

Box 3: Three types of intermediary services

Companies providing intermediary services, such as Google, Facebook and others, are the target of the DSA. Article 4 (1, p. 42) lists three subcategories: mere conduit service, caching service and hosting service. No examples are given, which would certainly help the reader to make sense of the text. A 'mere conduit service' is something like an internet service provider (ISP). It allows data to be sent and received but the ISP or conduit does not store the data. A 'caching service' stores the data temporarily in order to make transmission more efficient. Most important for my purposes are hosting services. These include web hosting, app stores, social media and business-to-consumer services where products are bought and sold.

have been merged): recipient (271), consumer (83), public (70), person (56), minor (47), individual (40), user (28), society (24), human (17), child (14) and citizen (7). 'User' is often deployed as an adjective, as in user-generated and user-friendly. 'Person' and 'minor' are very much legal terms in this context. 'Citizen' and 'public' relate to people's membership of the demos. The almost invisible 'citizen' in Figure 1 below illustrates how little attention people in their role as citizens is given in the DSA.

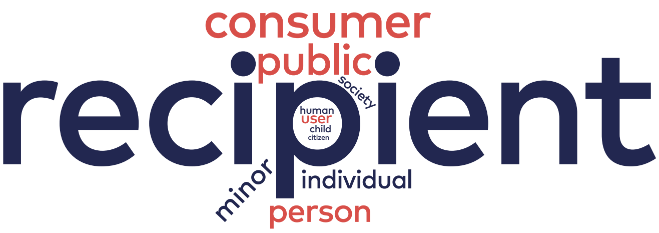


Figure 1 Word cloud of terms used in the DSA to describe people

This focus on the individual, while perhaps useful in consumer law, misses the broader social context in which such services are developed and used. Terms such as 'consumer' and 'recipient' also place people in a relatively passive role, though consumers often organise formally and informally to assert their rights. This focus on the individual who buys services obscures people's social and public roles as voters, residents, patients, carers, passengers, audiences, workers or other collectives. Of course, patients and passengers can also be rather passive, though they often organise to protect their interests. One policy recommendation is to support all kinds of collectives, such as unions, consumer and patient associations, to help their members defend their rights as laid out in the DSA.

There are exceptions to this focus on the individual, as the DSA does occasionally refer to citizens, society and publics. This tends to occur in two ways. The first is when the DSA refers to its objectives of reducing societal risks by protecting basic democratic rights, including freedom of expression, participation in civic discourse and freedom from discrimination. The latter is particularly important for meeting the objectives of the DSA, as most large platforms rely on profiling people, something which is prohibited in the constitutions of many member states. The second is when the DSA refers to the role of platforms in protecting people from criminal activities, such as hate speech and harassment. This is also relevant to those who do not use online platforms, the topic of the next section.

What about non-users?

Including users in discussions about new technologies has advantages, also in the DSA. It represents an important shift away from focusing only on those powerful social actors involved in the design and production of technologies, such as scientists, engineers, industrialists, entrepreneurs, politicians, marketers and financiers. Doing so opens up space to consider how users deploy technologies, in both intended and unintended ways. In the DSA, the focus is precisely on the need to control these powerful actors, referred to variously as online platforms, hosting services and intermediary services (see Box 2); particularly, as already mentioned, the very large ones with more than 45 million users within the EU.

Apart from the problem of focusing on the individual, the concept of user assumes that everyone, eventually will become a user. As I have argued elsewhere (5), this is the so-called addiction model of internet use: 'once a user, always a user'. It reflects a technocratic world view in which technology is central to human progress. Moreover, especially in its adjectival use, it focuses on the human-machine dyad, missing the broader social context in which computers and platforms are used, or not. Using other terms which capture the diversity of people's roles, practices and experiences provides more nuance to our understanding of technologies. Choice of terms should reflect the social context in which technologies are embedded and which crucially affect their functioning. In terms of policy, it is important to recognise that even non-users are affected by digital services, as data about them may be generated, processed and shared by third parties.

Digital technologies are increasingly central to economic and public life. However, large numbers of people remain excluded, through physiological, psychological, cognitive and socioeconomic limits. Even in a rich country such as the Netherlands with compulsory education until 16, two and a half million people are functionally illiterate (including functionally innumerate) (6). Such people might be limited in their possibilities to take part in society, to make full use of digital services and to understand their rights when dealing with online platforms. Nonetheless, such non-users may still be affected by what happens online, given the pervasiveness of such technologies. In many countries, the public sector rightly provides non-digital alternatives so that people can access public services, but data are almost certainly collected, stored and shared about such people, with or without their knowledge and consent. Governments have an obligation to ensure non-users are aware of the protections provided by the DSA.

Other groups may well be capable of using online services, but might resist, either partially on in full. When it comes to self-service in retail outlets, people are concerned that this will become the default, and some resist because they appreciate the personal touch and interactions with other people when shopping or eating out (7, p. 23).

Another example is the evidence that a growing proportion of young women hesitate to participate in online debate, having experienced or witnessed online abuse (8, p. 124). As mentioned above, the DSA wants online providers of intermediary services to protect people from harms and ensure their democratic rights. But as Allen (8) rightly points out, the work of content moderators is difficult, and it is extremely cumbersome to report online gender-based violence. Furthermore, Allen points to the risks of gendered misinformation, such as when journalists and politicians from gendered and racialised minorities are subject to personal attacks questioning their abilities for those roles. Gendered misinformation, according to Allen, 'is based on misogyny but can simultaneously intersect with discrimination based on racism, ableism, religious identity, etcetera and poses a risk to free expression, human dignity and to women's participation in civic discourse' (8, p. 128). These are all intended to be covered by the risk assessment provision of the DSA, but are difficult and costly to implement. Policy measures to support implementation could include providing assistance to feminist and anti-racist groups so that they can help individuals facing discrimination and violence via online platforms.

It is important to consider non-users as a diverse group. Rather than seeing use and non-use as an either/or choice, users need to be conceptualised along a continuum with degrees and types of involvement that may change, and may be voluntary or not. For example, people may occasionally buy goods online or deal with their bank but may not participate on social media platforms. Instead of emphasising people's role as consumers, it may be more beneficial to pay attention to the diversity of users and of domains of use, to practices of use and non-use and to the different roles people have when accessing digital services.

Policy recommendations

1. Digital services are becoming increasingly central to the economies of EU member states and to their administrative functioning. This has consequences not only for people as consumers of services but also for people as citizens. These collective interests and public values must also be considered and supported.
2. Non-users – both voluntary and involuntary – are also affected by the growth of digital services and online platforms. Policy-makers need to take seriously their needs and protections through providing easily available training and working with collectives, such as, but not restricted to, trade unions, consumer organisations and patient organisations.
3. The key recommendations of the DSA should be adapted to cover smaller providers, and those operating primarily in small language groups. This needs to go beyond suggesting them as good practice. Otherwise there is a risk that those people using digital services offered by small language groups will receive fewer protections.
4. The EU should provide summaries of its policy documents that are readable by large segments of the population and not only those with university degrees (see Box 1).

Author information:
Sally Wyatt is Professor of Digital Cultures at the Faculty of Arts and Social Sciences, Maastricht University. Her research delves into the societal dimensions of digital technologies, with particular attention to internet use, social exclusion, and how individuals leverage the internet for health-related information.

Contact: sally.wyatt@maastrichtuniversity.nl

'I am grateful to Jaime Simons and Arnoud Wils from the FASoS PLANT, for help with the content analysis of the Digital Services Act, using Voyant tools and wordclouds.com'

References

1. European Parliament and Council of the European Union. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending directive 2000/31/EC (Digital Services Act). Brussels: Official Journal of the European Union; 2022. 102 p. OJ L277/1.
2. Hoboken J van, Quintais JP, Appelman N, Fahy R, Buri I, Straub M. Foreword. In: Hoboken J van, Quintais JP, Appelman N, Fahy R, Buri I, Straub M, editors. *Putting the Digital Services Act into practice. Enforcement, access to justice, and global implications*. Berlin: Verfassungsbooks; 2023. p. 5-15.
3. Sinclair S, Rockwell G. *Voyant tools*; 2024. https://voyant-tools.org/
4. Poell T, Nieborg D, van Dijck J. *Platformisation*. Internet policy review. 2019; 8(4). doi: 10.14763/2019.4.1425
5. Wyatt S. *Non-users also matter: The construction of users and non-users of the Internet*. In: Oudshoorn N, Pinch T, editors. *How users matter: The co-construction of users and technology*. Cambridge, MA: The MIT Press; 2003. p. 67–79.
6. Netherlands Court of Audit. *Tackling functional illiteracy*. The Hague: Rekenkamer; 2016. https://english.rekenkamer.nl/publications/reports/2016/04/20/tackling-functional-illiteracy
7. Savanta European Computer Compass Q2 2024. Glasgow: Savanta. 2024. 29 p. https://info.savanta.com/Consumer_Compass_Report_Q2_2024.
8. Allen A. *An Intersectional Lens on Online Gender-Based Violence and the DSA*. In: Hoboken J van, Quintais JP, Appelman N, Fahy R, Buri I, Straub M, editors. *Putting the Digital Services Act into practice. Enforcement, access to justice, and global implications*. Berlin: Verfassungsbooks; 2023. p. 121-133.

Policy Brief

The Digital Services Act: not enough to protect democracy against populism

Prof. Tjalling Swierstra

Summary

An important goal of the Digital Services Act is to mitigate democratic risks posed by large online platforms. It aims to curb manipulation and misinformation and to protect fundamental civic rights, such as freedom of expression, media freedom, pluralism and protection against discrimination. However, the DSA is insufficient to protect public deliberation - the heart of democracy - against populism. Populism is an anti-democratic political programme which is partly facilitated by digital media. The DSA is rightly hesitant to limit the freedom of speech for politicians, but this severely restricts what it can do. It is important to acknowledge the limitations of the DSA in this respect, so as to be able to develop complementary policies to fight populism.

Digital media facilitate populism in at least three respects. First, by accelerating the production and consumption of news, New Social Media (NSM) make careful fact checking impossible and feed forms of impatience that are hard to reconcile with democratic procedures. Second, NSM play into our psychological bias favouring emotionally charged news over more bland and complex information. Third, NSM do away with gatekeepers and editorial filters guaranteeing civic norms of democratic deliberation. These three factors pave the road for populism.

The problem is that these democratic risks are intrinsically linked to features of NSM that are positively linked to democratic process. Acceleration allows the public to respond quickly, in real-time if need be, to problems and threats. Democratic activity/citizenship can also be enhanced by appealing to emotions; democracy is and should be a passionate affair. The absence of gatekeepers allows otherwise marginalised voices direct access to the public agora. And the so-called bubbles can enhance democracy by offering easy entrance points for people occupying marginal positions, who can then build their confidence and power in a safe environment.

Policy should aim to mitigate the populist risks of social media by social, legal and technical measures.

- Social measures include instructing people about these risks and investing in diverse offline spaces. In these spaces, citizens can practice and develop democratic skills which can inoculate them to the perverse effects of social media.
- Legal measures include design requirements for algorithms, but also safeguarding and subsidising independent news channels that function as a meeting ground for citizens to discuss the public good.
- Technical measures include designing algorithms to incentivise democratic deliberation, e.g., by safeguarding diversity and by indicating whether opinions are corroborated by facts, as well as smart voting.

Problem: populism threatening democracy

An important goal of the DSA is to mitigate democratic risks posed by large online platforms. It obliges such platforms to remove illegal content and forbids many forms of targeted advertising, including political advertising.[1] By extending greater democratic control over systemic platforms, it aims to curb manipulation and misinformation and to protect fundamental civic rights, such as freedom of expression, media freedom, pluralism and protection against discrimination.

By offering such democratic controls, the DSA aims to counter the threats posed to our democracies by the new (digital) social media. The rise of extreme right-wing populism, as will be argued below, constitutes such a threat.

The DSA must remain neutral vis-à-vis all legal political ideologies and parties, including extreme right-wing populist parties, when these have not been legally banned. But the DSA can address the digital conditions of origin for antidemocratic parties and ideologies in general. It is very questionable whether the EU can or even should require platforms to limit political speech. However, it can incentivise them to create better conditions therefore.

Democracy versus populism

The political scientist Cas Mudde defines populism as 'an ideology that considers society to be ultimately separated into two homogeneous and antagonistic groups, the "pure people" versus the "corrupt elite", and which argues that politics should be an expression of the *volonté générale* (general will) of the people' (2, p. 543). This general will is voiced by authoritarian leaders, posing as the champions of the common people. Populism is also marked by a particular political style. Populists typically pose as outsiders to the 'corrupted' system, always characterise the world in terms of 'crisis', and appeal to moral emotions like anger, fear and pride.

Populism often poses as the paragon of a democratic movement. After all, what else is democracy if not rule by the people? But there is no reason to accept this claim. Of course, democracy is a contested concept and there are many, incompatible conceptions of what a real democracy entails (3, 4). But at the heart of most conceptions of democracy is the idea that it is a political system that allows for the peaceful, non-violent organisation of diverse and conflicting values and interests and that avoids the tyranny of the majority by providing space to minorities. Any such acknowledgment of and appreciation for diversity and minorities is crucially lacking in populism, as it postulates a homogeneous 'people' in basic agreement on core issues opposed only by malicious elites undermining, frustrating and resisting this popular consensus. Affective polarisation, a strong opposition between 'us' and 'them', is in populism's DNA (5, 6, 7). As a result, populism attaches little value to key political practices that constitute and enable democratic systems (4).

Democracy is essentially the peaceful transformation of individual wills through processes like representation, shared deliberation and majority voting into collectively binding decisions (i.e., laws) that aim to realise the common good and are enforced in conformity to the rule of law. Democracy, it is true, is not exhausted by talking and voting. Another important democratic practice is that citizens recognise each other as members of the same civic community regardless of one's opinion. In a true democracy, agreeing with me, or with the leader, or with the majority, is expressly not a criterion for being treated as a fellow citizen with equal rights.

Other democratic practices are resisting (e.g., demonstrations, strikes, civil disobedience) and joining (into political collectives to exert more power, e.g., labour unions). To guarantee that citizens are free and safe to disagree about the common good, a functioning democracy also needs institutions like civil rights, a free press and an independent judiciary. Similarly necessary is a democratic ethos: a willingness to respect one another, to exchange ideas and reasons and to look for solutions that accommodate the interests of as many citizens as possible, including minorities. All this is absent in populism. Rather than diversity, it divides the world in 'us' and 'them', turning opponents into enemies. And enemies are to be fought and ostracised, if necessary with violence. In that sense, populism constitutes a very real and palpable threat to

[1] The DSA is not the only EU regulation serving this aim. There is for example also the Media Freedom Act aiming to safeguard pluralism and media freedom. See: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022PC0457&from=EN. Also, on political microtargeting: (1)

democracy – even if present-day populists pay lip service to democratic institutions and practices as long as they are not yet in power.

If the DSA is to protect democracy against threats posed by large and very large digital platforms, it also needs to discourage populism – irrespective of whether that is populism from the left (focusing on the rich elites) or from the right (focusing on immigrants and the liberal elites). Although the DSA certainly is a move in the right direction, further steps need to be taken.

Causes of populism's rise

There is ample debate about the causes behind the rise of populism, especially in its present right-wing, nationalist form. Current populist parties place their nation first and are inimical towards immigrants (8, 9). Some explain this reviving nationalism as a defensive reaction to the globalisation of the past decades: national political control has leaked away on one hand to international markets, and on the other hand to the often-opaque supranational structures of the EU. Citizens feel that they no longer have a grip on their environment through their national governments (10).

Secondly, neoliberalism prepared the way for populism by increasing income differences and decreasing economic security (11). This has left many worried about their economic status. Furthermore, neoliberal ideology puts a lot of responsibility on the individual; one modelled after a highly romanticised version of an entrepreneur as a strong individual taking care of themselves. As a result, many now long for a stronger, more protective and caring state that defends people against supposedly unfair competition (from people living in other nations, minority groups who get special privileges to redress past wrongs and migrants). Populism responds by promising to privilege its own, native population.

Thirdly, meritocracies have led to a new form of inequality between those with and those without theoretical education (12). In a meritocracy, only one's achievements, as opposed to inherited characteristics like gender, race and family, determine one's position on the societal ladder. It does this by creating equal opportunities for everyone to shine. In practice, they have helped shape economies in which diplomas are a necessary condition for success (13).

Furthermore, to the degree that equal opportunities have been realised, it becomes difficult to blame someone else for one's lack of success. Meritocracy places responsibility for failure squarely on the shoulders of the individual. This means that the lower socioeconomic classes can have trouble in keeping their self-respect (14). All populist parties promise to restore this self-respect by sowing distrust towards the supposedly contemptuous, educated elites, and stressing the worth and honour of ordinary, hard-working, honest people to whom they promise to give political power back.

There are other elements that can be mentioned to explain the surge of (right-wing) populist parties. For example, for decades technocratic governments have given citizens the impression that there is no alternative to the policies proposed by their experts (15), and that there has been a weakening of the ideological ties between political parties and the electorate. Additionally, a wide-spread feeling of malaise prevails in the West that it is losing its traditional privilege of (hard and soft) power. Furthermore, there is a sense that economic growth and progress are being thwarted by ecological boundaries. The COVID-19 measures further exacerbated a distrust in government.

And there is the explanation that is most relevant here: the rise of populism can be causally connected to digitalisation in general and to the dominance of the new social media in particular (16). This is where the DSA can make a difference. The question is: is that difference big enough? If not, what measures can we think of to make anti-populist policy more effective?

What the DSA can do

There is a strong correlation between digitalisation of the media, manipulation and the circulation of misinformation (16). There is ample evidence of deepfake news that is circulated on the Internet to disrupt democratic elections and deliberation; and the rise of populism is linked to, although not simply caused by, this misinformation and manipulation. Populists have very little patience for complex facts, preferring simple and emotionally charged messages. So, by fighting disinformation the DSA can contribute considerably to countering the populist attack on democracy.

The DSA is also opposing the micro-targeting of voters through NSM; the practice in which psychological profiles are used to bombard individual voters with political propaganda specifically geared towards their personality profile. By privately targeting individual voters, the propaganda cannot be scrutinised in the public sphere, a process which constitutes the core of any functioning democracy. (17) Democracies can be very tolerant towards the truth and untruth of political propaganda, as long as they are confident that these falsities and untruths will be brought to light in the democratic discussion itself. This form of voter-manipulation was made infamous by Cambridge Analytica, a company that boasted the election of Trump and the Brexit as their successes.

Populism cannot thrive where claims are subjected to public discussion. It is therefore cause for optimism that the DSA makes political micro-targeting more difficult by introducing restrictions on what personal data can be used and for what purposes and by forcing platforms to be transparent about how advertisers (political and others) approach their targets. More recently, in March 2023, the Council of the EU has published the Transparency and Targeting of Political Advertising Regulation (18).

What the DSA cannot do

Discouraging misinformation and preventing the abuse of personal data to politically manipulate citizens are both important and within the reach of the DSA. But the links between the new digital platforms and populism go deeper. There is more to the connection between digital media and populism than the fact that these media are venues for spreading misinformation and for manipulating citizens. The DSA constitutes an important starting point, but more is needed to help weaken the forces behind the rise of populism.

I want to mention three features of digital social media that constitute a threat to flourishing democratic practices and institutions by opening the door to populism, but that do not qualify as misinformation or manipulation. These digital risks to democracy deserve policy answers. The problem is, unfortunately, that the same features that undermine democracy can also enhance it.

The first democratic risk is the acceleration of news. Compared to newspapers, radio and television, digital media deliver news much faster. Developments in the world are sometimes even covered in real-time. The downside is that that there is often no time for careful fact-checking. False, or more likely, strategically incomplete information is allowed to spread with unprecedented speed. Corrections and contextualisation lag severely behind, thus usually receiving only a fraction of the attention that the original message received because everyone is already caught-up on today's unchecked news. From the perspective of an informed public - a requirement for democratic debate - this is highly problematic as established mechanisms for weeding out weak or false information are not able to function at this dizzying speed. Populism typically relies on quick and simple messages. Digital acceleration is therefore inherently biased in favour of populism, putting parties that rely on more informed and complex perspectives at a disadvantage. On the other hand, real-time coverage can also be an asset for democratic opinion formation. It often escapes government censuring and enables quick public responses, e.g., to disasters or acute forms of injustice.

Secondly, digital media play into our psychological bias favouring emotionally charged news over more bland and complex information. There is ample evidence that suggests that lies spread more easily and more quickly in digital space than complex and nuanced truths (19). They are usually easier and quicker to digest; they are tastier and spread by enthusiasts and critics alike. Populist leaders know that it is less important to be truthful, than to raise a ruckus. Furthermore, this psychological bias is exacerbated by algorithms that feed us what we like – in this case: more and more extreme material. On the other hand, appealing to emotions is not necessarily a bad thing. After all, democracy is not an academic discussion group but can and should be a passionate affair. If emotionally charged forms of communication can draw new citizen groups towards the democratic debate and engage them in civic activities, that is a plus for democracy.

The third feature is that social media eliminate gatekeepers and editorial filters. Citizens, at least those able to post on the Internet, have now immediate access to their fellow citizens. That seems nice: direct, unmediated communication! But in fact, gatekeepers and filters are not only necessary to ban misinformation, but equally important: they also embody and enforce democratic norms of civility, mutual respect and perspective-taking. These norms permit a real dialogue where parties aim to learn from each other. Ideally, in a dialogue – as opposed to a discussion – participants are willing to revise their opinions in the light of the reasons offered by their fellow citizens.

Some decades ago, in the early days of the Internet, enthusiasts indeed expected digital space to be a democratic utopia (20, 221). By now we have learned that real dialogues are hard to organise on the Internet. There are successful examples, e.g., patient or victim groups where people can help each other and learn from one another. But in politics such open and constructive interactions are exceedingly rare. A situation in which one is unable to observe another's face or body language and where participants usually remain anonymous, invites behaviour that people would never consider in their offline lives. Emotions and conflicts easily explode, resulting in hate speech.

Populism, thriving on antagonism, seems to be tailor-made for this type of polarising communication. On the other hand, social media can also enhance democracy by providing previously marginalised voices with a direct access to the public agora. Good manners have in the past all too often been a ploy to exclude fellow citizens from the democratic debate who lack training in the art of 'civil conversation'.

This analysis is far from exhaustive. These three mechanisms serve only as examples to demonstrate a more general point: it is not only the content spread by digital social media that can undermine democracy and pave the way for populism that refuses to peacefully organise pluralism and heterogeneity within the citizenry. There is also something in the digital technology itself that is biased against democratic deliberation and that favours forms of political communication typical of populism. These features are its speed, its being biased towards emotionally charged messaging, and its unmediated accessibility for everyone. Simply focussing on misinformation and manipulation will always lag behind the spread of populist messaging. This is all the more true because in the majority of cases our democratic laws don't prohibit lying and propagandising because they rightly entrust it to public debate to weed out falsehoods. The problem is, as argued above, that it is questionable whether an increasingly digitalised public sphere is still up to that traditional democratic task.

Policy recommendations

As we have seen populism has many causes, of which the digitalisation of the public sphere is only one. Looking for ways to ensure that digitalisation enhances democracy, rather than populism, is then at best a partial solution. However, that does not lessen its importance.

When we think about the links between digital platforms, digital media, democracy and populism; policy is entering largely unchartered waters. It will have to find a balance between installing safeguards for the quality of democratic opinion formation and remaining flexible and open enough to welcome in citizens who, thanks to the low thresholds of digital social media, are finding their way into the public sphere for the first time. Policy measures can therefore only be tentative and should be subject to democratic debate. The safeguarding of democracy and updating it for the 21st century is a task for democracy itself. But any democratic deliberation starts by someone making some suggestions. Here are some (not all of them original).

Possible measures can be social, legal and technical:

- We should organise a debate on the benefits but also on the risks of the speed and acceleration of news. A campaign for slow news may be an option.
- Norms for how to engage in a democratic discussion with one's fellow citizens should be explicated and translated into an etiquette for digital exchanges. People should be instructed and trained in them, both in offline and in online environments. They should also be made aware of how self-evident norms for offline behaviour are easily forgotten in online environments.
- There should be legal regulations for the design of algorithms governing the social media, so as to counter the current tendency to cater to our biases and emotional urges only. Interfaces and algorithms should be designed to enhance diversity of networks and opinions, rather than their homogeneity (22, 23). In this way they can help citizens to recognise each other as fellow members of the civic community.
- Interfaces and algorithms for social media should be so designed that they incorporate civic virtues. For instance, they could help users to determine to what extent opinions are evidence-based; they could ensure diversity and enhance inclusion; they could enhance empathy by stimulating taking the perspective of one's fellow citizens; and they could incentivise participants to not only send but also to listen. For all this it is necessary to expel the currently dominant market-logic: the more clicks the better, the more attention and time spent on our site the better.
- Governments should take care to protect believable, independent and prestigious channels for news and discussion, thus providing a meeting ground for citizens which can lure them away from their bubbles.
- Social media are about more than discussion. They also allow citizens to peacefully resist what they perceive as injustice and to organise into political communities. These digital spaces should be protected even if they do not conform to the ideals of a shared, collective, public sphere where all citizens meet.

Author information:
Tsjalling Swierstra is a Professor of Philosophy at Maastricht University's Faculty of Arts and Social Sciences. His work explores the ethical and political implications of emerging sciences and technologies.

References

1. Zuiderveen Borgesius FJ, Möller J, Kruikemeier S, Fathaigh RÓ, Irion K, Dobber T, Bodo B, De Vreese C. *Online political microtargeting: Promises and threats for democracy*. Utrecht Law Review. 2018 Feb 9;14(1):82-96.
2. Mudde C. *The populist zeitgeist. Government and opposition*. 2004 Jan;39(4):541-563.
3. Held D. *Models of democracy*. Redwood City: Stanford University Press; 2006.
4. Warren ME. *A problem-based approach to democratic theory*. American Political Science Review. 2017 Feb;111(1):39-53.
5. Rogowski JC, Sutherland JL. *How ideology fuels affective polarization*. Political behavior. 2016 Jun;38:485-508.
6. Kingzette J, Druckman JN, Klar S, Krupnikov Y, Levendusky M, Ryan JB. *How affective polarization undermines support for democratic norms*. Public Opinion Quarterly. 2021 Oct 1;85(2):663-77.
7. Müller JW. *What Is Populism?* Philadelphia: University of Pennsylvania Press; 2016
8. Mudde C. *The Far Right Today*. Cambridge: Polity; 2019.
9. Halla M, Wagner AF, Zweimüller J. *Immigration and voting for the far right*. Journal of the European economic association. 2017 Dec 1;15(6):1341-85.
10. Wetenschappelijk Raad voor het Regeringsbeleid. WRR-rapport nr. 108_Grip. *Het maatschappelijk belang van persoonlijke controle*. 2023 [cited 2024 June 22]. Available from: www.wrr.nl/publicaties/rapporten/2023/11/30/grip
11. Springveld N. *Neoliberalism, precarity, and precariousness*. Framework. 2017 Dec;30(2):25-39.
12. Young M. *The rise of the meritocracy.* New York: Routledge; 2017. (first published in 1958)
13. Bovens M, Wille A. Diploma democracy: *The rise of political meritocracy*. Oxford: Oxford University Press; 2017.
14. Kampen T, Elshout J, Tonkens E. *The fragility of self-respect: Emotional labour of workfare volunteering*. Social Policy and Society. 2013 Jul;12(3):427-38.
15. Jäger A. *Hyperpolitik: Extreme Politisierung ohne politische Folgen*. Frankfurt am Main: Suhrkampl 2023
16. Lorenz-Spreen P, Oswald L, Lewandowsky S, Hertwig RA. *systematic review of worldwide causal and correlational evidence on digital media and democracy*. Nature human behaviour. 2023 7(1):74-101.
17. Brkan M. *EU fundamental rights and democracy implications of data-driven political campaigns*. Maastricht Journal of European and Comparative Law. 2020 Dec;27(6):774-90.
18. Council of the EU. EU introduces new rules on transparency and targeting of political advertising. 2024 [cited 2024 June 22]. Available from: Https://www.consilium.europa.eu/en/press/press-releases/2024/03/11/eu-introduces-new-rules-on-transparency-and-targeting-of-political-advertising/
19. Vosoughi S, Roy D, Aral S. *The spread of true and false news online*. science. 2018 Mar 9;359(6380):1146-51.
20. Barlow JP. A *Declaration of the Independence of Cyberspace*. Davos: Electronic Frontier Foundation; 1996.
21. Negroponte N. *Being Digital*. New York: Alfred A. Knopf; 1995.
22. Bozdag E, Van den Hoven J. *Breaking the filter bubble: democracy and design*. Ethics and Information Technology. 2015 ;17(10.1007).
23. Helbing D, Mahajan S, Fricker RH, Musso A, Hausladen CI, Carissimo C, Carpentras D, Stockinger E, Sanchez-Vaquerizo JA, Yang JC, Ballandies MC. *Democracy by design: Perspectives for digitally assisted, participatory upgrades of society*. Journal of Computational Science. 2023 Jul 1;71:102061.

Maastricht University
STUDIO EUROPA MAASTRICHT

Policy Brief

How teenagers’ lifeworlds are shaped with snaps, streaks and social surveillance

Dr Katleen Gabriels, Emma Prebreza

Summary

With five million active users in the Netherlands, Snapchat is integral to the digital lifeworld of many people, including minors. This paper critically examines Snapchat’s design choices and their implications for underage users. Snap Inc. employs strategies such as algorithm-driven content curation, live location tracking and gamified interactions to maximise user engagement. While effective at increasing activity, these strategies often prioritise engagement metrics over the well-being of young users who are particularly sensitive to social feedback, validation and rejection.

This paper’s focus on a platform heavily used by minors aligns with the objectives of the Digital Services Act (DSA) which aims to better protect minors in the EU. In addition, Snapchat falls under the category of Very Large Online Platforms (VLOPs) of the DSA. Our paper argues that the responsibility of social media companies should extend beyond profit maximisation to include the well-being of their users, especially minors. In 2024, the European Commission acted against ByteDance (TikTok), Meta (Instagram and Facebook) and X Corp. (X, formerly Twitter) under the DSA for using so-called dark patterns: *design techniques that mislead users and prompt certain behaviours. Our analysis indicates that the DSA may not offer sufficient protection at this moment, also due to Snapchat's lack of age verification measures.*

While not all young users are equally affected, raising awareness about these practices is crucial. This paper advocates for comprehensive education programmes to help young people navigate social media responsibly and critically. It offers recommendations for policy-makers to enhance protections and urges parents and schools to play a proactive role in guiding young users. Creating environments where young people can openly discuss their online experiences and learn to manage them effectively is essential for their benefit and well-being.

Introduction

‘There was a sleepover that a classmate was not invited to. She saw it on Snap Map.’ ‘A friend was angry because I lost a Streak.’ These anecdotes from teenagers illustrate how design choices, in this case of Snapchat, shape their lifeworlds. The philosophical concept lifeworld refers to the everyday world of lived experience, from a subjective first-person perspective (1). Whereas the Snap Map is an example of how we have come to regard real-time location tracking as normal behaviour, the Streaks, which encourage users to interact daily with one another, are part of gamification strategies.

Gamification is the use of elements of game design in non-game settings, for instance, to motivate, engage and stimulate people to change their perception, attitude or behaviour (2). It is based on a reward system with levels, points, leader boards, badges and so on. An extrinsic reward system can lead to superficial engagement where users participate in activities solely to earn points and badges rather than engage in meaningful interactions. Companies implement gamified systems to encourage users to spend more time on their platforms. Gamification often requires the collection of user data to tailor experiences and track progress. On Snapchat, the Streaks, Snapscore and the earned badges (see Part 2) are examples of gamification.

Today, social surveillance (the systematic monitoring and analysis of someone's actions and behaviours within digital spaces) is increasingly regarded as an acceptable norm (3). Whereas live location tracking was still in its infancy less than a decade ago, it is now provided free and by default on widely-used devices. Apple offers the Find My app as a standard service to allow family members and friends to share their real-time locations. WhatsApp and Facebook Messenger (both from Meta) offer live location sharing. The Snap Map is a popular tool of Snapchat. As we will show, young people can feel pressure to share their location on Snap Map. Not participating by having the feature turned off can be perceived as a rule violation and lead to criticisms and suspicions which they would rather avoid. Of course, it can also be nice to see what your friends are doing and sharing your location can contribute to a feeling of safety. Yet, the Snap Map raises compelling questions concerning privacy, as it may reveal sensitive information, including personal details and habits of minors.

In this policy paper, we critically examine how these design choices shape behaviour, relations, communication and norms. The focal point of our analysis lies on Snapchat. As Snapchat has more than 45 million active users per month, it falls under the Very Large Online Platforms (VLOPs) category of the Digital Services Act (DSA). We have deliberately chosen a platform that has many underage users because the DSA aims to better protect minors in the EU. In May 2024, there was a scandal in a primary school in Kerkrade (Limburg, the Netherlands) when in several Snapchat groups, horrific pictures of terrorism and even sexual abuse of children were being shared (4). According to their Terms of Use, Snapchat allows anyone 13 and older to sign up (5). However, the platform has no means of age verification in place. Teenagers typically start using Snapchat for the first time when they are around 11-13 years old (6). They often report peer influence as a key motivator: their friends are using it and they feel pressured to follow the trend (7).

Our paper unfolds into four parts. First, we explain Snapchat and its most important features. This includes an academic literature study of how Snapchat's design shapes social dynamics of teenagers, and how it can impact family relations. Second, we describe Snapchat's privacy and advertisements policy in relation to the DSA. Third, based on our analysis, we formulate conclusions that we subsequently integrate in the fourth and final part: recommendations for policy-makers, parents and schools. The overall objective is to raise awareness about a platform used by most Dutch teenagers. Since Snapchat is integral to their daily lives, we believe it is crucial to gain a deeper understanding of its impact.

Snapchat

Launched in 2011 by three students at Stanford University, Snapchat is a popular social networking site (SNS) that draws in over 400 million daily users. Of these, 96 million come from Europe (8), including five million active users in the Netherlands (9). Snapchat's parent company is Snap Inc.; which claims to reach 90% of 13 to 24-year-olds in the Netherlands (10). On average, Dutch youth spend six hours a day on their mobile phone, including two and a half hours on social media (10, p. 3). Snapchat is available in a mobile version (app) and a web version. While young adults (18-24) comprise Snapchat's core demographic (38.1%), 18.6% of users fall within the 13-17 age group, demonstrating the platform's reach among adolescent users. In 2023, Snapchat's popularity among teen users surpassed that of Instagram (Meta) (11).

The basic and most popular version of Snapchat is free, with the company primarily generating revenue from targeted advertising. Users can subscribe to the platform's premium service called Snapchat+. For a monthly fee, they get access to exclusive features, including seeing who rewatched their stories, earning special badges that exemplify their premium status or restoring a lost Streak. The accounts of users under 18 may be linked to their parents' bank account. Parents might not pay attention to these smaller transactions or have been persuaded by their child to pay these costs.

How does Snapchat work?

Snapchat stands out from other popular SNS by prioritising moments of an ephemeral nature which are 'designed to erase communication artefacts after a short period of time' (12, p. 957). Snapchat's key communication artefact is the *Snap*: a photo or a video that a user can send to a chosen recipient who can only view its contents for up to ten seconds and replay it once. Each Snap can be accompanied by a short text description or edited using various filters powered by generative AI, including Snapchat Lenses (13). Communication with other users is limited to an individual's friend list. Friends can add each other by searching for usernames, importing phone contacts or scanning their unique profile QR codes (the Snapcode). With the algorithm driven Quick Add feature, Snapchat offers friend recommendations based on factors like mutual friends or the individual's enabled location setting (14).

If two users exchange Snaps (not chats) for three consecutive days, they initiate a *Snapstreak*. The score on this Streak increases if they continue to send at least one Snap to each other within a 24-hour window. The number of days the Streak has been active is shown, which can be up to hundreds and even thousands of days. The strength of a relationship is then marked by Friend Emojis, ranging from BFs to Super BFFs. The hourglass emoji is a notification feature that appears when one's Snapstreak with a friend is about to expire. The emoji serves as a reminder to send a Snap soon, otherwise the Streak will be broken. Users can also pay to restore a Streak; prices vary per country and users receive one free Snapstreak Restore. The Snapscore is a numerical score that represents the total number of Snaps sent and received, stories posted and other interactions on Snapchat. Each user's Snapscore is displayed on their profile and can be seen by their friends, serving to track engagement on the platform.

Another feature is the *Snapchat Story*: a collection of Snaps displayed on a user's profile for 24 hours. Unlike regular Snaps, these stories can be viewed without restriction. Users can also post on their private Snapchat Story, visible only to a selected group of contacts. The platform also serves as a messaging tool, where users can chat using text, stickers or initiate audio and video calls. Snapchat also allows users to create group profiles and group chat groups of up to 200 users.

Spotlight is a creator hub, where users can share their self-created videos (Snaps) with users worldwide. A Spotlight can be uploaded in the same way as a Snap, but it is always public. On the Spotlight page, videos from the entire Snapchat community can be watched. Spotlight is regarded as competition for TikTok and Instagram Reels.

The *Snap Map* is a live location sharing feature that allows users to keep track of the location of their friends. It displays the user's *Bitmoji*, a personalised cartoon avatar pinpointed on a world map. Other users can zoom in on this avatar to reveal the exact geolocation, including street names and movements, of the user. The feature works by utilising data from the GPS chip installed in the user's device (15). In addition to seeing their friends' location, the Snap Map also enables users to explore events in their local area and beyond by viewing *Live Stories*: collaborative Snapchat stories which feature videos from users who decide to upload their Snaps to the map. Moreover, *Heat Maps* can display popular events with a high rate of *Snapping* (Snapchat activity) at the location. Users can control their privacy settings (see Part 3), choosing to share their location with all friends, selected friends or use *Ghost Mode*. Ghost Mode means hiding your location on the map while still being allowed to see the locations of friends with the feature enabled (16).

Snapchat's latest feature addition (in 2023) is My AI, an in-app generative chatbot that mimics a virtual friend who can answer questions, provide recommendations or facilitate casual conversation.

Social dynamics of Snapchat use among teenagers

Unlike text messages, Snaps are often seen as a more personal form of communication, allowing users to convey emotions through a combination of text and image (17). Snapchat's ephemeral (i.e., disappearing) nature encourages users to exchange spontaneous, unfiltered moments from their daily lives, including candid images they might not post elsewhere (12). Déage (18) warns that the ephemeral nature of Snaps coupled with the perceived sense of safety makes Snapchat the perfect facilitator for sexual content such as nude imagery. This practice is concernedly common among teenagers, and the contents often circulate among peers or are exploited for purposes such as so-called revenge porn, blackmailing or bullying (7). Similarly, bullying can also be facilitated through private stories, where teenagers use this 'exclusive' feature to complain or gossip about others.

Although Snapchat notifies a sender when their message was read or screenshotted, acts like screen-recording or using another phone to take a picture of the Snap go unnoticed. Studies among teen users found that screenshotting private or intimate content can lead to feelings of discomfort and anxiety (7) and is often seen as a breach of peer trust (17).

Communication on Snapchat also has its unwritten rules. One study found that teen users often expressed frustration when these rules were not followed, including Snapping excessively or breaching implicit relationship boundaries (17). In addition, studies reveal a concerning trend regarding communication (see, e.g., 18); underage users did not consider most people in their friend list as someone they could trust. Similarly, teens often engage in Snapping with strangers or friends-of-friends they met through Snapchat's recommendations (7). Some feel pressured to add a 'friend' based on their number of mutual friends.

Furthermore, the practice of maintaining Streaks and earning special rewards represents gamification challenges (19, 20), which encourage users to spend more time with the app. Some teenagers, worried about losing their Streaks, might share their passwords with friends when they know they will not be able to use Snapchat for a while; for instance, when they go to a youth camp or school trip where smartphones are not allowed (18). Refusing to comply might raise suspicions that a friend is concealing something. Similarly, a lost Streak requires an explanation and restarting this reciprocal communication is an expected behaviour (20). Thus, with the sole purpose of maintaining their Streaks, teens commonly send each other content-poor images such as a black screen or so-called Mass Snaps addressed to several users at once (19). A study among 2483 Belgian early adolescents found that those who maintain Streaks show significantly higher problematic smartphone use (21).

While these design choices can be engaging, they can also contribute to *digital stress* (22); and the pressure to be readily available or maintain a certain status can lead to conflict or jealousy (21). An example is Snapchat's Mutual Besties badge, which reveals whether your best friend on Snapchat is also the best friend of someone else, potentially sparking insecurity. Most of Snapchat's users tend to prioritise social connectedness (23), an aspect that is particularly crucial during adolescence and is characterised by increased vulnerability and a desire for acceptance (6). Young people are particularly sensitive to both positive social feedback and rejection; leveraging these metrics to sustain platform engagement exploits their vulnerabilities and can likely result in problematic use (24, p. 2).

A qualitative study with 51 Belgian secondary school students showed that they experienced a connection overload from perceived pressure to be always available to their friends' notifications (25). For example, the Read function increased the pressure to respond quickly for both the sender and receiver. Finally, My AI has sparked significant controversy as it was found to provide sex advice for underage users (26) and appears to have access to location data without users' consent (27).

Effects of the snap map on teenagers

Location tracking often occurs without users' explicit knowledge or perception of its associated risks (28). The Snap Map can reveal significant details about a person's routines and habits, including potentially sensitive information related to health, ethnic affinity, socioeconomic status or religious beliefs (15). For example, someone's live location might disclose a visit to a sexual health clinic or attendance at a religious service. Such data can be leveraged by big tech and third parties for marketing and other revenue-generating purposes (28, 29). Beyond commercial exploitation, location data can be misused by criminals or other users for blackmail, harassment or stalking (29).

Monitoring the activity and whereabouts of others as part of social surveillance practices is a major motivator for adolescents' use of SNS, likely influenced by their identity development and desire to fit in with peer groups (3). Similarly, Sachs (6) argues that amid this 'impressionable stage of development', this age group is particularly vulnerable to its negative impacts (p. 64). While the ability to control who sees their location, stay connected to social events or pass time are seen as benefits of Snap Map, teenagers also tend to express concerns about their perceived lack of privacy (7). Specifically, they report feeling uneasy about others being able to track them constantly and feel pressured to be always accountable for their whereabouts (6). Despite these concerns, disabling the feature or turning on Ghost Mode may be seen as violating an established norm, as well as a sign of suspicious behaviour among peers (7).

Dunn and Langlais (30) discovered a positive correlation between higher use of Snapchat and a decrease in overall mental health regardless of age, ethnicity or gender. The surveillance of peers via Snap Map was associated with stress (e.g., monitoring ex-partners' whereabouts) as well as a fear of missing out (FOMO). Similarly, being excluded from social gatherings can cause feelings of exclusion, sadness and low esteem (6). The Snap Map and real-time location tracking can create a harmful loop: users increase their social media use to escape FOMO but intensify it due to exposure to endless social opportunities (31). Although monitoring others' seemingly 'fruitful social lives' can contribute to loneliness and depression, Vanherle et al. (3, web) suggest that young users might still feel compelled to engage in social surveillance to maintain their social image.

Snapchat and parents

Here, we include a literature study on how teenagers and parents regard Snapchat. We found a scarcity of existing studies on this topic, highlighting a gap in the literature. Addressing this gap in future research is important to better understand the impact of Snapchat on family dynamics.

Vaterlaus et al. (17) demonstrated generational differences in Snapchat use between teenagers and their parents. Teenagers frequently expressed frustration that their parents did not understand Snapchat's functionalities or benefits, and they found this gap a missed opportunity to foster a closer relationship with their parents. Snapchat's disappearing messages might be particularly appealing to teen users as it allows them more control over what their parents see (i.e., to escape parental control) (21). However, Snapchat can also be a tool for family connection. A study with American college students found that they frequently use Snapchat to communicate with close and distant family members (32). Moreover, its design features, such as the ability to reply to Snapchat Stories, made these interactions more engaging.

Parents can also try to control their teens' technology use. However, Yardi and Bruckman (33) found that as children grow older and enter their teenage years, enforcing limits on technology use and keeping up with their online activities becomes increasingly difficult. Likewise, parents can struggle with their own digital skills or lack of time.

Snapchat provides several features for parents to safeguard their children from potential safety threats. For example, they can control whether their teens communicate with My AI. If children try to use it, Snapchat will let them know it was disabled. Moreover, with the Family Center feature, parents can see who their children are snapping with or have in their friend list without seeing the content of this communication. Starting in 2024, the tools of the feature expanded, so parents can now see who their children share their location with on Snap Map, as well as the child's location settings (34).

Privacy and advertisement policy

The policy pages on Snapchat's Dutch website implement the legislative frameworks of the DSA and General Data Protection Regulation (GDPR); all information in this section is based on these webpages (35, 36).

Snap Inc. offers personalised advertisements to maintain a free service. The advertisements are based on information that users provide themselves (e.g., birth date, gender), what the company thinks a user may find interesting, users' activities on the platform and information that Snap Inc.'s partners and advertisers provide about users. For instance, if a user searches for a film on a partner website, an ad for that film will be shown on Snapchat. Advertisers can then measure to what extent their ads have been successful.

Snap Inc. tracks users' activity on Stories, Spotlight (e.g., the content of Snaps they post or view on Spotlight), Snap Map, My AI, and so on. On My AI, the ads are based on the context (content) of the conversation. The company also tracks which content users look at and which functions they use on Snapchat. Based on users' behaviour, they make several assumptions (i.e., profiling), which they call Lifestyle Categories. The company subsequently makes inferences about a user's interests and has Snapchat Content Categories to categorise the content that a user interacts with.

Furthermore, Snap Inc. also collects information about users' context, location and device including insight into its operating system, screen size, language selection, installed apps and other features. The ads also take location into consideration. For instance, if users are near a coffee shop, an advertiser might show them ads for coffee.

The company does not share information with advertisers that directly identifies users, such as a name, phone number or email address. Data are analysed for personalised experiences, including personalised ads and friend suggestions. Snap Inc. makes use of machine learning to optimise their services, including personalisation and advertisements. In line with the DSA and GDPR, they state they do not profile EU users under the age of 18 in order to personalise ads (37). Yet, as mentioned before, the company has no serious age verification in place.

EU users can opt out of activity-based ads, ads based on target groups and ads from third-party networks. They can also change the ad topics and set ads preferences, and they can report misleading ads (e.g., clashing with community guidelines). Users can adapt and delete information, such as Lifestyle Categories and Snapchat Content Categories, and can manage their basic account. Furthermore, they can control with whom they share information. They can download a copy of the data that Snap Inc. keeps of them and, if they want to remove their account, the procedure on how to do this is available. In line with the DSA and the rules for VLOPs, Snapchat users in the EU can opt out of personalised content altogether.

Concluding thoughts

Recently, there have been alarming publications about the influence of the smartphone on the mental well-being of young people (see, e.g., 38). However, so far, there is no convincing evidence that social media have long-lasting negative effects on their mental well-being. Findings of Project AWeSome, a Dutch research project that focuses on adolescents, well-being, and social media, show that young people who are vulnerable offline are also at greater risk online (10, p. 8). For instance, young people who are anxious or stressed and who tend to compare themselves socially are more vulnerable online than other young people. Conversely, youth who are resilient offline also seem to be at lower risk online.

Young people are a heterogeneous group, so we must be careful with one-sided conclusions. The insights based on our literature study confirm that Snapchat design can provoke certain behaviour, especially at an age when teenagers want to be affirmed by their peers. Although the effects will not be negative for everyone, there is a vulnerable group that must be protected better. For instance, as we have shown, Snapchat shapes interpersonal dynamics and emotional well-being, as minor disruptions in online engagements can escalate into conflicts, such as having an argument over losing a Streak, which can be considered a breach of commitment. Other design choices, such as the Mutual Besties Emoji, can contribute to insecurity, competition and jealousy. Also, teenagers have been naive about the promise of supposedly disappearing Snaps. The desire to belong to their peer group often supersedes the careful consideration of terms and conditions, leaving them susceptible to the design choices.

Since 2018, the GDPR has prohibited the profiling of minors on platforms; however, there is sufficient evidence that platforms do not abide by this rule (10, p. 12). Because Snapchat is a VLOP, specific rules and obligations, including risk-assessment, are in place. VLOPs also undergo independent audits. Compared to the GDPR, the DSA includes a stronger enforcement procedure to hold platforms more accountable and to be more transparent about their terms and conditions. Based on their Dutch website (see previous section), Snap Inc. complies with the DSA, at least in terms of the information they provide. Yet, it is impossible for us to check to what extent minors are profiled and targeted, and to what extent their data are collected, especially without effective age verification in place.

The DSA explicitly prohibits so-called dark patterns, such as manipulative design practices and addictive patterns of use. Regarding dark patterns, the European Commission (EC) has started formal proceedings against ByteDance (TikTok) (39), Meta (Facebook and Instagram) (40) and X (X Corp.) (41). According to the EC, ByteDance and Meta do not do enough to protect minors online and to avoid gamification strategies (the formal proceedings against X do not focus on minors). The EC is concerned that TikTok, Facebook and Instagram, including their algorithms, may stimulate behavioural addictions in children. Additionally, the EC has reservations about the age-assurance and verification methods implemented by Meta and ByteDance. Based on our analysis, we can raise similar concerns about Snap Inc. In our understanding, the risks regarding minors are not sufficiently mitigated and the company cannot simply place the responsibility on the user (and the user's parent). Parents often face significant challenges in this regard. They may lack awareness of how features on Snapchat operate, struggle to monitor their children's interactions effectively or be constrained by time issues that prevent thorough supervision. These gaps in parental oversight can leave minors vulnerable to various risks, including peer pressure.

By joining forces, policy-makers can implement robust regulations that prioritise the protection of minors online, while schools can integrate digital literacy and online safety into their curricula. Parents can be supported through workshops and resources that enhance their understanding of digital platforms and equip them with strategies to guide their children's online activities.

Policy recommendations

Policy-makers:

1. Invest in comprehensive education programmes, such as digital literacy initiatives, to ensure young users understand how design choices influence their behaviour and how they can adjust their privacy settings.
2. Mandate Snap Inc. to implement and control mandatory age verification.

Parents:

1. Familiarise yourself with the online platforms important to your child. Be aware that it is in fact against the rules for children under 13 to have a Snapchat account.

2. Educate yourself and your child about the importance of privacy settings and the potential consequences of sharing personal information, such as location data.

3. Review Snapchat's privacy and advertising policies together with your child, and discuss their implications and possible consequences. Collaboratively adjust the settings (users in the EU can opt out of personalised content altogether) to encourage critical reflection and awareness, including about peer pressure.

4. Establish clear guidelines for screen time, social media use and maintaining healthy offline relationships, ideally before your child receives their first smartphone.

5. Engage in open and ongoing conversations with your child about digital experiences, including live location sharing (social surveillance): what are its advantages and disadvantages?

6. Educate yourself about problems such as cyberbullying and sexting so that you can provide informed advice or know where to seek help if your child encounters problems.

7. Learn about family tools and resources that can help manage your child's online activities (e.g., Snapchat's Family Center, Snapchat's guide for parents, knowledge centres for media literacy).

Schools and teachers:

1. Integrate discussions on digital literacy and online safety into the curriculum, dedicating specific lessons to platforms like Snapchat and their features.

2. Invest in creative methods to encourage critical thinking and responsible digital behaviour.

3. Foster open communication channels where students feel comfortable discussing their online experiences, including peer pressure related to Streaks or privacy concerns with the Snap Map. Offer guidance on establishing healthy boundaries in digital interactions.

4. Support and inform parents by organising workshops and seminars.

5. Share online resources with students to enhance their understanding and awareness.

Author information:

Katleen Gabriels is an Associate Professor at the Faculty of Arts and Social Sciences at Maastricht University. She is specialised in moral philosophy and the philosophy of technology, with her research focusing on the interplay between morality and computer technologies.

Emma Prebreza is a graduate of the Maastricht University BA Digital Society programme and is currently an MSc student in International Public Management and Policy at Erasmus University Rotterdam.

References

1. Eberle TS. *Exploring Another's Subjective Life-World: A Phenomenological Approach*. Journal of Contemporary Ethnography. 2015;44(5):563-579.
2. Shahri A, Hosseini M, Phalp K, Taylor J, Ali R. *Towards a Code of Ethics for Gamification at Enterprise*. In: Frank U, Pastor O, Loucopoulos P, Petrounias I, editors. *Proceedings of The Practice of Enterprise Modeling*. 7th IFIP WG 8.1 Working Conference, PoEM 2014, Manchester, UK, November 12-13, 2014. Berlin – Heidelberg: Springer-Verla; 2014. p. 235-245.
3. Vanherle R, Trekels J, Hermans S, Vranken P, Beullens K. *How It Feels to Be "Left on Read": Social Surveillance on Snapchat and Young Individuals' Mental Health*. Cyberpsychology: Journal of Psychosocial Research on Cyberspace. 2023;17(5):Article 3.
4. Koeyvoets C. *Lijken en kinderporno; leerlingen delen gruwelijke beelden*. L1 [Internet]. 2024 May 7 [cited 2024 Sept 12]. Available from: https://www.l1nieuws.nl/nieuws/2651225/lijken-en-kinderporno-leerlingen-delen-gruwelijke-beelden
5. Snap Inc. *Snap Inc. Terms of Service 2024* [Available at: https://snap.com/en-US/terms]
6. Sachs J, editor *Psychological Repercussions of Location-Based Social Networks in Today's Youth 2018*.
7. Green B. *Streaks, Stories, and Social Capital: A Bourdieusian Approach to Teenagers' Use of Snapchat*: University of Guelph; 2020.
8. Dean B. *Snapchat Demographic Stats: How Many People Use Snapchat?* Backlinko [Internet]. 2024 May 18 [cited 2024 Sept 12]. Available from: https://backlinko.com/snapchat-users
9. Author unknown. *5 miljoen Nederlandse Snapchatters en de teller loopt door*! Newsroom Snap [Internet] 2023 September 13 [cited 2024 Sept 12]. Available from: https://newsroom.snap.com/nl-NL/5-million-dutch-snapchatters-and-counting
10. Valkenburg PM, van der Wal A, Beyens I. *Schermgeluk en schermverdriet: De invloed van social media op de mentale gezondheid van jongeren* (Essay 4). Unicef-essayreeks Kinderrechten in de digitale wereld. 2023. [Available at: https://www.project-awesome.nl/]
11. Pew Research Center. *Teens, Social Media and Technology 2023*. 2023.
12. Bayer JB, Ellison NB, Schoenebeck SY, Falk EB. *Sharing the Small Moments: Ephemeral Social Interaction on Snapchat*. Information, Communication & Society. 2016;19(7):956-77.
13. Bhatti D. *5 Ways Snapchat Uses Artificial Intelligence and Machine Learning*. Medium [Internet]. 2023 May 23 [cited 2024 Sept 12]. Available from: https://daanishbhatti.medium.com/5-ways-snapchat-uses-artificial-intelligence-and-machine-learning-a885d29eed66
14. Firoozabadi Dehghani I. *What Is Quick Add on Snapchat: A Comprehensive Guide in 2023*. Medium [Internet]. 2023 October 14 [cited 2024 Sept 12]. Available from: https://idehghani28.medium.com/what-is-quick-add-on-snapchat-a-comprehensive-guide-in-2023-ed455253b96c
15. Zreik J-P. *Geo-location, Location, Location*. Rutgers Computer & Tech LJ. 2019;45:135.
16. Snap Inc. *How Do I Share My Location with All of My Friends on Snapchat?* [Available at: https://help.snapchat.com/hc/en-gb/articles/7012270909972-How-do-I-share-my-location-with-all-of-my-friends-on-Snapchat]
17. Vaterlaus JM, Barnett K, Roche C, Young JA. *"Snapchat Is More Personal": An Exploratory Study on Snapchat Behaviors and Young Adult Interpersonal Relationships*. Computers in Human Behavior. 2016;62:594-601.
18. Déage M, editor *'Don't You Trust Me? 'Teenagers Challenging Friendship on Snapchat*. ECSM 2019 6th European Conference on Social Media; 2019.
19. Hristova D, Dumit J, Lieberoth A, Slunecko T, editors. *Snapchat Streaks: How Adolescents Metagame Gamification in Social Media*. GamiFIN; 2020.
20. Hristova D, Jovicic S, Göbl B, de Freitas S, Slunecko T. *"Why Did We Lose Our Snapchat Streak?". Social Media Gamification and Metacommunication*. Computers in Human Behavior Reports. 2022;5:100172.
21. van Essen CM, Van Ouytsel J. *Snapchat Streaks—How Are These Forms of Gamified Interactions Associated with Problematic Smartphone Use and Fear of Missing out among Early Adolescents?* Telematics and Informatics Reports. 2023;11:100087.
22. Steele RG, Hall JA, Christofferson JL. *Conceptualizing Digital Stress in Adolescents and Young Adults: Toward the Development of an Empirically Based Model*. Clinical Child and Family Psychology Review. 2020;23(1):15-26.

23. Grieve R. *Unpacking the Characteristics of Snapchat Users: A Preliminary Investigation and an Agenda for Future Research*. Computers in Human Behavior. 2017;74:130-8.
24. American Psychological Association, April 2024. *Potential Risks of Content, Features, and Functions. A Closer Look at the Science behind how Social Media Affects Youth*. APA.org
25. De Groote D, Van Ouytsel J. *Digital Stress within Early Adolescents' Friendships – A Focus Group Study from Belgium*. Telematics and Informatics. 2022;73:101877.
26. Fowler AG. *Snapchat Tried to Make a Safe AI. It Chats with Me about Booze and Sex*. The Washington Post [Internet]. 2023 March 14 [cited 2024 Sept 12]. Available from: https://www.washingtonpost.com/technology/2023/03/14/snapchat-myai/
27. Author unknown. *Snapchat's My AI: Why Is It Controversial?* BBC [Internet]. [cited 2024 Sept 12]. Available from: https://www.bbc.co.uk/bitesize/articles/zck4jfr
28. (28) Baron B, Musolesi M. *Where You Go Matters: A Study on the Privacy Implications of Continuous Location Tracking*. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2020;4(4):Article 169.
29. Karanja A, Engels DW, Zerouali G, Francisco A. *Unintended Consequences of Location Information: Privacy Implications of Location Information Used in Advertising and Social Media*. SMU Data Science Review. 2018;1(3):13.
30. Dunn TR, Langlais MR. *"Oh, Snap!": A Mixed-Methods Approach to Analyzing the Dark Side of Snapchat*. The Journal of Social Media in Society. 2020;9(2):69-104.
31. David ME, Roberts JA. *The Dual Nature of Social Media: Examining the Direction of Causal Flow Between Fear of Missing Out and Social Media Use*. Cyberpsychology, Behavior, and Social Networking. 2023;26(12):881-5.
32. LeBouef S, Dworkin J, Park E. *Family Cohesion Using Snapchat: A Qualitative Study of College Students' Experiences*. Journal of Applied Youth Studies. 2023;6(4):197-211.
33. Yardi S, Bruckman A, editors. *Social and Technical Challenges in Parenting Teens' Social Media Use*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2011.
34. Snap Inc. *Expanding Our In-App Parental Tools 2024* [Available at: https://values.snap.com/en-GB/news/expanding-our-in-app-parental-tools-2024]
35. Snap Inc. *Transparantie over Snapchat-advertenties 2024* [Available at: https://values.snap.com/nl-NL/privacy/ads-privacy]
36. Snap Inc. *Privacybeleid. Met ingang van: 26 februari 2024*. [Available at: https://values.snap.com/nl-NL/privacy/privacy-policy] A summary of this information is also available, as well as a video in which their privacy policy is summarised. They also provide information about privacy per product (e.g., Snap, Chat, My AI, and so on).
37. Snap Inc. *Veelgestelde vragen over de European Digital Services Act 2024* [Available at: https://help.snapchat.com/hc/nl/articles/18416565314964-Veelgestelde-vragen-over-de-European-Digital-Services-Act
38. Haidt J. *The Anxious Generation. How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness*. New York [USA]: The Penguin Press; 2024.
39. Author unknown. *Commission Opens Proceedings against TikTok under the DSA regarding the Launch of TikTok Lite in France and Spain, and Communicates its Intention to Suspend the Reward Programme in the EU*. European Commission Press Release [Internet] 2024 April 22 [cited 2024 Sept 12]. Available from: https://ec.europa.eu/commission/presscorner/detail/fen/ip_24_2227
40. Author unknown. *Commission Opens Formal Proceedings against Meta under the Digital Services Act Related to the Protection of Minors on Facebook and Instagram*. European Commission Press Release [Internet]. 2024 May 16 [cited 2024 Sept 12]. Available from: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2664
41. Author unknown. *Commission Sends Preliminary Findings to X for Breach of the Digital Services Act*. European Commission Press Release [Internet] 2024 July 12 [cited 2024 Sept 12]. Available from: https://ec.europa.eu/commission/presscorner/detail/ov/ip_24_3761

Maastricht University
STUDIO EUROPA MAASTRICHT

Policy Brief

Reducing misinformation and conspiracy theories on social media

Prof. Jan-Willem van Prooijen

Summary

Policy-makers often focus on algorithms in their attempts to reduce the spread of misinformation and conspiracy theories online. However, the main reason why misinformation and conspiracy theories proliferate on social media is because human users decide to share it. One of the main goals of the Digital Services Act is to compel providers of digital services, specifically Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs), to enhance measures against misinformation online. To do so effectively, it is important to understand the motivations of people to be active on social media.

People often share false information to serve their identity needs and appease their in-group. For example, false information that disparages political opponents may gain so-called likes and other forms of social approval from like-minded others. Research suggests that believing and sharing misinformation is often not due to incompetence or the intention to purposefully mislead others; rather, it is mostly due to people's attention being focused on social connections instead of accuracy. Shifting people's focus on the possible accuracy or inaccuracy of information, therefore, can reduce their belief in misinformation and their willingness to share it.

This policy brief reviews interventions that successfully shift people's focus to accuracy. One intervention is warning labels that are presented simultaneously with misinformation. Such warning labels can be quite effective if implemented correctly. Moreover, interventions either before (prebunking) or after (debunking) encountering misinformation can be effective, although the effects of these interventions tend to be small and decrease over time. Even though emotions and identity needs are primary reasons why people believe misinformation and conspiracy theories, raising public awareness of possible inaccuracies, as well as rationally refuting and correcting such false information, does make a difference for many citizens.

Reducing misinformation and conspiracy theories on social media

Misinformation and conspiracy theories proliferate on social media. Common examples of misinformation and conspiracy theories are health-related, such as the frequently posted false claim that childhood vaccines can cause autism. Relatedly, during the COVID-19 pandemic many messages circulated asserting that mRNA vaccines can change people's DNA, or that pharmaceutical companies and national governments deliberately hide dangerous side effects of vaccination (1-3).

Misinformation and conspiracy theories have a myriad of harmful consequences. The above examples undermine public health, as for instance is reflected in reduced vaccination rates and a revival of dangerous childhood diseases (e.g., whooping cough and measles). Also, online misinformation and conspiracy theories can polarise and radicalise citizens. For instance, a common meme on far-right message boards is the Great Replacement Theory – the conspiracy theory that there is a secret plot to gradually replace the White population in Western nations with immigrants and Muslims. This conspiracy theory has been associated with various far-right terrorist attacks, including the El Paso shootings, Breivik's attacks in Oslo and Utøya, and the 2019 Christchurch terrorist attacks (4).

A key goal of the Digital Services Act (DSA) is to prevent such harmful consequences by reducing the spread of misinformation online. The DSA more generally seeks to create a safe digital space across the EU where the basic rights of all users of digital services are protected. This includes stipulating specific rules that compel Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) to enhance measures against online misinformation. Together with related regulatory instruments, most notably the 2022 Strengthened Code of Practice on Disinformation, the DSA provides a legal framework to reduce misinformation online. To function effectively, however, it is important to understand the main reasons why misinformation and conspiracy theories circulate in the online environment.

Policy-makers, journalists and the general public often attribute the online spread of misinformation and conspiracy theories to algorithms (i.e., the filter bubble). This is misguided, however. While algorithms do contribute to the spread of misinformation, accumulating research underscores that their role is quite limited. A large-scale study on YouTube (analysing over 300,000 accounts) showed that both users' own online search behaviour and links received from other users on social media platforms are far better predictors of their decision to watch extremist videos that contain misinformation and conspiracy theories than algorithmic recommendations (5). Another study indicated that algorithms mostly recommend dubious content from alternative or extremist YouTube channels to users that are regularly active on those channels anyway. But even among these users, links shared by other human users on social media were more important than algorithmic recommendations in predicting what content they decided to watch (6). A recent study on the online consumption of far-right content across a broad range of social media channels also concluded that the role of algorithms is more limited than commonly assumed (7).

Instead, the main reason why misinformation and conspiracy theories circulate online is because they are being spread by human users. For instance, a large-scale study of misinformation on Twitter (now X) underscored that false news spreads farther, faster and more broadly than true news. This was not due to bots, however. False news did so well on Twitter because many human users decided to share it (8). Moreover, a study analysing over 10 million United States (US) Facebook accounts concluded that human choices far exceeded algorithmic ranking in shaping users' decisions about what content to watch and share. People specifically were likely to select content that aligned with their own beliefs and political ideologies, while avoiding content that opposed their views (9). The goal of the current contribution, therefore, is to illuminate the reasons why people share false information online, and what interventions may reduce the spread of misinformation and conspiracy theories.

Why do people share misinformation online?

The DSA seeks to increase security by reducing the amount of misinformation that circulates online. To achieve this goal, I propose that it is important to understand the psychological motivations of people to be active on social media. While seeking and sharing information is one of the reasons why people are on social media, another (and possibly more important) reason is to acquire, maintain, and perpetuate social connections with others. For example, one study examined the motivations of users in anti-vaccination Telegram groups by analysing their messages. Besides sharing information, also emphasising a shared identity appeared to be an important motivation for people to be active on this platform (2).

Such a shared identity appeals to people's desire to form social connections and be part of meaningful groups. This *need to belong* has been referred to as a basic psychological need that helps to explain both prosocial and antisocial behaviour (10). People's need to belong for instance drives them to spend time and resources to help other group members; however, it may also drive them to display hostility towards different groups, particularly if these groups have goals that are incompatible with one's own group (e.g., anti-vaccination group members sharing hostile messages about national health authorities [2]).

The design of social media platforms helps users satisfy their need to belong in various ways. It is relatively easy to form new connections with like-minded people, it is relatively easy to stay in touch with acquaintances, and it is possible to give each other signs of social approval for their contributions (e.g., likes). It can be inferred that seeking or sharing truth is not people's only (and arguably not people's main) motivation to be active on social media.

Consequently, people may sometimes share dubious or false content for social identity motives. For instance, a member of a far-right online group may be tempted to post harmful (but false) information about left-wing politicians to acquire social approval from like-minded other users (e.g., likes, reposts and positive comments). In one set of studies, US participants (including both Democrats and Republicans) judged a range of news headlines, some true and some false (11). These headlines were either Democrat-leaning (e.g., supporting Democrats or disparaging Republicans) or they were Republican-leaning (e.g., supporting Republicans or disparaging Democrats). After each headline, participants were asked if the headline was accurate or inaccurate, and if they would share it on social media. Half of the participants could earn a bonus payment depending on how many news headlines they correctly classified as accurate or inaccurate. Results indicated that financially incentivising truth increased people's accuracy and decreased their willingness to share false news headlines. In another of their studies, these researchers also included a condition that incentivised participants to share articles that they believed would be liked by their political in-group; this intervention decreased accuracy.

These findings suggest that the motivations for truth and for a positive social identity do not always align. People sometimes share false information not because they are incapable of identifying it as false, but because they are primarily motivated to appease their in-group. People often are tempted to share information of dubious quality if it reflects well on their own group or poorly on a competing group.

Other studies suggest that believing false information is associated with a general tendency to believe weak claims (i.e., reflexive open-mindedness [12]). For instance, one study revealed that low levels of analytic thinking and a high tendency to see a deeper meaning in nonsense statements predict an increased likelihood of believing false information. Accumulating research indicates, however, that such reflexive open-mindedness does not necessarily imply an impaired mental capacity to distinguish true from false news, nor does it imply the malevolent intention to purposefully mislead others. Instead, it largely seems to imply a lack of attention for the possibility that a news item may be false and increased attention for other factors (such as appeasing other users that are part of one's identity). For instance, asking people to judge the veracity of news headlines (thereby shifting their attention to the possible accuracy or inaccuracy of the news) subsequently increases the quality of news that they decide to share (13).

To summarise, while the Digital Services Act focuses on reducing misinformation on social media, an important motivation for people to be active on social media is social connections; and such social identity motives sometimes undermine the quality of information that people decide to share. Importantly, sharing misinformation is often not due to incompetence or bad intentions but to a lack of attention to the possibility that a news item may be false. This insight provides important tools for reducing the spread of misinformation and conspiracy theories online.

Interventions

The above insights suggest that making people actively reflect on the veracity of information before they share it is likely to reduce the amount of misinformation and conspiracy theories online. In one study, researchers first selected 5,379 Twitter users who had previously shared links of far-right sites that are known for frequently producing false information (Breitbart.com and Infowars.com). They subsequently sent these users a message asking them to judge the veracity of a single, non-political headline. Results indicated that this intervention increased the correctness of news that these users shared in the subsequent 24 hours (13). These findings support the notion that shifting people's attention to accuracy decreases their tendency to share false news online. Importantly, a single accuracy prompt was sufficient to reduce the amount of misinformation that these users shared in the next 24 hours. This study suggests that interventions that (even if only occasionally) invite users to consider the possibility that news items may be false decreases their likelihood of sharing less accurate information.

In practice, such attention to accuracy may be accomplished by making modifications to the design of social media platforms. One way in which social media companies implement this is by flagging dubious content with warning labels. This practice is also addressed in the 2022 Strengthened Code of Practice on Disinformation, which provides extensive legal commitments allowing users to flag unreliable content as false. Warning labels do not censor any information but highlight information that fact-checkers have identified as likely false. Such warning labels thus shift users' attention to the possible inaccuracy of information before they share it.

The effectiveness of these warning labels has often been debated by policy-makers, journalists and influencers, but what does scientific research have to say about them? As it turns out, accumulating research supports the notion that warning labels to flag dubious content are generally effective in reducing the likelihood that people believe in misinformation or decide to share it, including misinformation that appeals to their political in-group. However, the size of the effect depends on how the warning labels are implemented (14).

For example, warning labels are more effective if they are specific and precise. This includes flagging a specific news item as false or flagging a specific news source as unreliable. Consistently, labelling information as specifically 'false' is more effective than labelling it with a more general term like 'disputed'. Warning messages that are insufficiently specific (i.e., those that alert people to the general possibility that some information they encounter online might be false) have the drawback that they can lead people to also judge true information as false (15). Likewise, lack of precision (e.g., incorrect warning labels) may backfire and reduce people's trust in true information (16). Moreover, warning labels are more effective to the extent they are more clearly visible. Finally, users are sensitive to the source of the warning label (e.g., fact-checkers, crowd judgments or AI). Generally, warning labels produced by fact-checkers are most effective, although crowd judgments or AI can also produce effective warning labels (for a more elaborate overview of these guidelines see [14]).

A defining feature of warning labels is that people receive them simultaneously with the misinformation. Other interventions, however, may focus people's attention on the accuracy of information either after or before encountering it. A common intervention after encountering false information is *debunking*, defined as rationally refuting misinformation and conspiracy theories using facts, logic and reason. Such debunking can be implemented in many ways, including informing the public why popular pieces of misinformation are false, providing links to fact-checking sites, spreading accurate information through social media and engaging in a personal correspondence with concerned citizens.

Debunking regularly raises concerns among policy-makers due to the possibility of it backfiring; specifically, that a debunking attempt can actually strengthen people's belief in misinformation (e.g., due to a deep-rooted distrust in authorities, people may perceive the debunking effort as a deliberate attempt to mislead them). However, a large body of research has indicated that this concern is mostly unwarranted. Backfire effects do exist but are extremely rare, and most typically, debunking efforts have a small but consistently positive effect in reducing people's belief in misinformation and conspiracy theories (17). This suggests that fact-checking and other forms of debunking, combined with efforts to ensure that many people encounter the debunking of false information online, contributes to the goal of the Digital Services Act to reduce citizens' belief in false information.

A common intervention before encountering false information is so-called *prebunking*. Prebunking is generally seen as a psychological form of inoculation; by receiving small doses of misinformation and conspiracy theories, people subsequently become less susceptible to such false information when they encounter it. Prebunking may, for instance, prepare citizens for specific misinformation while making them aware how and why such false information is misleading. Prebunking may also focus on teaching people to recognise the manipulative tricks that are generally used by unreliable platforms or influencers when spreading misinformation and conspiracy theories. As such, prebunking efforts address the goal of the 2022 Strengthened Code of Practice on Disinformation to enhance media literacy and critical thinking among citizens. Prebunking often (but not necessarily) takes place online, and existing initiatives include apps that train people's skills in recognising misinformation.

Research supports the effectiveness of prebunking, but a drawback is that these effects tend to wear off over time (18). Moreover, little is yet known about the specific strengths and weaknesses of different prebunking initiatives. While speculative at this point, possibly training people to recognise false information at a younger age (e.g., in high school curricula) might lead to more internalisation, and hence, a longer-lasting effectiveness. Future research would need to examine this possibility.

Concluding remarks

This policy brief has argued that not algorithms, but human users are the main reason why misinformation and conspiracy theories circulate online. Interventions hence need to take the psychology of human online behaviour into account. Two take-aways of this policy brief are particularly important. First, many people who share misinformation and conspiracy theories online are neither incompetent nor have bad intentions. Instead, their attention is focused on different needs that they seek to satisfy online, particularly the need for connectedness and a shared identity with like-minded others (2, 11).

Second, interventions can seize on this insight by focusing people's attention on the accuracy or inaccuracy of information (13). An awareness that information might be false reduces people's tendency to believe or share it, even if it is information that would appeal to their like-minded social network. Such interventions can take place simultaneously with receiving misinformation (e.g., warning labels that flag misinformation as false [14]), but also before or after encountering misinformation and conspiracy theories (prebunking and debunking [17, 18]).

These insights and recommendations contain a paradox; it is well-known that not reason but emotions (e.g., frustration, anxiety) increase people's susceptibility to misinformation and conspiracy theories. Moreover, people often cherry-pick information online to be able to uphold their false but ideologically convenient beliefs, and some citizens have fallen so deep down a rabbit hole of misinformation and conspiracy theories that they do not appear sensitive to any form of scientific evidence, reason or logic (19-23). How can refuting false information through fact-checking and reason persuade people given these considerations?

This paradox can be resolved by not making a caricature of citizens susceptible to false information and conspiracy theories. One needs to keep in mind that only a small minority is responsible for disseminating a large majority of misinformation online (so-called super-spreaders [18]). These super-spreaders for instance include influencers that are popular in radical groups, or people active on conspiracist message boards. As these people are heavily committed to spreading a distorted view of reality – possibly due to genuine belief, or possibly due to other incentives (e.g., financial, status) – the current recommendations are unlikely to change these super-spreaders.

However, and importantly, the current recommendations are likely to influence the many more moderate users who encounter the messages of these super-spreaders online. Most of these users are regular citizens who are not incompetent or have malicious intent but may have understandable questions, uncertainties and worries about distressing societal events such as a pandemic or a war. These people may be susceptible to the misinformation and conspiracy theories that they encounter online, but they may also be susceptible to the voices of science and reason. For this silent majority, raising awareness of possible inaccuracies and providing rational arguments does make a difference.

Policy recommendations

1. While policy-makers often are concerned with algorithms, the main reason why misinformation and conspiracy theories proliferate online is because human users decide to share such false information. The psychology of human online behaviour is hence important to consider when developing policy to reduce misinformation online.

2. A key reason why many human users share false information is neither incompetence nor bad intentions, but inattentiveness to the possibility that the information they share might be false. Often people are focused on identity needs instead (e.g., gaining likes or other signs of social approval from like-minded others). Research has found promising results for interventions that shift people's attention to the possible accuracy or inaccuracy of information.

3. Warning labels that flag content as false are effective in reducing the spread of misinformation and conspiracy theories online if they are implemented well. The 2022 Strengthened Code of Practice on Disinformation provides extensive legal commitments allowing users to flag unreliable content as false. Warning labels are most effective if they are specific, precise, clearly visible and produced by fact-checkers.

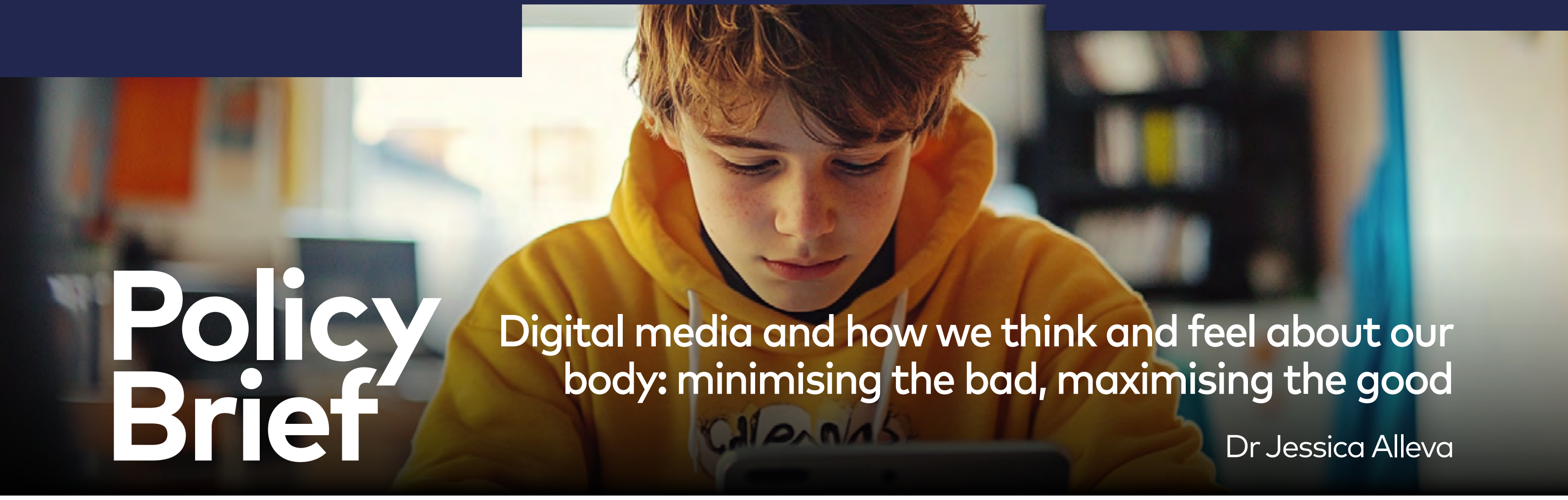
4. Rationally refuting false information through facts, logic and reason (debunking) can also be an effective intervention, although often the effects are small. Debunking efforts can take many forms, including informing the public why popular pieces of misinformation are false, providing links to fact-checking sites, spreading accurate information through social media and engaging in a personal correspondence with concerned citizens. Contrary to common concerns, debunking efforts rarely backfire.

5. Teaching people how to recognise falsehoods before they encounter it (prebunking) also can be effective. The effects of prebunking decrease over time, however, suggesting that a continuous effort is needed. Including prebunking training in high school curricula might be a promising intervention that requires further research.

Author information:
Jan-Willem van Prooijen is Professor of Radicalisation, Extremism and Conspiracy Thinking at Maastricht University. He also holds affiliations with the Department of Experimental and Applied Psychology at Vrije Universiteit Amsterdam, and the Netherlands Institute for the Study of Crime and Law Enforcement (NSCR).

References

1. Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018). *The psychological roots of anti-vaccination attitudes: A 24-nation investigation*. Health Psychol, 37, 307-315. DOI: 10.1037/hea0000586
2. Schlette, A., Van Prooijen, J.-W., Blokland, A., & Thijs, F. (2023). *Information, identity, and action: The messages of the Dutch anti-vaccination community on Telegram*. New Media Soc. https://doi.org/10.1177/14614448231215735
3. Van Prooijen, J.-W., & Böhm, N. (2024). *Do conspiracy theories shape or rationalize vaccination hesitancy over time?* Soc Psychol Personal Sci, 15, 421-429.
4. Obaidi, M., Kunst, J., Ozer, S., & Kimel, S. Y. (2022). *The "Great Replacement" conspiracy: How the perceived ousting of Whites can evoke violent extremism and Islamophobia*. Group Process Intergroup Relat, 25, 1675-1695. https://doi.org/10.1177/13684302211028293
5. Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). *Examining the consumption of radical content on YouTube*. Proc Natl Acad Sci, 118, e2101967118.
6. Chen, A., Nyhan, B., Reifler, J., Robertson, R. E., & Wilson, C. (2023). *Subscriptions and external links help drive resentful users to alternative and extremist YouTube videos*. Sci Adv, 9, eadd8080.
7. Verwey-Jonker Instituut (2023). *Rechtsextremistische radicalisering op sociale mediaplatforms? Ontwikkelingspaden en handelingsperspectieven*. WODC rapport.
8. Vosoughi, S., Roy, D., & Aral, S. (2018). *The spread of true and false news online*. Science, 359, 1146-1151.
9. Bakshy, E., Messing, S., & Adamic, L. A. (2015). *Exposure to ideologically diverse news and opinion on Facebook*. Science, 348, 1130-1132.
10. Baumeister, R. F., & Leary, M. R. (1995). *The need to belong: Desire for interpersonal attachments as a fundamental human motivation*. Psychol Bull, 117, 497-529.
11. Rathje, S., Roozenbeek, J., Van Bavel, J. J., & Van der Linden, S. (2023). *Accuracy and social motivations shape judgements of (mis)information*. Nat Hum Behav, 7, 892-903.
12. Pennycook, G., & Rand, D. G. (2020). *Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking*. J Pers, 88, 185-200.
13. Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). *Shifting attention to accuracy can reduce misinformation online*. Nature, 592, 590-595.
14. Martel, C., & Rand, D. (2023). *Misinformation warning labels are widely effective: A review of warning effects and their moderating features.* Curr Opin Psychol, 54, 101710.
15. Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G. … & Nyhan, B. (2019). *Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media*. Polit Behav, 42, 1073-1095.
16. Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J. R., Olin, J., Quinn Ross, M., & Szafran, J. (2021). *Fake claims of fake news: Political misinformation, warnings, and the tainted truth effect.* Polit Behav, 43, 1433-1465.
17. Nyhan, B. (2021). *Why the backfire effect does not explain the durability of political misperceptions.* Proc Natl Acad Sci, 118, e1912440117.
18. Van der Linden, S. (2022). *Misinformation: Susceptibility, spread, and interventions to immunize the public*. Nat Med, 28, 460-467. doi: 10.1038/s41591-022-01713-6
19. Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). *Understanding conspiracy theories*. Adv Pol Psychol, 40, 3-35.
20. Hornsey, M. J., Bierwiaczonek, K., Sassenberg, K., & Douglas, K. M. (2023). *Individual, intergroup and nation-level influences on belief in conspiracy theories*. Nat Rev Psychol, 2, 85-97. https://doi.org/10.1038/s44159-022-00133-0
21. Van Prooijen, J.-W. (2022). *Injustice without evidence: The unique role of conspiracy theories in social justice research*. Soc Justice Res, 35, 88-106.
22. Van Prooijen, J.-W. (2024). *Group-oriented motivations underlying conspiracy theories*. Group Process Intergroup Relat. https://doi.org/10.1177/13684302241240696
23. Wagner-Egger, P., Bangerter, A., Delouvée, S., & Dieguez, S. (2022). *Awake together: Sociopsychological processes of engagement in conspiracist communities*. Curr Opin Psychol, 47, 101417. https://doi.org/10.1016/j.copsyc.2022.101417



Policy Brief

Digital media and how we think and feel about our body: minimising the bad, maximising the good

Dr Jessica Alleva

Summary

The rapid digitalisation of society has transformed how individuals view the world and themselves. One area where this transformation is particularly clear is with regards to body image, with digital media playing a role in how people think and feel about their our own bodies and what they consider to be beautiful or not.

This policy brief first describes what body image is and how it is related to physical and mental health. Next, it explores the complex relationship between digital media and body image based on decades of research, including how and why digital media can negatively affect body image, but also how and why digital media can positively affect body image. This knowledge is used to create the Body Image Decision Tool for Digital Media which stakeholders can apply to help determine the impact of digital content on body image. Next, applications to the Digital Services Act and additional considerations are outlined. This policy brief concludes with policy recommendations that will guide efforts towards minimising negative body image and optimising positive body image for a greater number of people.

What is body image and why should we care?

Body image describes the thoughts and feelings that people have about their body, which may pertain to how their body looks but also to how their body functions (see Box 1) (1, 2). Importantly, body image is subjective: how a person thinks and feels about their body is not necessarily the same as how their body objectively looks and functions.

Negative body image refers to having negative or unfavourable thoughts and feelings about one's body (e.g., dissatisfaction, disgust). Negative body image has been identified as a global health concern because it is prevalent among people around the globe and has serious consequences (3). For example, negative body image is related and contributes to withdrawal from joyful physical activity and meaningful life activities (e.g., hobbies, education), depression, anxiety and low self-esteem; and it is the number one risk factor for eating disorders (3).

Box 1: Appearance vs. functionality

Most people know what is meant by *appearance* or *looks*. This pertains to visible physical characteristics of a person such as their body size, muscularity, height and skin colour.

Body functionality is a less familiar term. It refers to all of the things the body is able to do. Body functions may fall into one of six domains: (1) physical capacities (e.g., walking, stretching), (2) internal processes (e.g., digesting food, healing from a cold), (3) bodily senses and sensations (e.g., seeing, hearing), (4) creative endeavours (e.g., dancing, drawing), (5) communication with others (e.g., hugging, body language) and (6) self-care activities (e.g., showering, bathing).

Traditionally, researchers have focused on negative body image given these links with ill-being. However, more and more research is being conducted on positive body image, too. Positive body image refers to having an overall sense of gratitude, acceptance and respect for one's body (2). Mounting evidence shows that positive body image is related and contributes to numerous aspects of physical and mental health, such as engagement in joyful physical activities, adaptive eating behaviour, self-esteem, positive mood and quality of life (4, 5).

Importantly, negative body image and positive body image are not opposite ends of the same spectrum. That is, people can experience aspects of both negative body image and positive body image concurrently. For example, someone can feel dissatisfied with their weight and yet feel grateful for the health of their body. This has important implications for interventions because we can strategise both about how to lower negative body image and how to enhance positive body image—thereby optimising the opportunities to impact health and well-being.

In this spirit, I first describe how digital media contribute to negative body image. Then, I turn to how digital media contribute to positive body image. Using this knowledge, I identify potential applications to the Digital Services Act (DSA) and key policy recommendations.

To bring the research evidence to life, Box 2 presents a case study of Greta, highlighting how digital media contributed to her negative body image early in life, and how digital media also helped her transition to a more positive body image over time.

Box 2: Greta's story

When Greta was 10 years old she developed a skin condition called vitiligo. At first, she felt proud of the way that she looked. She liked cheetahs and could run fast, so she felt the white patches on her skin likened her to this animal. Then, her peers began making fun of the way that she looked, and she started to feel insecure about her appearance. In addition, everyone on social media looked perfect to her—not only the celebrities she followed, but also the 'flawless' selfies of her peers. In comparison, Greta felt different and ugly. When she posted pictures of herself on social media, some of her peers made hurtful comments. To make matters worse, when she began gaining weight in puberty, her grandmother warned her, 'You already have vitiligo. You can't be fat, too. No one will want to date you'.

For years, Greta felt unhappy with her appearance and this led her to withdraw from her hobbies, like running and swimming, because she did not want to be seen. She covered her skin with baggy clothing and heavy makeup. She started an extreme diet and excessively exercised to try and lose the weight that she gained in puberty.

One day at school, Greta collapsed in gym class. After a consultation, the school nurse suspected that Greta had an eating disorder. Together with her parents and general practitioner, Greta found the help of a qualified clinical psychologist.

In treatment, Greta began thinking critically about the messages she had learned about beauty growing up and started challenging her unhealthy thoughts and behaviours. Her therapist also suggested finding online communities of people with visible differences. Greta followed influencers on Instagram who also had vitiligo and other visible differences, and who promoted body acceptance and care. She realised that she was not alone, and that if other people could accept their bodies, she could too. She formed friendships with other young people with visible differences, and they created a group to keep in touch and support one another.

Over time, Greta started feeling more positively about her body, and realised that she was so much more than her appearance. She created her own blog about her experiences of vitiligo and eating disorder recovery, began writing for other public outlets (e.g., the school newspaper) and did some public speaking engagements. Other people felt inspired by Greta's story and encouraged her to keep doing the important work that she is doing.

Source: Greta's story is based on a synthesis of narratives from my research on how women transition from a negative body image toward a

How do digital media worsen body image?

To date, hundreds of studies have investigated how digital media can contribute to negative body image. Typically, these studies have focused on image-based social media platforms such as Facebook, Instagram and TikTok.

Overall, time spent on digital media is related to negative body image. Breaking this down further, exposure to appearance-related content on digital media is particularly problematic (see Box 3).

Box 3: Appearance-related content

Appearance-related content refers to content that portrays and upholds societal body ideals or expectations for how a body should look and/or function. Body ideals have become increasingly globalised, with many people around the world striving for very thin/lean and very muscular/toned bodies (3). Body ideals also typically emphasise youth and being able-bodied. Appearance-related content may pertain to images (e.g., images of models who are lean and muscular) and/or text (e.g., a model describing their desire to lose weight, comments that express admiration for a model's leanness).

Research has shown that exposure to appearance-related content on digital media is related to negative body image. For example, people who report higher exposure to appearance-related content on digital media also report higher levels of negative body image.

Importantly, the research has also provided evidence for causality: exposure to appearance-related content on digital media can cause negative body image. For example, in laboratory experiments, people who are exposed to appearance-related content report increases in negative body image compared to people in control groups who are exposed to neutral content. In longitudinal studies that monitor groups of people across time, exposure to appearance-related content on digital media predicts increases in negative body image later in time. For recent reviews of this research area, see (6–10).

Interestingly, some research has also investigated the effects of viewing idealised images of oneself (for example, using so-called beautifying filters on digital platforms like Instagram) and these types of images also contribute to negative body image (11, 12).

Understanding the underlying mechanisms of the effects of digital media on negative body image is crucial to knowing how to intervene. Based on leading theories in the body image field and extensive research (9, 10, 13–17), there are three key processes that can explain this relationship.

1. Internalisation of body ideals: internalisation of body ideals extends beyond the awareness of societal body ideals to taking on these ideals as one's own personal ideal (e.g., knowing that thinness is considered beautiful vs. striving to become thin). Appearance-related content leads to the internalisation of body ideals and, in turn, the internalisation of body ideals leads to a more negative body image.

2. Body comparisons: appearance-related content leads to comparing one's own body to the body ideal. In turn, body comparisons contribute to a more negative body image.

3. Self-objectification: self-objectification refers to the tendency to view and evaluate one's body based predominantly on how it looks (rather than on what it can do). More broadly, it also refers to viewing and evaluating one's overall self from an appearance-based perspective rather than based on other qualities (e.g., personality, education). Appearance-related content can reinforce an emphasis on appearance, which leads to higher self-objectification and, consequently, to a more negative body image.

It is important to underscore that feeling poorly about one's own body does not motivate individuals to engage in adaptive health activities (e.g., physical activity, adaptive eating behaviours). In contrast, as described above, negative body image leads to disengagement from adaptive health activities (3).

How can digital media improve body image?

Given that research on positive body image has emerged relatively recently, there is comparatively less research on how digital media can affect positive body image. The initial evidence supports that exposure to appearance-related content on digital media contributes to decreases in positive body image (18, 19). Promisingly, there is evidence to show that some types of digital media can contribute to higher positive body image. This evidence pertains mainly to research on body positivity content (see Box 4) (20). Overall, the research shows that exposure to body positivity content on digital media is related and contributes to a more positive body image (and to reductions in negative body image). For example, in laboratory experiments, participants who are exposed to body positivity content on digital media report increased positive body image compared to participants in control groups who are exposed to neutral content or to traditional appearance-related content. For reviews of this research, see (6, 21).

Box 4: Body positivity content

Body positivity content refers to content (images and/or text) that depicts body diversity and promotes respect and care for all bodies regardless of characteristics such as size, shape, age, gender and ability. Body positivity content often promotes an emphasis on appreciation for the body's functionality and for valued aspects of the overall Self such as personality and education.

Not all body positivity content is created equal! Research has shown that body positivity content tends to align with the description above, but some content may contain conflicting messages (e.g., pairing the message 'Every body is worthy of respect' with only images of very thin women). This is because in digital media spaces essentially anyone can create content and label it as 'body positivity'. Therefore, when it comes to creating or disseminating content that is likely to promote positive body image, it is important to ensure that the content truly aligns with the definition of body positivity content.

Qualitative research has also shown that digital media can play an important role in helping people to transition from a predominantly negative body image to a predominantly positive body image over time (22)(see also Box 2). In this research, participants described body positivity content, as well as seeing greater representation and diversity across the media landscape overall (e.g., in the music industry, television shows), as influential. People also used digital media to find and connect with other people who shared similar experiences and characteristics (e.g., visible difference, sexual orientation). This helped them to feel less alone and led to meaningful social connections where people felt supported and accepted for who they are (regardless of how they look). Digital media was also used to advocate and help other people, for example, via blogging about mental health and posting one's own body positivity content online.

The three processes described above can also explain the beneficial effects of exposure to body positivity content online. Namely, this content has the potential to decrease levels of (1) internalisation of body ideals (e.g., via the portrayal of diverse bodies or messages that critique body ideals), (2) body comparisons (e.g., by encouraging viewers to accept and appreciate their own body) and (3) self-objectification (e.g., by directing viewers to appreciate their body functionality and other aspects of themselves) thereby leading to lower negative body image and to higher positive body image.

An important theory from the field of positive body image (i.e., the acceptance model of intuitive eating) (23) highlights an additional mechanism called unconditional body acceptance by others. That is, when people feel that others unconditionally accept their body, they are more likely to develop a positive body image and engage in adaptive health behaviours. Applied to the context of digital media, body positivity content can lead viewers to perceive that others would accept their body, thereby helping them to feel more positive about their body.

Last, a common myth about body positivity content is that it may promote disengagement from health behaviours. The assumption is that seeing diverse people who are comfortable and happy in their bodies could encourage viewers to 'let themselves go'. As described above, body positivity content contributes to a more positive body image, and positive body image contributes to taking care of one's body, for example, via joyful physical activity and adaptive eating behaviour (4,5). There is no empirical evidence to support that body positivity content will discourage engagement in health behaviours.

The body image decision tool for digital media

Box 5 presents the Body Image Decision Tool for Digital Media which is grounded in the extensive evidence base summarised above. Stakeholders can use the questions in Box 5 to determine how digital content is likely to affect users' body image. Answers on the left side tip the scale toward the likelihood of higher negative body image and lower positive body image, and those on the right tip the scale toward the likelihood of lower negative body image and higher positive body image. Ideally, stakeholders should aim for 'no' to Questions 1-3 and 'yes' to Question 4. Box 5 also presents four imagined scenarios to illustrate how the decision tool can be applied.

Box 5: The body image decision tool for digital media

The Questions

	Is the digital content (including images and/or text) likely to...	
Yes	1. Reinforce the internalisation of body ideals?	No
Yes	2. Encourage body comparisons?	No
Yes	3. Reinforce self-objectification?	No
No	4. Convey unconditional body acceptance?	Yes
Higher negative body image Lower positive body image		Higher positive body image Lower negative body image

Example applications In the first set of scenarios below, users are likely to experience higher negative body image and lower positive body image (i.e., Questions 1-3 would be 'yes', Question 4 would be 'no'). In the second set of scenarios, users are likely to experience lower negative body image and higher positive body image i.e., (Questions 1-3 would be 'no', Question 4 would be 'yes').

Likely to promote higher negative body image and lower positive body image:

1. Across their website and social media, a sportswear brand portrays images of models that mainly reflect globalised body ideals; there is little variation in body size and other physical characteristics. The models are often passively posed, with an emphasis on specific body parts (e.g., abdominals, buttocks). The text emphasises the importance of sportswear that is physically flattering. The text frames physical activity as a means to lose weight, gain muscularity and to look better. The sportswear is not available in larger sizes.
2. A teenager starts following one of their favourite celebrities on Instagram and TikTok. The celebrity films their workout regimes and the supplements they use to achieve rapid weight loss and greater muscle tone. Many of these posts are sponsored content with links to purchase supplements with a promotional code. Soon, the teenager begins noticing that their Instagram and TikTok feeds are inundated with suggested accounts of other influencers who promote a similar lifestyle and products with little variation in content.

Likely to promote lower negative body image and higher positive body image:

1. Across their website and social media, a sportswear brand portrays images of models that are diverse in terms of body shape and size, skin colour, age, visible difference and physical ability. The models are pictured enjoying a range of physical activities. The text emphasises the importance of sportswear that is functional and comfortable. The text frames physical activity as a means to have fun, de-stress, develop a skill and connect with others. The sportswear is available in a wide range of sizes.
2. A teenager starts following an influencer who posts body positivity content on Instagram and TikTok. Across the influencer's account, the images portray body diversity, for example, including people of various body sizes, skin colours, ages and physical abilities. The text emphasises the importance of body respect and care, appreciating one's body functionality and investing in valued life domains such as education and friendships. The influencer integrates aspects of media literacy into their posts, such as how body ideals are unrealistic. Suggested accounts on Instagram and TikTok promote similar content, as well as content that pertains to the user's other, non-body-related interests (e.g., photography, music).

A note on generative artificial intelligence

At the time of writing, the capacities and spread of generative artificial intelligence (AI) are rapidly expanding. Generative AI is being used to create images of people (e.g., AI influencers), and there is reason to believe that many of these images reflect globalised body ideals (24). While research on these technologies and their impact on body image is still needed, the extant evidence base on the effects of digital media on body image and its underlying mechanisms can provide useful knowledge as we navigate these rapid changes. That is, content created via generative AI that depicts globalised body ideals and reinforces internalisation, body comparisons and self-objectification (and minimises perceived body acceptance) will contribute to a more negative body image and lower positive body image.

In contrast, if the content depicts body diversity and minimises internalisation, body comparisons and self-objectification; and in addition, portrays unconditional body acceptance, it will contribute to lower negative body image and higher positive body image. Optimistically, there are efforts emerging that take these points into consideration. For example, Dove recently launched the 'Real Beauty Prompt Playbook' that contains guidelines on creating images of people that are diverse and representative.

Applications to the Digital Services Act

Stakeholders can apply the Body Image Decision Tool for Digital Media (Box 5) to any type of digital content to help determine whether it is likely to minimise harm and maximise good.

Turning to the Digital Services Act (DSA) specifically, many tenets are important for body image. Below, I highlight three key examples with additional considerations for positive impact.

1. The DSA ensures that there is transparency in how users' algorithms are determined and will enable users to opt out of personalised feeds. Further, this information is expected to be given in a way that is easy for users to understand. This is helpful because it will enable users to adjust an algorithm that could be contributing to higher exposure to appearance-related content (and to lower exposure to body positivity content), thus giving them greater control over their user experience and well-being.

Additional considerations: rather than require users to actively seek information about algorithms, expect them to fully understand this information and make an informed decision, it is better to design adaptive environments from the start as default. This is especially important considering that people are not very good at recognising body image problems or know what to do about them (25).

An open question is whether the default experience should be an intentionally adaptive algorithm. For example, adaptive algorithms could promote body positivity content (see Box 4) as well as content that is unrelated to appearance and the body (e.g., an individual's hobbies and education).

2. The DSA makes it illegal to expose people under the age of 18 to targeted and/or sensitive advertising content. In the body image context, examples of particularly sensitive content could be weight loss supplements, fitness supplements and cosmetic procedures.

This regulation is important because it can prevent young people from being exposed to appearance-focused content during a life stage where their body image is being developed, where they are particularly sensitive to others' approval and may lack the sharpened critical thinking skills to resist appearance pressures (26).

Additional considerations: In many contexts, such as cosmetic procedures, the regulation of advertisements for youth is clear. Yet, it becomes challenging to regulate advertisements when body ideals are embedded in advertisements for other products and services. For example, advertisements for a television series may contain exclusively idealised bodies. Unfortunately, it is unlikely that the DSA can regulate content to this extent, and this points towards broader systemic changes that must be made to the media landscape.

Furthermore, one might wonder why individuals above 18 years old should also not be, by default, protected from exposure to advertisements that are very likely to harm their body image according to the evidence base.

3. Very Large Online Platforms (VLOPs) are required to make risk assessments of the impact of their platform's content on users' health. As described in this brief, body image is a central determinant of health. VLOPs should therefore incorporate body image as one of their indicators of users' health. The requirement to create risk assessments can ensure that user health stays at the forefront of VLOP's radars. Over time, risk assessments can be compared, creating accountability. For example, if an assessment indicates that a VLOP is likely contributing to poorer body image multiple years in a row, it could flag that a VLOP is not taking sufficient or effective action and signals the necessity to try other strategies.

Additional considerations: Ideally, risk assessments should be part of every platform's routine no matter how large. Further, an unanswered question is whether VLOPs (and other online platforms) can objectively conduct self-assessments, especially when a negative result could equate to bad publicity and potentially a reduction in profit. It is hoped that stakeholders see these results as an opportunity to make positive, impactful changes.

Conclusions

The DSA aims to protect individuals' fundamental rights and to create a safer online environment. Reimagining the digital media landscape to minimise the exposure to appearance-related content and to maximise the exposure to body positivity and other adaptive content can give individuals greater freedom to live fuller lives that are not encumbered by body concerns. It is hoped that the present policy brief provides guidance in this direction.

Policy recommendations

Based on the extensive evidence base and the preceding discussion pertaining to digital media and body image, the following overarching recommendations are offered:

- Recognise that body image is a key aspect of mental and physical health and prioritise body image.
- When creating and disseminating digital content, ensure that exposure to appearance-related content (Box 3) is minimised and, if applicable, incorporate body positivity content (Box 4). Use the Body Image Decision Tool for Digital Media (Box 5) as a guide.
- Design the features of online environments so that the default minimises the risk of negative body image and maximises the opportunities for positive body image (e.g., feeds that include exposure to body positivity content and diverse content creators).
- Support users in taking steps to protect their body image online (e.g., obtaining information about why specific advertisements are shown to them and content moderation tools).
- Relatedly, where applicable, provide content that increases digital media literacy (e.g., within school curricula). Specifically, digital media literacy programmes provide individuals with the knowledge and skills to analyse, evaluate, produce and participate in social media (27).
  There are evidence-based media literacy programmes that improve body image and may protect users from the effects of appearance-related content, including programmes that focus on social media specifically (28–31).
  Note that research has also investigated the effectiveness of including disclaimer labels on appearance-related content (e.g., reminding viewers that the images have been digitally edited). Overall, the evidence shows that these types of messages do not work and, in some cases, may even worsen body image (for a review, see [28]). Therefore, their use is not recommended.
- Invite body image experts and individuals with lived experience to the table, for example, to minimise blind spots when designing digital content and the features of online platforms.
- Generative AI is a rapidly evolving technology that must be considered with respect to protecting users' body image online (e.g., use of guidelines to create content that minimises harm and maximises positive impact).
- Invest in collaborations between researchers and digital media platforms (e.g., Instagram, TikTok) to continue to investigate the relationship between digital content, body image and health, including how digital content can foster positive body image.

Author Information

Jessica Alleva is an Assistant Professor in the Faculty of Psychology and Neuroscience, Maastricht University. Her research investigates body image (i.e., how individuals think and feel about their bodies), the factors that impact body image (e.g., social media, social relationships), and techniques for improving body image.

Contact: jessica.alleva@maastrichtuniversity.nl

References

1. Alleva JM, Martijn C, Van Breukelen GJP, Jansen A, Karos K. *Expand Your Horizon: A programme that improves body image and reduces self-objectification by training women to focus on body functionality*. Body Image. 2015;15:81–9.
2. Tylka TL, Wood-Barcalow NL. *What is and what is not positive body image? Conceptual foundations and construct definition*. Body Image. 2015;14:118–29.
3. Rodgers RF, Laveway K, Campos P, de Carvalho PHB. *Body image as a global mental health concern*. Cambridge Prisms: Global Mental Health. 2023;10:1–8.
4. Linardon J, Messer M, Tylka TL. *Functionality appreciation and its correlates: Systematic review and meta-analysis*. Body Image. 2023 Jun 1;45:65–72.
5. Linardon J, McClure Z, Tylka TL, Fuller-Tyszkiewicz M. *Body appreciation and its psychological correlates: A systematic review and meta-analysis*. Body Image. 2022 Sep 1;42:287–96.
6. Vandenbosch L, Fardouly J, Tiggemann M. *Social media and body image: Recent trends and future directions*. Curr Opin Psychol. 2022 Jun 1;45:101289.
7. Fardouly J, Vartanian LR. *Social media and body image concerns: Current research and future directions*. Curr Opin Psychol. 2016 Jun 1;9:1–5.
8. Saiphoo AN, Vahedi Z. *A meta-analytic review of the relationship between social media use and body image disturbance*. Comput Human Behav. 2019 Dec 1;101:259–75.
9. Faelens L, Hoorelbeke K, Cambier R, van Put J, Van de Putte E, De Raedt R, et al. *The relationship between Instagram use and indicators of mental health: A systematic review*. Computers in Human Behavior Reports. 2021 Aug 1;4:100121.
10. Ryding FC, Kuss DJ. *The use of social networking sites, body image dissatisfaction, and body dysmorphic disorder: A systematic review of psychological research*. Psychology of Popular Media. 2020 Oct;9(4):412–35.
11. Dijkslag IR, Block Santos L, Irene G, Ketelaar P. *To beautify or uglify! The effects of augmented reality face filters on body satisfaction moderated by self-esteem and self-identification*. Comput Human Behav. 2024 Oct 1;159:108343.
12. Kleemans M, Daalmans S, Carbaat I, Anschütz D. *Picture perfect: The direct effect of manipulated Instagram photos on body image in adolescent girls*. Media Psychol. 2018 Jan 2;21(1):93–110.
13. Fioravanti G, Bocci Benucci S, Ceragioli G, Casale S. *How the exposure to beauty ideals on social networking sites influences body image: A systematic review of experimental studies*. Adolesc Res Rev. 2022 Sep 1;7(3):419–58.
14. Fredrickson BL, Roberts TA. *Objectification theory: Toward understanding women's lived experiences and mental health risks*. Psychol Women Q. 1997 Jun 24;21(2):173–206.
15. Holland G, Tiggemann M. *A systematic review of the impact of the use of social networking sites on body image and disordered eating outcomes*. Body Image. 2016 Jun 1;17:100–10.
16. Moradi B, Huang YP. *Objectification theory and psychology of women: A decade of advances and future directions*. Psychol Women Q. 2008;32:377–98.
17. Thompson JK, Heinberg LJ, Altabe M, Tantleff-Dunn S. *Exacting beauty: Theory, assessment, and treatment of body image disturbance*. American Psychological Association; 1999.
18. Brown Z, Tiggemann M. *A picture is worth a thousand words: The effect of viewing celebrity Instagram images with disclaimer and body positive captions on women's body image*. Body Image. 2020 Jun 1;33:190–8.
19. Alleva JM, Grünjes C, Coenen L, Custers M, Vester P, Stutterheim SE. *A randomized controlled trial investigating two protective filtering strategies to mitigate the effects of beauty-ideal media imagery on women's body image*. Comput Human Behav. 2024 Jun 1;155:108178.

20. Cohen R, Irwin L, Newton-John T, Slater A. *#bodypositivity: A content analysis of body positive accounts on Instagram*. Body Image. 2019 Jun 1;29:47–57.
21. Cohen R, Newton-John T, Slater A. *The case for body positivity on social media: Perspectives on current advances and future directions*. J Health Psychol. 2021 Nov 1;26(13):2365–73.
22. Alleva JM, Tylka TL, Martijn C, Waldén MI, Webb JB, Piran N. *"I'll never sacrifice my well-being again:" The journey from negative to positive body image among women who perceive their body to deviate from societal norms*. Body Image. 2023 Jun 1;45:153–71.
23. Avalos LC, Tylka TL. *Exploring a model of intuitive eating with college women*. J Couns Psychol. 2006 Oct;53(4):486–97.
24. The Bulimia Project. *Scrolling into bias: Social media's effect on AI art* [Internet]. 2023 [cited 2024 Aug 5]. Available from: https://bulimia.com/examine/scrolling-into-bias/
25. Hewitt J, Murray K. *Negative body image mental health literacy in women: Exploring aesthetic and functional concerns and the role of self-objectification*. Body Image. 2024 Mar 1;48:101657.
26. Gattario KH, Frisén A. *From negative to positive body image: Men's and women's journeys from early adolescence to emerging adulthood*. Body Image. 2019 Mar 1;28:53–65.
27. Tamplin NC, McLean SA, Paxton SJ. *Social media literacy protects against the negative impact of exposure to appearance ideal social media images in young adult women but not men*. Body Image. 2018 Sep 1;26:29–37.
28. Tiggemann M. *Digital modification and body image on social media: Disclaimer labels, captions, hashtags, and comments*. Body Image. 2022 Jun 1;41:172–80.
29. Wilksch SM, O'Shea A, Taylor CB, Wilfley D, Jacobi C, Wade TD. *Online prevention of disordered eating in at-risk young-adult women: A two-country pragmatic randomized controlled trial*. Psychol Med. 2018 Sep 1;48(12):2034–44.
30. Bell BT, Taylor C, Paddock D, Bates A. *Digital Bodies: A controlled evaluation of a brief classroom-based intervention for reducing negative body image among adolescents in the digital age*. British Journal of Educational Psychology. 2022 Mar 1;92(1):280–98.
31. McLean SA, Wertheim EH, Masters J, Paxton SJ. *A pilot evaluation of a social media literacy intervention to reduce risk factors for eating disorders*. International Journal of Eating Disorders. 2017 Jul 1;50(7):847–51.

Maastricht University
STUDIO EUROPA MAASTRICHT

Policy Brief

Transparency of personal health data sharing and the Digital Services Act

Dr Visara Urovi

Summary

Personal health data boosts healthcare research – yet, a big amount of these data is controlled by big corporations. This raises transparency and trust issues as customers have little or no information on how their data is collected, stored and re-used.

The Digital Services Act (DSA) aims to enhance digital service safety, accountability and transparency, particularly for large platforms. However, its application to digital health services requires further elaboration. Digital health services handle personal health data and require clear practices to ensure privacy and trust. In this article, we explore how there are new opportunities to shine a light on these services and their use of personal health records via the transparency reporting of the DSA.

Concerns over digital health services include those related to giant tech companies dominating the health data market, the potential of data breaches and the inability to leverage these data at a societal level. A standardised transparency report would support interoperability and improve user trust. As such, it needs to clarify key customer questions on data usage, collection, processing and sharing. Users should be enabled to identify the use of artificial intelligence (AI) and how AI decisions may impact them, to understand secondary data use details and their ability to decide on secondary data sharing.

In conclusion, standardising transparency reporting can become an important tool for digital health services under the DSA to combat data monopolies, ensure informed user consent and support innovation.

Introduction

The Digital Services Act (DSA) is legislation proposed by the European Union aimed at creating safer and more accountable digital services (1) (DSA Legislation). The rules of the DSA are designed to protect consumers and their fundamental rights, define the responsibilities of online platforms, deal with illegal content and end products, achieve greater transparency and encourage growth and innovation.

In the context of transparency requirements of the DSA, standardising the transparency reporting requirements is likely to influence transparency reporting legislation in countries beyond the EU (2).

In this work, we define transparency reporting as a way to provide internal information on matters of public concern (3). Such reporting is made by the companies owning digital services towards the external audience.

A lot of the initial focus of the DSA seems to have been placed on requiring more transparency from large user platforms such as social media or digital market platforms, as these have been found to be potential channels for disinformation and/or customer deception. For example, the DSA transparency database has been established (4) with online platform providers reporting on so-called statements of reason which detail motives for the removal, restriction or rejection of content. There is risk, however, that this is a narrow view of transparency reporting. There is a sensitive category of digital services dealing with personal health information which are not well represented in the DSA transparency discourse. The DSA identifies the basis of auditing for service compliance, but how can digital health services create transparent information which serves auditors, citizens/patients as well as society?

Transparency, as mandated by the DSA, can be a great opportunity to ensure that digital health services operate in line with user privacy requirements, foster trust and ensure that user data are not monopolised by a handful of players; rather that they enable research and contribute to innovation.

In this article, I will focus on digital services that operate with personal health data. These services can vary in nature, from data storage systems to health recommender systems. They rely on large data collections to improve and finetune their digital services; some share or sell valuable information. Most of these activities may be performed without customers/users being able to weigh the costs and benefits to themselves.

Personal health data, artificial intelligence and data sharing

Health data is an important avenue to accelerate healthcare research. As a key element of current healthcare innovations, health data supports the development of important AI predictive models as well as decision support systems. Although the research benefits from data reuse are widely acknowledged, research and innovation on today's health data have not reached their highest potential, primarily because health data are kept in silos. Several reasons contribute to data being siloed, of which two are prominent: lack of data interoperability often due to a high incentive to maintain customised, proprietary solutions (5); and legal requirements or lack of clarity on the accountability and risks deriving from data sharing (6).

Despite being collected in silos, health data is generated in high volumes. This has driven the development of digital platforms which can be broadly divided into two categories: the electronic medical/health record (EHR/EMR) and the personal health record (PHR) (7). A common feature broadly distinguishing EHR/EMR from PHR systems is that EHR/EMR systems are intended for the use of healthcare professionals. On the other hand, the primary purpose of PHRs is to grant individuals a way to collect and control their personal health information (8).

Thus, PHRs can be defined as health and wellness data of individuals who exclusively control them across their lifespan (9). Via PHR systems, individuals can take a more active role in their own health, they can contribute to their health data repository and can rely on their decision-support capabilities (9). The information contained in PHRs can include medical history, medication and wearable/patient generated health information (i.e., weight, blood pressure, glucose, questionnaires, etc.) as well as wellness and lifestyle information (diet, exercise, mental health, reproductive, etc.). Due to their sensitive nature, PHRs are protected by data privacy regulations. The overall governing law applicable to PHRs in the European Union (EU) is the Data Protection Directive (10), specifying data protection rules about the processing of personal data and their free movement.

PHRs are increasingly valuable in the current demographic shift towards an aging population (11) as they can fuel the creation of novel digital health services that support patients to make health decisions and to self-manage their health (12). Having recognised the potential, big corporations are in an advantageous position to collect personal health records and to create competitive digital health products and services[1]. Powered by the network effect (13), the concentration of personal health data in the hands of the digital service providers is leading towards data monopolies (14) (also known as *data-opolies* (15)).

The network effect sees that markets which rely on large data collections experience positive feedback loops: more data implies a better product (14). In personal health data terms, this means that novel AI models and future health apps are more likely to be created and sold by the company that possesses the largest specialised health datasets. The use of AI leads to even more personal health data being concentrated into a handful of corporations.

[1] For example, while we do not know exactly how many PHR data points Apple currently stores and processes, Apple is currently the biggest wearable vendor, accounting for 29.7% of the market share in 2022, followed by Fitbit (Google).

Testimony to this trend is the fact that the latest ground-breaking AI models have Google (Bert Model – (16), Meta (Llama and Llama 2- (17)) and Microsoft[2] (OpenAI's GPT models – (18)) supporting them.

Digital health services may use AI models for important decision making. For example, they may use wearable information, state of wellness, voice data or PHRs to detect disease progression or decide on a treatment. On their own, they can bring a lot of benefits to customers; and novel regulation, such as the EU AI Act (19), are being proposed to protect end users. However, the DSA is an opportunity to improve digital services by requiring the disclosure of the use of AI algorithms and possibly their certification.

Moreover, via data monetisation (cumulatively the market value is estimated to reach US$707.86 billion by 2025[3]), and without patients being aware that their data is shared and used for commercial purposes, digital health services profit from health data. Increasing transparency of health data sharing practices has raised awareness about the transparency required within digital health services. However, if left on its own, trust in the medical system may be sacrificed to uncontrolled market forces.

Finally, while GDPR (20) is intended to protect data privacy, data ownership has yet to be clearly attributed (21,22). This topic merits its own debate as there are quite some repercussions that appear. Personal health records can be of a societal value, but this would require the companies controlling the data to let other researchers and innovators access such data (provided that users consent). This could happen only if personal health data become a shared resource in support of research innovations. Currently, each company has their own policies for researchers who want to analyse their data. Who they accept, or not, is usually not publicly known. Related to this, is the fact that individuals have been enabled by GDPR to data portability, yet data portability is hindered by lack of user awareness and format standardisation for the data exports (23). Therefore, even if an individual is able to download their personal data and carry them onto other platforms, the other platforms are often unable to read their data as it is not formatted following a standard. Thus on its own, data portability within GDPR has not been able to contrast the above-mentioned network effect.

Next, we will discuss cases that show how transparency in the DSA context can add value to individuals and to research and innovation. We will discuss the possibilities of public reporting resources to ensure that consumers and their fundamental rights are protected, to transparently account for digital health services and encourage growth and innovation.

Health apps as a case study

Driven by technology advancements and customer demand, the digital health service market has experienced significant growth and innovation[4]. The COVID-19 pandemic significantly accelerated the adoption of telehealth services, leading to increased demand for virtual consultations, remote monitoring and digital health platforms. This has also been reflected into an increased demand for remote monitoring technologies and wearable devices which can support managing chronic conditions and promote wellness. Moreover, large data collections have increased the opportunities for AI models to be integrated into digital health services to enhance diagnostic accuracy, personalise treatment plans, and improve patient outcomes. An easy way to increase customer use of digital health technologies has been via health apps. Apps provide an easy way to collect PHRs and to integrate wearable and AI technology.

To create some perspective on the PHR collection in the digital service space, we can examine data from some of the largest wellness/health apps in the market. Table 1 lists the owner, the user base (in 2023), annual revenue (in 2023), business model and data breaches (if any):

[2] GPT models are owned by Open-AI who has a large commercial agreement with Microsoft who has been sponsoring Open-AI since 2016 (see [17] for more details).
[3] Data Monetization Market https://www.transparencymarketresearch.com/data-monetization-market.html
[4] In 2023 the digital health market was estimated at US$ 241.3 billion and a further 30% growth is estimated by 2030 https://www.grandviewresearch.com/horizon/outlook/digital-health-market-size/global

Table 1 Overview of health apps with a large user base.

Health App	Owner	Service	Users	Business Model	Revenue	Data Breaches
My FitnessPal	Francisco Partners	Fitness tracking – collects activity, weight and dietary data	200 million	App purchases, subscriptions, advertising	US$310 million	In 2018 – 150 million user accounts
Fitbit	Google (Alphabet Inc)	Health and fitness tracking – collects demographic data, fitness data, stress and mindfulness data and manually-entered data, achievements[5]	128 million	Purchases of wearable devices and accessories, advertising, subscriptions	US$1 billion (estimated value as Google does not specify for Fitbit)	In 2021 - 61 million users, mostly Fitbit and Apple costumers were exposed. The leak generated from GetHealth, a platform that syncs data from popular fitness and health apps. In 2023, an advocacy group Noyb filed complaints against Fitbit (Austria, the Netherlands, and Italy) alleging that the company is in violation of EU data privacy regulations.
Apple Health App	Apple	An app for the organisation and sharing of health information – collects health records, medications, labs, activity and sleep.	Unclear (estimated *100 million users of Apple watch in 2020)	Purchases of Apple devices, integrations into corporate wellness programmes, research partnerships	Unclear (*US$38 million is the revenue shared for Apple watch)	In 2021 – 61 million users mostly Fitbit and Apple customers were exposed. The leak generated from GetHealth, a platform that syncs data from popular fitness and health apps.

The above information was collected using public sources. These services enjoy a large share of the health app market, yet there is no structural way to for a customer to be well informed on what happens to their PHRs. Events such as data breaches of PHRs can be of great impact for end users. Similarly, it is important to know how data is exploited within the companies storing customers' PHRs. Such information can be important decision-making points for customer choices. While most companies indicate that PHRs are not sold in exchange for money, each company has its own data monetisation strategies. For example, all three example companies share data with third parties. Two of which were exposed to data leaks via a third-party player.

Customers are affected yet they have no clear information over these data sharing processes. When individuals become users of a digital service, their consent regarding what the company does with the collected data does not reflect users' preferences on data sharing. Rather the user is forced to accept

[5] The data collection in Fitbit is much larger. The above is describing only data collected for the core service. More info is provided here https://support.google.com/product-documentation/answer/14811751?hl=en-IN

data sharing as part of the service agreement. The data sharing with third parties is mostly decided by the companies collecting the data. Researchers who want to perform research must apply to the companies in question who will decide if to enable the research. For example, Fitbit has an application programming interface (API) in place for researchers to access the data of their study cohort. In order to conduct a study, researchers buy the wearable devices from Fitbit and use them in their study protocol.

Each company has their own different application protocols. No public information could be found on how many research-related data requests are made to these services and how many are rejected. This does not mean that the companies do not engage in research and innovation, for example, since 2021, Fitbit has a health equity research initiative to promote health research[6]. Rather, the customers are not aware nor have the possibility to engage directly with such initiatives.

DSA transparency and digital health services

Transparency reporting in relation to personal health records should answer several questions for the customer:

1. If I use this service, what am I consenting to with regard to my data?
2. How are my data collected, processed and shared?
3. Is there any important decision taken by AI and how can it impact me?
4. If I agree to secondary use of data, how many and which type of researchers are allowed to work with the data?
5. Can I personalise data sharing with my own preferences on consent towards secondary research?

It is important that the answers to these questions are understood by customers and are monitored by authorities. When dealing with PHRs, various factors are of concern:

1. Even beyond personal health records, tech giants (Google, Apple, Amazon, Microsoft and others) are quickly penetrating the healthcare and pharmaceutical markets with a focus on inpatient and outpatient care, pharmaceutical R&D and intermediary fitness markets (24). Access to health and other personal records is essential to this step and is creating a large competitive advantage for these companies (25). To date, customers seem to understand and accept that they are paying for a service with their own data. Yet, there are societal ramifications to these decisions as well as greater consequences for customers when sensitive health data are at stake.

   For example, Apple is reportedly planning to provide a health insurance in the United States[7]. Having collected large datasets via their health app and in combination with artificial intelligence models, Apple is in a good position to estimate risk and pricing. There are regulation and reputation factors for which most insurers may not use data in this way (26). Independently from the hypothetical possibilities of what Apple may end up pursuing, one may argue that in general digital service markets have all the mechanisms needed to create behaviour-based services deriving from the appropriation of personal health data. This highlights an important aspect; users need to know how data is going to be used or monetised.

   Thus, clearly indicating the business model, especially how user data is monetised (27) by services under the DSA definition, is an important component of transparency and should be clearly reported to end users.

2. Large collections of PHRs can be vulnerable to data breaches and have serious consequences for customers (28). In a recent report commissioned by Apple, by October 2022, almost 12 billion online customer accounts and their data were compromised in data breaches, with cloud-based storage having seen a significant increase in the last two years (29). Data collection, data processing and sharing all expose user data to a degree of risk.

   Thus, clarifying how these three processes (data collection, data processing and sharing) are handled and the security mechanisms that are in place should be part of the transparency reporting. This should include how, where and which data are collected; how they are protected (i.e., encrypted in transit and/or in storage); for which purposes the data can be further

[6] https://blog.google/technology/health/meet-our-2023-health-equity-research-initiative-awardees/
[7] https://www.forbes.com/sites/barrycollins/2022/10/18/apple-will-launch-health-insurance--in-2o24-says-analyst/

analysed and which are the third parties who are also given access to them.

3. As AI is advancing at a rapid speed, it is starting to be used for important decision making (30). For example, Fitbit will use Gemini (a generative AI model created by Google AI) to coach users with personalised advice[8]. While such innovations are important and can positively impact citizens, the use of AI models in critical decision making and their limits should be made transparent (31). A simple way to achieve some level of transparency in relation to AI models are so-called model cards (32). These can be seen as simple documents accompanying AI models to describe how they work. Importantly, model cards also enable reporting that explains how AI decisions are taken (from the field of explainable AI – see (33)), making them an interesting instrument to understand complex AI models. They are already used by the AI industry and, with small extensions may be used to clarify how a digital health service is using a specific AI model, on which data and how well this model performs (34).

4. Successful digital health services benefit from the network effect, enabling the service provider to offer improved services while their competitors remain behind. As data becomes concentrated into one main player, more and more data are controlled by the company. At the least the situation becomes asymmetric and, as indicated in the use case section, a monopoly on information can occur. A key to breaking these data asymmetries is to enforce interoperability (35). Interoperability describes the ability of a system to exchange or process information from another system. Enabling access to patient data must be a paramount guiding principle to approach the problem of data asymmetry. When interoperable, health data must be accessible to physicians and patients and in a secured and format that will benefit research (36).

   Risks to privacy are present with and without data sharing (35). These risks can be eliminated or mitigated by new secure data analytics techniques such as confidential computing (37) or homomorphic encryption (38). Moreover, additional blockchain platforms have been proposed as a mechanism to enable data sharing with transparency regarding data access and control over what happens to users' data (39,40). Interoperable protocols or systems should be adopted within the DSA transparency report to enable secondary health data sharing and to foster innovation.

5. Faced with services whose data harvesting remains opaque; health app users make important decisions without adequate understanding. If properly informed, users have shown to select apps with fewer data collection points. Moreover, for users, the purpose of use has a high importance, especially with personal health records (41). There has been work in capturing consent and purpose of use in blockchain-based data sharing networks (39,42), the idea being to capture a decentralised data management process with a blockchain network. Users, or data owners, can decide how and when to enable others to use their data. The possibilities are numerous; should data ownership of PHRs be clarified, this can open novel opportunities for digital health services of the future.

Dynamic consent (43) and management of PHRs should enable customers to decide on secondary data reuse. Accompanied with standard approaches to data portability, this would foster more data distribution, more competition and innovation. Several studies propose including users in dynamically providing consent using blockchain-based networks (42,44,45). Blockchain networks are distributed and transparent in defining what happens between the participants who transact in the network. Such solutions can enable users to decide and choose what to share, with whom and for what purpose. Yet, while promising, this technology needs further developing to be used at scale.

In conclusion, the legitimate advantage that large tech players have acquired in digital services is providing them with access to the sphere of health and medicine (46). This access raises numerous risks that are not captured by privacy regulations, such as GDPR, but can be partially mitigated via standardised transparency reporting within the DSA framework. If these risks are not mitigated the digital health services offered by tech corporations will soon be delivering our health services, shaping

[8] https://www.pcmag.com/news/fitbit-using-google-gemini-for-new-ai-that-could-become-your-fitness-coach

health policy and gaining decision-making power across the healthcare domain. Transparency reporting should be mandatory for products and services dealing with personal data. Transparency reporting should be regulated in ways that does not hinder but contributes to innovation, thus further analysis should be made on the proposed requirements to identify how they can positively contribute to health innovation products and services within the EU.

Table 2 Transparency challenges and possible reporting strategies

Transparency challenges	Reporting	Existing technical solutions	Policy
Profit models	Which profit models are used by the service provider?	Data monetisation models – At an advanced level, blockchain technology and smart contracts can transparently log PHR data sharing across service providers and enforce or support monetisation models.	Requirement to report data monetisation models used by the service provider.
Data collection	Where is the data stored?	Cloud-based platforms, servers, personal devices, data vaults, etc	Requirement to describe data collection at a level of detail that helps decision making of users.
	How is it protected?	Encryption/ decryption models can protect data in transit and in storage.	
	Data breaches	Public data logs – blockchain technology can increase trust in identifying data breaches if by design all the data transfers are logged in a shared blockchain reporting system.	
Data processing	Data processing that may affect the user	(AI) models deciding for aiding decisions for/about the user	Reporting on data processing using tools such as model cards to help end users to decide how data processing affects them or the society.
	Data processing that may group/ categorise the user	(AI) models classifying the user into a group or a category	
	Data processing that goes beyond the user	(AI) models making population-level decisions	
Data sharing	With whom are the data shared?	Public data access logs identifying data processors – blockchain-based systems can record data sharing and record data transactions across service providers.	Explicitly name additional data processors (business partners, researchers) and the purposes of secondary use
	For what purpose?	Interoperability protocols: secondary uses can be based on consent models and PHR data standardisation efforts. Researchers may be given API access to data.	

Transparency challenges	Reporting	Existing technical solutions	Policy
Ownership	Who is the owner of the data?	User or service identifiers	Clearly identify the data owner and what the user (owner or not) can do with their personal health records, including their rights.
	What can the owner can do to the data?	A variety of techniques: if the owner is the service provider, then profit, collection, processing, sharing is as described in the cells above.	
	What can the user consent to?	Several data consent models exist, including dynamic consent whose basis is to enable users to agree or not to share data for different processing purposes.	
	What are the rights of users?	Thanks to GDPR, for EU users, downloading, removing and updating PHRs are technical functions that should always be provided. Standard formatting for the downloaded PHRs is still missing.	

Policy recommendations

The DSA should provide a standardised transparency reporting format for personal records, including those that are health-related. Such a format will store transparency information from the digital service/platform. The main components of services transparently reporting on personal health records are included in Table 2. The table proposes five core aspects to be clarified by standardised transparency reporting as well as identifies well-known and emerging technologies that can support reporting in a decentralised way.

In particular there are five main points to report on:

- **Profit models:** which data monetisation is used
- **Data collection**: highlighting the security levels for the stored data and eventual data breaches
- **Data processing**: the use of data processing techniques and resulting (AI) models; the details of AI models can be reported via model cards or a similar format
- **Data sharing**: clarifies for the users and for the potential research community if/how/with whom/for what the data are shared
- **Data ownership**: identifies who owns the data and what users can consent to when data is being shared with third parties. Users should know their data portability options and be provided with a portable format.

The DSA should disincentivise the creation of health data monopolies. There should be a clear understanding of the increasing and ubiquitous influence of tech giants in digital health platforms. This understanding can be achieved with direct requests for information.

The DSA should ask digital service platforms to report on data sharing collaborations between researchers and its users (if any). Informed users, standardised formats for data portability and standardised transparency reporting should help to combat data monopolies and support innovation with novel products generated from new market players.

As an extension of the DSA transparency database, the DSA should consider enabling data sharing through the enforcement of interoperability platforms. The use of shared (blockchain or similar decentralisation technology) networks can aid information sharing and generate trust among users. Such transparency generated by the DSA will help users to decide whether to trust companies with their own personal data.

Author Information

Visara Urovi is an Associate Professor at the Institute of Data Science within Maastricht University's Faculty of Science and Engineering. Her research encompasses predictive AI models and data sharing techniques applied to healthcare.

References

1. Digital Services Act [Internet]. 2024. Available from: https://eur-lex.europa.eu/EN/legal-content/summary/digital-services-act.html
2. Urman A, Makhortykh M. *How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms*. Telecommun Policy. 2023;47(102477,).
3. Cotterrell R. *Transparency, mass media, ideology and community*. J Cult Res. 1999;3(4):414–26.
4. DSA Transparency Database [Internet]. 2024. Available from: https://transparency.dsa.ec.europa.eu/
5. Szarfman A, Levine JG, Tonning JM, Weichold F, Bloom JC, Soreth JM, et al. *Recommendations for achieving interoperable and shareable medical data in the USA*. Commun Med. 2022 Jul 18;2(1):86.
6. Geneviève L, Martani A, Perneger T, Wangmo T, Elger B. *Systemic fairness for sharing health data: perspectives from Swiss stakeholders*. Front Public Health. 2021 May;(9:669463).
7. Brandt R, Rice R. *Building a better PHR paradigm: Lessons from the discontinuation of Google HealthTM*. Health Policy Technol. 2014 Sep;3(3):200–7.
8. Flaumenhaft Y, Ben-Assuli O. *Personal health records, global policy and regulation review*. Health Policy. 2018 Aug;122(8):815–26.
9. Tang, PC, Ash, JS, Bates DW, Overhage JM, Sands DZ. *Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption*. J Am Med Inform Assoc. 2006;13(2):121–6.
10. General Data Protection Regulation [Internet]. [cited 2024 Dec 1]. Available from: https://gdpr-info.eu
11. *World report on ageing and health*. World Health Organization; 2015.
12. Karampela M, Ouhbi, S, Isomursu, M. Personal health data: *A systematic mapping study*. Int J Med Inf. 118:86–98.
13. Soma JT, Davis KB. *Network effects in technology markets: Applying the lessons of Intel and Microsoft to future clashes between antitrust and intellectual property*. J Intell Prop L. 2000;8, 1.
14. McIntosh D. *We need to talk about data: how digital monopolies arise and why they have power and influence*. J Technol Law Policy. 2018;
15. Stucke ME, Grunes AP. *Introduction: big data and competition policy*. Big Data Compet Policy. 2016;
16. Devlin J, Chang MW, Lee K, Toutanova K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018.
17. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D. Llama 2: *Open foundation and fine-tuned chat models*. ArXiv Prepr ArXiv230709288.
18. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. *GPT-4 Technical Report* [Internet]. arXiv; 2024 [cited 2024 Dec 11]. Available from: http://arxiv.org/abs/2303.08774
19. EU AI Act [Internet]. European Union; 2024. Available from: https://eur-lex.europa.eu/eli/reg/2024/1689/oj
20. Voigt P, Von dem Bussche A. *The EU general data protection regulation (gdpr). A Practical Guide*,. Springer Int Publ. 2017 Aug;
21. Black ST. *Who Owns Your Data?*. Ind Rev. 2021;54:305.
22. Kahn SD, Terry SF. *Who owns (or controls) health data?* Sci Data. :February 2024 Feb.
23. Kranz J, Kuebler-Wachendorff S, Syrmoudis E, Grossklags J, Mager S, Luzsa R, et al. *Data portability*. Bus Inf Syst Eng. 2023 Oct;65(5):597–607.
24. Schuhmacher A, Haefner N, Honsberg K, Goldhahn J, Gassmann O. *The dominant logic of Big Tech in healthcare and pharma*. Drug Discov Today. 2023 Feb 1;28(2):103457.
25. Gleiss A, Kohlhagen M, Pousttchi K. *An apple a day–how the platform economy impacts value creation in the healthcare market*. Electron Mark. 2021 Dec;31(4):49–76.

26. Jeanningros H, McFal L. *The value of sharing: Branding and behaviour in a life and health insurance company*. Big Data Soc. 2020 Sep;7(2).
27. Birch K, Cochrane DT, Ward C. *Data as asset? The measurement, governance, and valuation of digital personal data by Big Tech*. Big Data Soc. 2021 May;8(1).
28. Kröger JL, Miceli M, Müller, F. *How data can be used against people: A classification of personal data misuses*. 2021;
29. Stuart E, Madnick S. *The rising threat to consumer data in the cloud* [Internet]. 2022 [cited 2024 Apr 1]. Available from: https://www.apple.com/newsroom/pdfs/The-Rising-Threat-to-Consumer-Data-in-the-Cloud.pdf
30. Rajpurkar P, Chen E, Banerjee O, Topol EJ. *AI in health and medicine*. Nat Med. 2022 Jan;28(1).
31. Fletcher R, Nakeshimana A, Olubeko O. *Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health*. Front Artif Intell. 2021 Apr;
32. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. *Model cards for model reporting*. Proc Conf Fairness Account Transpar. 2019 Jan 29;220–9.
33. Holzinger A, Saranti A, Molnar C, Biecek P, Samek W. *Explainable AI methods-a brief overview. InInternational workshop on extending explainable AI beyond deep models and classifiers*. Springer Cham. 2022;13–38.
34. Lohachab, A, Urovi V. A *Blockchain-Based Approach for Model Card Accountability and Regulatory Compliance*. In: Cham: Springer International Publishing. 2024.
35. Pernagallo G. *Overcoming asymmetric information: A data-driven approach*. Elgar Companion Inf Econ. 2024 Mar 12;135–53.
36. Agrawal R, Prabakaran S. *Big data in digital healthcare: lessons learnt and recommendations for general practice*. Heredity. 2020 Apr;525–34.
37. Mulligan DP, Petri G, Spinale N, Stockwell G, Vincent HJ. *Confidential Computing—a brave new world*. Int Symp Secure Priv Exec Environ Des. 2021 Sep 20;132–8.
38. Yi X, Paulet R, Bertino E. Homomorphic encryption. Springer Int Publ. 2014;
39. Urovi V, Jaiman V, Angener A, Dumontier M. Luce: *A blockchain-based data sharing platform for monitoring data license accountability and compliance*. Blockchain Res Appl. 2022 Dec 1;3(4):100102.
40. Coelho R, Braga R, CDavid J, Dantas M, Stroele V, Campos F. *Integrating blockchain for data sharing and collaboration support in scientific ecosystem platform*. 2021;
41. Van Kleek M, Liccardi I, Binns R, Zhao J, Weitzner DJ, Shadbolt N. *Better the devil you know: Exposing the data sharing practices of smartphone apps*. In: Proceedings of the 2017 CHI. 22017 May 2. p. 5208–20.
42. Jaiman V, Urovi V. *A consent model for blockchain-based health data sharing platforms*. IEEE Access. 2020 Aug;
43. Kaye J, Whitley EA, Lund D, Morrison B, Teare H, Melham K. *Dynamic consent: a patient interface for twenty-first century research networks*. Eur J Hum Genet. 2015 Feb;23(2):141-6.
44. Khalid M, Ahmed M, Helfert M, Kim J. *Privacy-First Paradigm for Dynamic Consent Management Systems: Empowering Data Subjects through Decentralized Data Controllers and Privacy-Preserving Techniques*. Electronics. 2023 Dec 12;(12(24):4973.).
45. Merlec MM, Lee YK, Hong SP, In HP. *A smart contract-based dynamic consent management system for personal data usage under GDPR*. Sensors. 2021 Nov 30;
46. Sharon T. *Blind-sided by privacy? Digital contact tracing, the Apple/Google API and big tech's newfound role as global health policy makers*. Ethics Inf Technol. 2021 Nov;23(Suppl 1):45–57.

Maastricht University

STUDIO EUROPA MAASTRICHT

Policy Brief

Synthetic media and reality engineering: policy solutions for the EU

Dr Thomas Frissen

Summary

In this policy brief, I explore synthetic media: digital artefacts created entirely with generative artificial intelligence (GenAI). These can be visual, auditory, audiovisual or textual, such as deepfake videos or output from large language models (LLMs), such as ChatGPT. Due to the rapid developments in GenAI technology, it is becoming increasingly easy for anyone to engineer a 'reality' with synthetic media. Unlike traditional forms of forgery, synthetic media require no original source and are algorithmically crafted. This makes them powerful tools for both creativity and deception.

Synthetic media are everywhere, from viral social media hoaxes to malicious deepfake campaigns. In 2024, fake images of celebrities at the Met Gala fooled millions, while realistically-sounding deepfake robocalls tried to disrupt primary elections in the United States. Such misuses fuel a growing epistemic crisis in societies, eroding trust in democratic processes by blurring the line between fact and fiction. The real threat lies not just in the convincing nature of synthetic media, but in their rapid spread across digital platforms, particularly very large online platforms (VLOPs). Understanding these dynamics is essential to mitigating harm at both individual and societal levels.

This brief offers a number of policy recommendations to address these challenges under the EU Digital Services Act (DSA) and AI Act. Key proposals are:

Fortifying investments in digital forensics for early detection of harmful deceptive media.
Holding social media platforms accountable for their role in enabling and amplifying synthetic media.
Building public resilience through psychological inoculation strategies.

These policies aim to address synthetic media, their enablers and their societal impact in a responsible manner—without throwing the baby out with the bath water. As such, the policy recommendations support the safeguarding of creativity and democratic integrity while fortifying trust and safety in the digital age.

Why Katy Perry's ultra-extravagant dress matters

On 6 May, the 2024 edition of the Met Gala took place. Fashion's biggest night out of the year is most known for its red carpet with stars in excentric and extravagant outfits. But this year, right after the opening, pictures of celebrities in truly unprecedented and extraordinary outfits started to appear on social media. Images of Katy Perry in a princess dress that looked like a wreath of moss with a careful but seemingly organic arrangement of flowers, leaves and butterflies and Rihanna in a white puffy satellite dish dress with embroidered green floral ornaments circulated on X and Instagram and received millions of views and comments in only a few minutes (1,2). Thousands of social media users expressed their excitement as well as disgust about the stars and their outfits. Remarkably, however, both of these superstars were not present at the Gala that evening. These photos turned out to be deepfakes (i.e., synthetic images): artefacts forged with the help of generative artificial intelligence technology and deep learning. These images seemed so true that millions of social media users couldn't tell whether they were real or fake.

To some this story may sound mundane or trivial, and admittedly, the Met Gala is very exclusive and some outfits are so extravagant that they could blur the line between fantasy and reality. Why should we be bothered by this story? Because in a different context, synthetic media may be deadly serious. Voice synthesis and audio deepfakes, for example, have been widely used in scamming activities (3,4) as well as for spreading political disinformation. In January 2024, thousands of Americans received a realistic sounding deepfake robocall from Joe Biden with the request to not vote in the presidential primary (5). It was unclear who was behind this action, but is was evident that their objective was to suppress votes and interfere in elections. Clearly, it seems safe to claim that these kinds of synthetic media may have harmful consequences for democratic processes and democracies in general.

An epistemic crisis and synthetic media: a combustible combination

According to Benkler, Faris and Roberts (6), many democratic societies are going through an epistemic crisis in which they are 'buckling under the pressure of technological processes that [overwhelm] our collective capacity to tell truth from falsehood and reason from its absence' (6, p4). With the rapid developments in GenAI technology, it is becoming only more difficult to differentiate between real and fake (7). Currently however, many synthetic media, like deepfake videos, still have many imperfections and are relatively easy to spot. Yet, these technologies are becoming increasingly accessible, easy-to-use and advanced; and the results are becoming better and more realistic.

The potentially combustible combination of an epistemic crisis and technological advances in synthetic media have led to the argument that the EU and other democratic societies may be heading towards an 'infopocalypse' (8) if we do not rethink the possibilities of regulating the production and dissemination of synthetic media. In this policy brief, a short overview of the technologies underpinning synthetic media is provided. Then, the state of the research on the uses and implications of synthetic media is revisited before several policy recommendations are made in relation to the EU Digital Services Act (DSA) and the EU AI Act.

From pixels to perfection

The concept of synthetic media refers to a specific set of media artefacts that are created (or manipulated) with GenAI technology. Synthetic media can be visual, auditory, audiovisual and/or textual. Perhaps the best-known forms of synthetic media are deepfake videos and images—like the images discussed in the opening anecdote of this policy brief—but texts generated through large language models (LLMs), such as OpenAI's ChatGPT, are good examples of synthetic media as well. The origins of these deepfakes can be traced back to 2017, when videos began circulating on Reddit that used machine learning technology to swap the faces of Hollywood actresses onto the bodies in pornographic video footage (8–10). These media artefacts are called synthetic because they are the end-product of a compiling process that uses algorithmic, mathematical and stochastic procedures in which fragmented digital substances are merged together into a new digital whole—a whole that bears close resemblance to reality.

A good example of a synthetic medium is a deepfake video created with one of the most common AI techniques for media synthesis, a generative adversarial network (GAN). In a GAN, two contesting, computational agents, one called the 'generator' and the other 'discriminator', work against each other in an iterative loop. The discriminator is supplied with a set of target images, and the generator compiles a first image by rearranging differently coloured pixels at random on a blank canvas. This image is sent to the discriminator who tries to differentiate between the generated artefact and the target image by comparing the two compositions at pixel level. If the discriminator succeeds, the generator has failed and is tasked with compiling a new image with a new random order of pixels based on the feedback from the discriminator. This loop is repeated over and over and with each iteration the generator 'learns' which compositions are more adequate. This loop continues until the generator has produced an image or video (i.e., compositions of coloured pixels) that the discriminator cannot differentiate between the synthesised artefact and the target image.

More recently, synthetic media are increasingly created with a new generation of deep-learning techniques, specifically diffusion models. A diffusion model works a bit differently than a GAN. Where in a GAN two adversarial agents play a zero-sum game (i.e., when one agent gains, the other fails), a diffusion model 'learns' from the process of gradually adding noise to an image. If the model 'understands' that process, it can also learn to reverse it. A diffusion model then starts with random noise, and through a step-by-step approach it transforms the noise into realistically looking footage. This technology works for static images (e.g., with Stable Diffusion[1] or Midjourney[2]) as well as for videos (e.g., OpenAI's Sora[3]).

Looking at these techniques, it is important to note that synthetic media are not necessarily manipulations of existing media. They differ fundamentally from more classical image manipulation techniques, such as so-called photoshopping, in terms of mode and speed of creation, the required human skills and know-how, the non-human and algorithmic agency in the creation process and the experienced reality of the artefacts. Indeed, the computational technology underpinning synthetic media demands much less human skill than was needed previously for image forgery or manipulation.

Furthermore, in contrast to older forms of image manipulation, synthetic media do not require a so-called original (11). Especially in the case of diffusion models; if they are provided with any form of text-based input, the models will create credible-looking footage of anyone or anything in any imaginable simulated situation (12). In this way, synthetic media also differ from previous forms of forgery because of the algorithmic autonomy. A substantial part of the creation of the synthetic medium is outsourced to a set of autonomous stochastic, mathematical and/or statistical processes beyond the creativity of a human. This leads to infinite possibilities in the engineering of seemingly credible realities.

Reality engineering

Currently, the production of synthetic media still requires a lot of computational resources, but the tide is turning. With the advent of LLMs like ChatGPT, and diffusion models like Sora or Stable Diffusion, the availability, accessibility and accuracy of synthetic media is only improving. It is not unlikely that this will also increase the different uses of synthetic media in both benign and malignant contexts.

Benign engineered realities

Synthetic media, generated with either GANs or diffusion models, are extensively used in art, film and documentary production, mental healthcare and medicine. Although visual effects are not uncommon in Hollywood, the desired result can be achieved much faster, cheaper and more realistically through media synthesis. Very recently, for example, film director Robert Zemeckis used synthetic media technologies to age and de-age the character played by Tom Hanks in the 2024 film Here (9). Furthermore, in a documentary about political protestors in Hong Kong, synthesised faces were used for interviewees to ensure their safety and anonymity (13).

In medicine, synthetic media are adopted in the processes of interpreting complex MRI results as well as in the training of surgeons and ophthalmologists (14). Similarly, deepfakes have been effectively utilised in therapeutic settings, aiding individuals with post-traumatic stress disorder (PTSD) and assisting those coping with the sudden loss of a loved one (15,16). Synthetic media are also believed to support creativity and the realisation of new modus operandi in the arts, communication and marketing (17,18).

Malignant engineered realities

While there are these benign examples, the main point of concern in the scientific and public debate are the malicious ones. This concern is mainly rooted in the idea that the core principle of synthetic media is deception or the intention to knowingly mislead another person (19,20). In fact, synthetic media such as deepfakes have indeed been used with this objective in the contexts of politics and monetary scams. The Joe Biden robocall example from the introduction of this policy brief is such a case (5). There have also been reported instances of individuals falling victim to scams perpetrated through

[1] https://stability.ai/
[2] https://www.midjourney.com/home
[3] https://openai.com/index/sora/

voice synthesis in which perpetrators extort money with deepfake voices of distressed family members. In a similar vein, immediately after the 2023 Turkish earthquakes, a famous AI-generated image of a Greek firefighter rescuing a young child was posted on X alongside links to crypto wallets with the request to donate money (21).

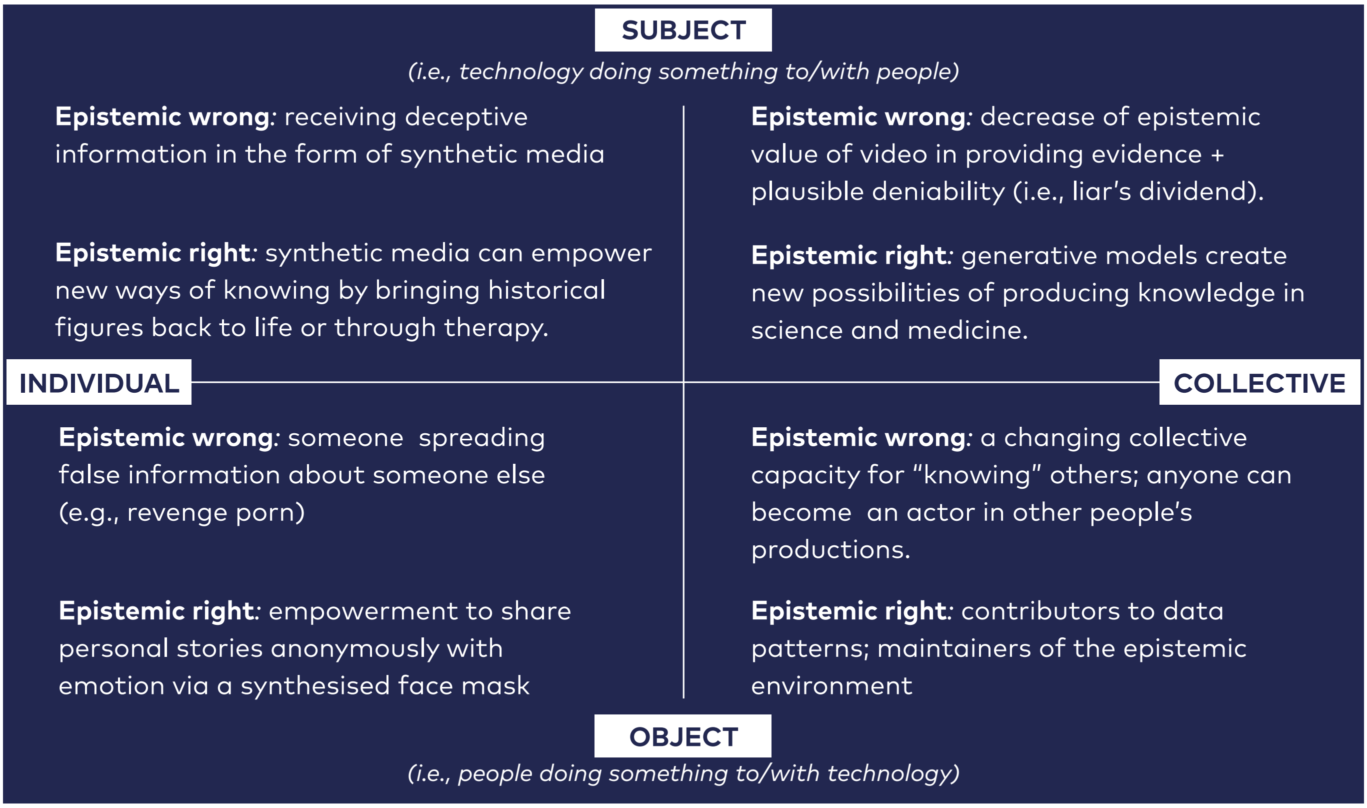
In political arenas, synthetic media are increasingly more common and being used both in support of and against politicians. Notably, Pakistan's former prime minister, Imran Khan, who has been sentenced to several years in jail and is prohibited from engaging in political activities, employed GenAI voice synthesis to create support for his political party (22). Similarly, in a remarkable attempt to demonstrate Russia's synthetic media capabilities, Vladimir Putin was interviewed by a synthetic body double of himself during a news conference last year (23). Likewise, immediately after the Russian invasion of Ukraine, a deepfake video of Volodymyr Zelensky surfaced on social media in which he demands Ukrainians lay down their arms and surrender (24). More recently, Tom Cruise's 'voice' was used to narrate a documentary called Olympics Has Fallen, which was part of a Russian-led disinformation campaign to discredit Emmanual Macron and the Paris Olympics (25).

Crucial to note here is that synthetic media can be created by isolated individuals fairly easily. However, their dissemination depends heavily on digital networks and the online information ecosystem, and on VLOPs in particular. More specifically, we could say that VLOPs are the wind in the sails of synthetic media. They propel them far into the networks of different audiences. Without the propulsion generated by likes, shares and comments, synthetic media would remain adrift and their impact would remain relatively insignificant. Rather than thinking of synthetic media as influencing us directly like the old saying 'seeing is believing', we should consider their impact to be more like a ripple effect: something that grows as the media artefacts spread through networks and online communities. In fact, it is less about whether the synthetic media look convincingly real and much more about how we interact with them. In that sense, the true danger of synthetic media lies in the combination of seemingly realistic outputs together with the collective engagement in digital networks.

The four quadrants of epistemic consequences of synthetic media

While the science on the effects of synthetic media is still in its infancy, some relevant initial findings can shed a light on the implications of synthetic media in

Figure 1 Visualisation of the proposition by Kerner and Risse (10)



SUBJECT

(i.e., technology doing something to/with people)

Epistemic wrong: receiving deceptive information in the form of synthetic media

Epistemic right: synthetic media can empower new ways of knowing by bringing historical figures back to life or through therapy.

Epistemic wrong: decrease of epistemic value of video in providing evidence + plausible deniability (i.e., liar's dividend).

Epistemic right: generative models create new possibilities of producing knowledge in science and medicine.

INDIVIDUAL

COLLECTIVE

Epistemic wrong: someone spreading false information about someone else (e.g., revenge porn)

Epistemic right: empowerment to share personal stories anonymously with emotion via a synthesised face mask

Epistemic wrong: a changing collective capacity for "knowing" others; anyone can become an actor in other people's productions.

Epistemic right: contributors to data patterns; maintainers of the epistemic environment

OBJECT

(i.e., people doing something to/with technology)

current societies. Firstly, most synthetic media still have imperfections which makes it relatively easy to detect forged footage. In line with this, Vaccari and Chadwick (26), found in their experimental study that most people were in fact not deceived by a deepfake video of Barack Obama, especially if they were exposed to a disclaimer stating that the video was fake.

Having said that, synthetic media can have ramifications on epistemic processes in society. In other words, they can affect how people and communities come to know things or believe in what is true or false. In a resourceful article, Kerner and Risse (10) map the epistemic 'rights' and 'wrongs' in the context of synthetic media on two axes: individual vs collective and object vs subject (see Figure 1). With the distinction between subject and object, the authors refer to the difference between the technology doing something to people or communities (i.e., subject) and the people/communities doing something to the technology (i.e., object).

These four quadrants will now be explained, starting with the part in which people are subjected to synthetic media. As individual epistemic subjects, we have an epistemic right to information - to correct information. This right may be violated when we receive synthetic media that provide deceptive or misleading information. At the same time, synthetic media can also empower epistemic rights of individual subjects. For example, historic figures can be brought to life or people can talk to beloved ones that have passed away (10).

In the arena of collective subjects, a significant epistemic wrong is the devaluation of video footage as undisputable testimony or evidence. Up until now, video has played a crucial role in human inquiry and court cases. This was because, historically, videos could not be easily tampered with. However, with the possibility that video footage can be synthetic and fake, our collective trust in the epistemic value of video decreases. Furthermore, the existence of synthetic media now means that anyone could claim that any footage could be a deepfake. This plausible deniability is what is often called the liar's dividend (27). At the same time, there are also positive aspects of synthetic media for the collective epistemic subject. As discussed previously, media synthesis can facilitate new ways of generating knowledge in the context of science and medicine (14).

The lower half of the figure addresses when people become objects instrumental to technology. More specifically, as individual epistemic objects, people can experience epistemic harm in their self-determination regarding how they wish to be perceived by others. Synthetic media can misrepresent individuals, depicting them in contexts or activities they did not consent to. For example, deepfake pornographic videos have impeded Indian journalist Rana Ayyub's ability to be recognised appropriately in her professional capacity (10). Yet, synthetic media can also empower the individual epistemic object. Through filters and face masks on social media platforms, people can share personal or intimate experiences (e.g., about abuse) with real emotional expressions, without revealing their true identities.

Finally, in the context of collective epistemic objects, synthetic media impact the ways in which we get to 'know' each other. People's fantasies can be externalised into synthetic media, which means that anyone can become an actor in other people's fabrications (10). At the same time, the collective epistemic object contributes to data patterns in the digital realm and therefore fortifies our collective understanding of society. More specifically, through the creation of synthetic media we collectively supply vast amounts of data about our private lives which reveal patterns of human behaviour, thoughts and feelings like never before.

This approach is meaningful as it focuses on the epistemic ramifications of synthetic media rather than on the technologies underpinning them and is therefore a robust framework for the development of a long-term policy. If we wish to regulate synthetic media in Europe in an adequate manner, we should work on a policy that caters to each of these quadrants. In the next section, some suggestions will be made.

Policy recommendations

Based on the examples and discussions above, we can derive a few principles that are meaningful for current and future EU regulations for synthetic media. First of all, the genie is out of the bottle: the technology and synthetic media are here to stay. In fact, the GenAI industry is currently expanding, and future synthetic media will only become more sophisticated and realistic. Second, synthetic media, their underlying technology (e.g., GANs and diffusion models) and their corresponding engineered realities are not intrinsically benign or malignant. Both on the individual as well as the collective level, synthetic media may have harmful as well as empowering capacities—even when it comes to deception. EU regulations should therefore be careful in their attempt to mainly prohibit certain AI systems (the DSA and Chapter II, art. 5 of the AI Act) simply because of certain technological capacities. Moreover, prohibition requires enforcement, and enforcement may be difficult. That being said, the following steps could be taken to monitor, restrict and embrace synthetic media in the EU.

One way to tackle this problem could be technologically: the EU could invest in better digital forensics to detect false, misleading and deceptive media artefacts faster. This would immediately address the epistemic wrongs in the collective-subject quadrant of the framework above (10). It would help to fortify the truth and our collective trust in video footage again (28). This is also in line with the society-wide goal of 'mitigation of systemic risks, such as manipulation or disinformation' as stated in the DSA.

The technology to create synthetic media often produces results with imperfections. These imperfections can be detected with adequate software. For example, earlier deepfakes could often be detected by looking at the physiological discrepancies between the deepfake and the real person (e.g., a person's eye blinking patterns), but this method has also drawbacks. As soon as detection technology improves, the software to produce media synthesis also improves. Newer deepfakes now mirror, for example, authentic blinking behaviour. The risk is that it will become a cat-and-mouse game in which the cat never catches the mouse. In line with the DSA and the EU AI Act, one possibility could be to require providers of certain (high-risk) GenAI systems to build in methods to fingerprint their products. This would enable quick verification of the authenticity of a certain image or video clip.

Additionally, an EU policy could also consider addressing the enablers and distributors of synthetic media, especially VLOPs. This is immediately in line with the DSA goal of 'greater democratic control and oversight over systemic platforms'. In this context, VLOPs in particular play an important role. Although the psychological effects of exposure to a certain synthetic medium may be limited (26), including a disclaimer that indicates whether content is fake or not can still be beneficial when a video clip or image is posted. Many social media platforms are already including such warnings.

Finally, and perhaps most importantly, given the epistemic crisis many democratic societies are going through, the EU should invest in increasing trust in political and democratic institutions as well as fortifying media and digital literacy. Benkler and colleagues (6), argue that societal polarisation precedes the emergence of the internet—and by extension the emergence of synthetic media—and not the other way around. Therefore, it may be most fruitful to address the societal polarisation or the quadrant of individual epistemic subjects first and foremost.

While the DSA has as an ambition to better protect citizens from illegal online content, it is less clear on how this should be done. In this context, it may be interesting to look at the science behind psychological inoculation (29,30). This technique has proven particularly meaningful for fighting disinformation and misinformation. In a process similar as to vaccination, psychological inoculation works through exposing people to a small dose of the dis- and/or misinformation in order to make them more resilient and immune to it. The EU could support the development of a similar technique for the technologies underpinning synthetic media. More specifically, the expectation is that by training people in creating their own synthetic media, they would become more proficient in potentially detecting other synthetic media, including malignant ones. Next year's Met Gala would be a good test case for this. Would the creation of one's own deepfake extravagant outfit build sufficient 'antibodies' to differentiate a real truth from an engineered truth? Only time will tell.

Author Information

Thomas Frissen, an Assistant Professor at the Faculty of Arts and Social Sciences at Maastricht University, investigates the relationship between digital technologies and human psychology in his research. Thomas is is the director of studies for the Bachelor Digital Society.

References

1. Mendez II M. *Don't Be Fooled by These AI-Generated Met Gala Looks*. Time. 2024; Entertainment-Internet Culture.
2. Crescenzi C. *Met Gala Deepfakes Are Flooding Social Media*. Wired [Internet]. 2024 May; Available from: https://www.wired.com/story/met-gala-deepfakes-katy-perry-rihanna/
3. Chen H, Magramo K. *Finance worker pays out $25 million after video call with deepfake 'chief financial officer.'* CNN. 2024 Feb 4;
4. Khatsenkova S. *Audio deepfake scams: Criminals are using AI to sound like family and people are falling for it*. Euronews. 2023 Mar 23;
5. Steck E, Kaczynski A. *Fake Joe Biden robocall urges New Hampshire voters not to vote in Tuesday's Democratic primary*. CNN. 2024 Jan 22;
6. Benkler Y, Faris R, Roberts H. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. New York, NY: Oxford University Press; 2018.
7. Twomey J, Ching D, Aylett MP, Quayle M, Linehan C, Murphy G. *Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine*. PLoS One [Internet]. 2023;18(10 October):1–22. Available from: http://dx.doi.org/10.1371/journal.pone.0291668
8. Fallis D. *The Epistemic Threat of Deepfakes*. Philos Technol. 2020;
9. Lees D. *Deepfakes in documentary film production: images of deception in the representation of the real*. Stud Doc Film [Internet]. 2024;18(2):108–29. Available from: https://doi.org/10.1080/17503280.2023.2284680
10. Kerner C, Risse M. *Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds*. Moral Philos Polit. 2021;8(1):81–108.
11. Fasoro A. *The ontological quandary of deepfakes*. AI Soc [Internet]. 2024;(0123456789). Available from: https://doi.org/10.1007/s00146-024-01902-6
12. Meikle G. *Deepfakes*. Polity Press; 2023.
13. Sparks M. *BBC documentary used face-swapping AI to hide protesters' identities*. New Scientist [Internet]. 2022 Nov; Available from: https://www.newscientist.com/article/2348197-bbc-documentary-used-face-swapping-ai-to-hide-protesters-identities/
14. Waisberg E, Ong J, Masalkhi M, Lee AG. *OpenAI's Sora in ophthalmology: revolutionary generative AI in eye health*. Eye. 2024;(April):8–9.
15. Haselhoff R. *Deepfakes & Mental Health: Using Artificial Intelligence for Good*. Beeld and Geluid. 2021.
16. van Minnen A, ter Heide FJJ, Koolstra T, de Jongh A, Karaoglu S, Gevers T. *Initial development of perpetrator confrontation using deepfake technology in victims with sexual violence-related PTSD and moral injury*. Front Psychiatry. 2022;13.
17. Miller E, Dupont T, Wang M. *Enhanced Creativity and Ideation through Stable Video Synthesis*. 2024; Available from: http://arxiv.org/abs/2405.13357
18. Whittaker L, Letheren K, Mulcahy R. *The Rise of Deepfakes: A Conceptual Framework and Research Agenda for Marketing*. Australas Mark J. 2021;29(3):204–14.
19. Hancock JT, Bailenson JN. *The Social Impact of Deepfakes*. Cyberpsychology, Behav Soc Netw. 2021;24(3):149–52.
20. Natale S. *Deceitful Media*. Oxford University Press; 2021.
21. Gelbart H. *Scammers profit from Turkey-Syria earthquake*. BBC. 2023 Feb 14;

22. Ray S. *Imran Khan—Pakistan's Jailed Ex-Leader—Uses AI Deepfake To Address Online Election Rally*. Forbes. 2023 Dec 18;

23. Henley J. *AI-generated Putin asks Putin about his rumoured body doubles – video*. The Guardian [Internet]. 2023 Dec 14; Available from: https://www.theguardian.com/world/video/2023/dec/14/ai-generated-vladimir-putin-rumoured-body-doubles-video

24. Allyn B. *Deepfake video of Zelenskyy could be "tip of the iceberg" in info war, experts warn*. NPR [Internet]. 2022 Mar 16; Available from: https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

25. Milmo D. *Russia targets Paris Olympics with deepfake Tom Cruise video*. The Guardian. 2024 Jun 3;

26. Vaccari C, Chadwick *A. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News*. Soc Media Soc. 2020;6(1).

27. Chesney B, Citron D. *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*. Calif Law Rev [Internet]. 2019;107. Available from: https://www.californialawreview.org/print/deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security

28. Gregory S. *Fortify the Truth: How to Defend Human Rights in an Age of Deepfakes and Generative AI*. J Hum Rights Pract. 2023;15(3):702–14.

29. Roozenbeek J, van der Linden S, Goldberg B, Rathje S, Lewandowsky S. *Psychological inoculation improves resilience against misinformation on social media*. Sci Adv. 2022;8(34):1–11.

30. Roozenbeek J, van der Linden S. *Fake news game confers psychological resistance against online misinformation*. Palgrave Commun [Internet]. 2019;5(1):1–10. Available from: http://dx.doi.org/10.1057/s41599-019-0279-9

Maastricht University
STUDIO EUROPA MAASTRICHT

Policy Brief

The AI risks in very large online platforms and search engines

Dr Konstantia Zarkogianni

Summary

This policy brief focuses on the integration of AI technologies in Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) which serve more than 45 million users monthly under the Digital Services Act (DSA). It highlights the key AI technologies, including recommender systems, information retrieval systems, and generative AI, which shape user experiences and impact the dissemination of information. Recommender systems leverage user data to deliver personalised content, while information retrieval systems prioritise relevant documents in response to user queries. Generative AI, a transformative technology, enriches digital content but introduces risks such as hallucinations and the proliferation of misinformation, including deepfakes. Despite the benefits of AI integration, VLOPs and VLOSEs face significant systemic risks. These include privacy and security vulnerabilities, threats to user autonomy, dissemination of harmful content and the addictive nature of these platforms.

The brief discusses the growing concerns about AI-driven rabbit holes which steer users toward extreme content, as well as the potential for generative AI to spread disinformation. Addressing these risks requires a human-centred approach to AI regulation, emphasising ethical design, transparency, user control and human oversight. The policy brief calls for improved training of AI systems with diverse, high-quality data, collaboration among stakeholders and the implementation of explainability techniques to support user control and human oversight. It underscores the need for ongoing policy development to harmonise the DSA with the AI Act, ensuring that platforms and AI systems operate within a legal framework that protects individuals and society while enabling technological innovation.

Introduction

According to the Digital Services Act (DSA), online platforms and search engines are classified as 'very large' when reaching more than 45 million users per month. In general, online platforms primarily serve as social networks, marketplaces, content sharing sites or communication tools; while search engines are designed to help users find information on the internet by indexing websites and delivering relevant search results based on user queries. The widespread use of very large online platforms (VLOPs) and search engines (VLOSEs) enables a massive amount of data collection gathered upon users' registration (i.e., demographic) and through users' interaction (e.g., search queries, browsing history, user behaviour, location, social connections, cookies and tracking).

This huge amount of data has made feasible the development of AI algorithms capable of shaping user experiences and influencing information dissemination. Therefore, AI has become an integral part of VLOPs and VLOSEs with the aim to personalise content recommendations, prioritise search results and optimise the overall user experience. However, the reliance on AI-based technologies, such as recommender systems and information retrieval mechanisms, poses risks and challenges in terms of spreading illegal and harmful content and proliferating misinformation (1). Furthermore, the recent technological advancements in generative AI pose new risks that are linked to so-called hallucinations. Generative-AI can trigger hallucinations through generating and disseminating false information, including the widely known deepfakes, while manipulating services that can mislead voters (2).

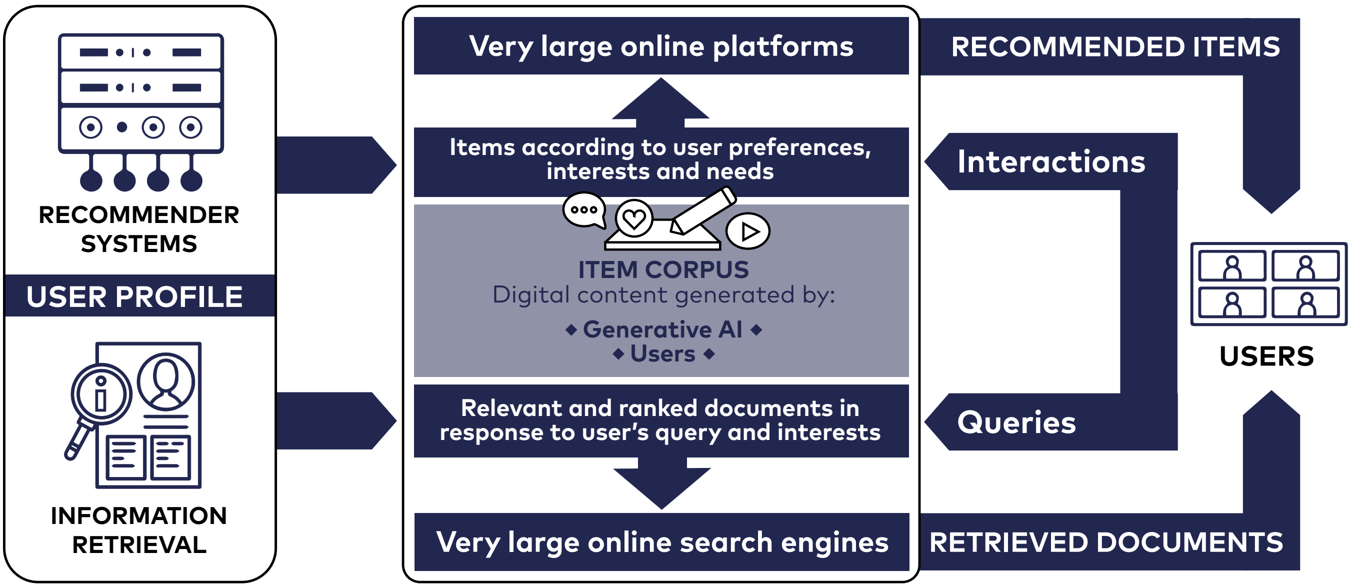
Although the benefits of integrating AI into VLOPs and VLOSEs are many and well recognised, there are challenges and risks for individuals and societies that should be thoroughly investigated in order to identify appropriate risk mitigation measures and develop meaningful and applicable regulations. In view of these challenges, there is the urgent need to bridge the DSA legislation framework that focuses on regulating intermediary services (including online platforms) with the one defined in the AI Act that governs the emerging AI technologies. Taking into consideration that the DSA was issued before the explosion of AI technologies with high societal impact, it lacks important legislation pieces regarding the integration of AI in these intermediary services. On the other hand, the AI Act constitutes the first-ever legal framework defining a risk-based approach concerning the use of AI. Although the DSA and AI Act were enacted independently, the regulation of platforms and the utilisation of AI systems are becoming more interconnected, as recognised in the preamble of the AI Act. Nevertheless, determining the appropriate legal framework for issues where AI intersects with platform regulation may necessitate reconciling two distinct and often parallel pieces of legislation (3).

The objective of this policy brief falls within the overall scope of the European Centre for Algorithmic Transparency (ECAT) that was launched in April 2023 to support the enforcement of the Digital Services Act (DSA). ECAT's mission is to delve into the AI algorithms that are implemented in favour of VLOPs and VLOSEs to understand their functionalities and assess their long-running impact. Within the frame of the proposed policy brief, a risk assessment strategy of these AI algorithms will be presented contributing to one of ECAT's main objectives: to develop practical methodologies towards fair, transparent and accountable algorithmic approaches.

Dominant AI technologies in VLOPs and VLOSEs

Personalisation and adaptation to specific users' needs and preferences constitute the main objectives of VLOPs and VLOSEs. This can be achieved by harnessing the power of AI, as depicted in Figure 1. The mission is to connect people with information available in the form of online digital content. Recommender systems are the dominant intelligent technologies integrated into VLOPs, recommending items to be presented to specific users based on their interactions with the platform. VLOSEs utilise information retrieval systems to identify relevant documents in response to user queries and prioritise them according to users' interests. The emerging and rapidly growing technology of generative AI has already penetrated the market, enriching the digital item corpus with more content. Each of these technologies are further analysed below.

Figure 1 The important role of AI in VLOPs and VLOSEs


RECOMMENDER SYSTEMS
USER PROFILE
INFORMATION RETRIEVAL
Very large online platforms
Items according to user preferences, interests and needs
ITEM CORPUS
Digital content generated by:
◆ Generative AI ◆
◆ Users ◆
Relevant and ranked documents in response to user's query and interests
Very large online search engines
RECOMMENDED ITEMS
Interactions
USERS
Queries
RETRIEVED DOCUMENTS

Recommender systems

According to the DSA, a recommender system is defined as follows:

'a fully or partially automated system used by an online platform to suggest in its online interface specific information to recipients of the service, including as a result of a search initiated by the recipient or otherwise determining the relative order or prominence of information displayed' (Regulation 2022/2065, Art. 2).

From a technological perspective, recommender systems leverage advanced data analysis and machine learning techniques to deliver personalised recommendations. A typical recommender system is constructed to retrieve items from an item corpus according to individual user preferences, interests and needs. This is achieved through incorporating a user profile module that is responsible for capturing user preferences, behaviours and demographics. The user profile collects and maintains information about the users' interactions with the online platform, including items they have viewed, rated, purchased or engaged with in any way.

Furthermore, recommender systems employ sophisticated algorithms to analyse the user profile data and identify patterns in user behaviour. These algorithms may include collaborative filtering which recommends items based on similarities between users' preferences, or content-based filtering which recommends items similar to those previously liked or interacted with by the user. Additionally, hybrid approaches that combine multiple recommendation techniques are often utilised to enhance the accuracy and effectiveness of recommendations. This intricate process of data analysis and algorithmic computation enables recommender systems to generate personalised recommendations that cater to the unique preferences and interests of each user, thereby enhancing user satisfaction and engagement with the platform.

Information retrieval

Information retrieval in VLOSEs refers to the process of efficiently and effectively retrieving relevant information from massive collections of documents in response to user queries. These VLOSEs, such as Google, Bing and Yahoo, index billions of web pages and other types of content, requiring sophisticated algorithms and infrastructure to handle the scale and complexity of the data. The information retrieval process includes two tasks. In the first task a set of relevant documents is retrieved based on the user's query. Recent techniques exploit the use of pre-trained language models like BERT to perform this task. In the second task, the objective is to rank the retrieved documents through estimating relevance scores based on the user's query criteria. Within the frame of the ranking task, different models are deployed compared to those utilised in the retrieval task in order to improve the effectiveness of the results. These models harness the power of reinforcement learning, contextual embeddings and attention mechanisms.

Generative AI

Generative AI constitutes a transformative technology for VLOPs and VLOSESs by enhancing user experience and providing innovative services. There are several key areas where Generative AI impacts VLOPs and VLOSESs. The most promising are the following:

Automated Content Generation: Generative AI can produce a wide range of content types, including text, images, music and videos. This capability can be used to constantly update and enrich the recommender's item corpus. Currently, the automated content generation is initiated by the users through manual prompts. However, the next generation of recommenders most probably will embed generative AI to automatically create personalised content in real-time with the aim to reduce the dependence on user-generated content while improving user experience.

Conversational agents: Generative AI powers sophisticated conversational agents that engage users in a natural and dynamic manner. These AI-powered interfaces leverage large language models to deliver comprehensive and contextually relevant responses across a broad range of tasks. From answering trivia questions and assisting with trip planning to generating personalised advice and handling complex, nuanced conversations, these systems offer a versatile and adaptable user experience. By seamlessly integrating into various VLOPS applications (e.g., My AI on Snapchat), they provide users with tailored interactions that enhance both convenience and engagement, making them valuable tools in both personal and professional contexts.

In view of the rapid technological advancements in generative AI and recognising their impact on VLOPs and VLOSEs, the DSA requested information from six VLOPs (Facebook, Instagram, Snapchat, TikTok, YouTube and X) and two VLOSEs (Bing and Google Search) about their mitigation measures for risks associated with generative AI. This request was published in a press release on March 2024 (4). DSA placed particular focus on clarifying and analysing the impact of generative AI on electoral processes, dissemination of illegal content, protection of fundamental rights, gender-based violence, protection of minors, mental well-being, personal data protection, consumer protection and intellectual property.

Systemic Risks

The recently introduced concept of high-reach AI refers to those AI systems whose widespread use may generate significant risks for both individuals and societies (1). VLOPs and VLOSEs fall into this category, posing certain risks and challenges as elaborated below.

Security vulnerabilities and privacy issues: The effectiveness of recommender systems in VLOPs in terms of offering users a personalised and unique experience relies on the collection and processing of users' personal and behavioural data. On the other hand, the storage and processing of vast amounts of such data make high-reach AI systems susceptible to attacks, manipulation or misuse which can lead to security breaches or malicious activities. Besides the security vulnerabilities, an important privacy issue is raised by user profiling techniques that consider personal and behavioural data, which might reveal personality features for which the user has not intentionally provided consent. Therefore, there is a trade-off between personalisation and privacy.

Impact on autonomy: High-reach AI systems incorporate sophisticated inference algorithms that facilitate user modelling and pave the way to predict and influence users' behaviour. This poses a risk to the users' fundamental right to control their access to information and communication flow. More specifically, the users' autonomy is being challenged by AI's power to steer their attention towards content that does not align with their needs and preferences, but rather serves the interests of the company (e.g., advertisements, new features).

Dissemination of harmful and illegal content: Systemic risks emerge when AI algorithms prioritise engagement and user satisfaction without sufficient regard for the legality or harmful nature of the content being promoted. This can lead to the dissemination of misinformation (false information regardless of indent), disinformation (false information with the intent to deceive), hate speech, extremist ideologies and other forms of harmful and illegal content. It is worth mentioning that the creators of fake news utilise sensational headlines that attract the attention of many users in order to boost ad revenue. A relevant study demonstrated that false news items spread more and faster than true ones on X (5). Furthermore, it has been argued that recommender systems in social platforms can provoke social polarisation and contribute in fragmented political discourse.

Addictive use: The AI algorithms in the VLOPs are designed to increase users' engagement and promote their retention. In the long-term this feature can cause addiction in the form of internet overuse and social isolation. Mental health can also be affected leading to issues such as anxiety, depression and decreased self-esteem. The constant exposure to curated content and social comparisons can exacerbate these mental health challenges, making it difficult for users to disconnect and engage in offline activities. Furthermore, the addictive nature of these platforms can impact productivity and overall well-being, creating a cycle that is hard to break.

Rabbit holes effect: The phenomenon of AI-driven rabbit holes refers to the tendency of algorithmic recommendations to steer users towards increasingly extreme or polarised content. This can lead users down paths characterised by confirmation bias, tribalism and ideological polarisation; this can ultimately undermine the quality of public discourse and erode trust in democratic institutions. Although the rabbit hole effect applies to all ages, the DSA places particular emphasis on children and young people. More specifically, on May 2024, DSA opened formal proceedings against Meta related to the protection of minors on Facebook and Instagram (6). Within the framework of the investigation, the potential addictive impacts of both platforms will be explored, particularly where an AI algorithm feeds young people negative content, such as unrealistic body images, driving them towards a rabbit hole effect.

Hallucinations: Hallucinations in generative AI are considered to be instances where the AI generates content that is plausible-sounding but factually incorrect or nonsensical. These hallucinations can have significant implications in the context of VLOPs and VLOSEs. As generative AI technology expands, so does its potential for wide-ranging misuse in creating affordable, highly convincing large-scale disinformation campaigns (2). According to a recently published white paper (2), the current landscape already presents damaging examples of AI-generated disinformation. Specific examples include deepfake videos in Facebook ads aiming to influence voters in Moldova, political deepfake ads on YouTube and AI-generated images spreading false information about Gaza and anti-immigrant narratives. Audio deepfakes, primarily involving fake political statements and conversations, have also been reported. Russian disinformation campaigns have exploited generative AI to create deceptive content, such as videos and conversations featuring the Ukrainian president.

Policy advice

Key technological challenges
There are certain key technological challenges that can be addressed to improve the performance of the AI algorithms incorporated into VLOPs and VLOSEs while simultaneously reducing the systemic risks.

AI algorithms are usually vulnerable to different types of biases: data, model and human. Data biases refer to the skewed or unrepresentative training data that can lead to biased outcomes. Model biases arise from the design and architecture of the AI system itself, potentially amplifying existing biases in the data. Human biases involve the prejudices and assumptions of the developers and users which can inadvertently influence the AI's behaviour and outputs. The presence of such biases may lead to discriminatory and misleading outcomes in content recommendation and presentation.

The operation of AI algorithms lacks accountability and transparency due to their complex and sophisticated nature. More specifically, the opacity surrounding AI algorithms in VLOPs and VLOSEs makes it difficult for users, policy-makers and researchers to understand how content recommendations are generated and to hold platforms accountable for the spread of illegal or harmful content.

Policy recommendations

This policy brief recommends following a human-centred approach that can help in balancing the benefits of AI with the need to protect individuals and society from its potential risks (Figure 2). The focus should be placed in the following aspects:

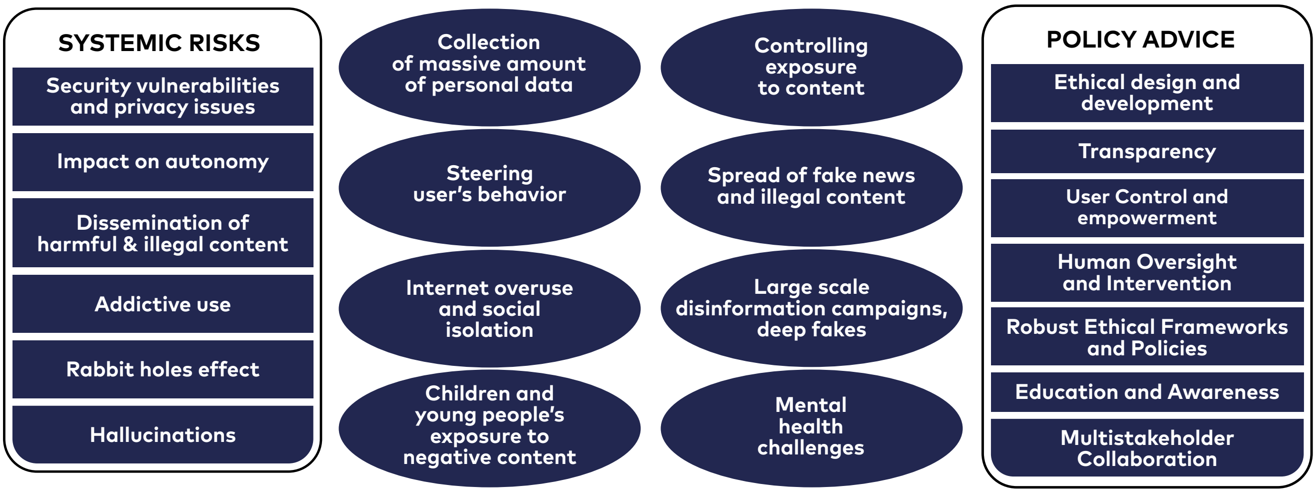
Ethical design and development: An inclusive data collection ensuring that training data is diverse and representative of various populations can reduce biases and improve fairness. Even in the case of VLOPs and VLOSEs that include millions of users, biases can emerge if certain groups are under-represented or overrepresented. For example, linguistic minorities receive limited content tailored to their language and culture. Mechanisms for detecting and mitigating biases in AI-based algorithms, including regular audits, transparency requirements and algorithmic impact assessments should be implemented. Platforms should prioritise Diversity, Equity and Inclusion (DEI) in algorithmic design and optimisation processes. However, defining DEI is challenging within the global frame of the VLOPs and VLOSEs where different values are applied to users across various countries, regions and cultures. The definition of DEI requires a balance between universal principles and localised approaches. One way to face this challenge is to include people with diverse perspectives in the process of building, developing and using these AI systems.

Transparency: The AI systems should incorporate effective interpretable and explainable mechanisms capable of generating reasoning of their decisions and processes in a human understandable way. This is of particular importance since transparency and explainability help users trust and verify the outputs of the AI systems.

User control and empowerment: The users should have the control of their data and be aware of how their data is used by the AI systems. The implementation of human-in-the-loop technologies that enable feedback from the users in response to the AI outputs can provide valuable information in continuously improving the algorithms and correcting errors.

Human oversight and intervention: Combining AI with human oversight by means of hybrid intelligence systems to review and validate user-generated content can prevent harmful disinformation and errors. Content moderation systems are human-machine hybrids, and their integration into very large online platforms is essential for detecting inappropriate content. Depending on the degree of automation, these systems utilise AI to flag potential violations in user-generated content and decide whether to take immediate action or escalate for human review. Higher degrees of automation allow for quicker actions, although human oversight is crucial in complex cases. The Digital Services Act (DSA) regulates content moderation procedures by requiring providers of VLOPs to include internal complaint-handling systems that allow end users to easily and effectively contest moderation decisions. This creates a demand for large moderation workforces which has become impractically large and overloaded. A better approach to human supervision is to improve automated moderation by training AI models with high-quality data gathered through the involvement of highly qualified workers (such as policy-makers) to review a representative sample of moderation decisions. Additionally, AI content moderation systems should incorporate explainability techniques to provide human reviewers with sufficient information about the reasons for flagging content [7].

Figure 2 Overview of the systemic risks associated with AI in VLOPs and VLOSEs along with policy advice to develop strategies for mitigating these risks. The concepts in circles provide further details and examples about the risks.


SYSTEMIC RISKS
Security vulnerabilities and privacy issues
Impact on autonomy
Dissemination of harmful & illegal content
Addictive use
Rabbit holes effect
Hallucinations
Collection of massive amount of personal data
Steering user's behavior
Internet overuse and social isolation
Children and young people's exposure to negative content
Controlling exposure to content
Spread of fake news and illegal content
Large scale disinformation campaigns, deep fakes
Mental health challenges
POLICY ADVICE
Ethical design and development
Transparency
User Control and empowerment
Human Oversight and Intervention
Robust Ethical Frameworks and Policies
Education and Awareness
Multistakeholder Collaboration

Education and awareness: Educating users about the capabilities and limitations of AI can help them better understand and critically evaluate AI-generated content. Training developers on ethical AI practices and the potential impacts of their work can lead to more responsible AI development.

Multi-stakeholder collaboration: Foster collaboration between platforms, policymakers, regulators, researchers and civil society organisations to address systemic risks associated with AI-based algorithms in VLOPs and VLOSEs. Multi-stakeholder dialogue and cooperation are essential for developing holistic and effective policy responses.

Conclusions

The AI-based algorithms in VLOPs and VLOSEs are used to identify, categorise, rank, suggest and present information to users. Therefore, they often replace humans in decision-making processes. Within the frame of the proposed policy brief, the functionality of the AI-based algorithms are presented and connected with their potential in spreading illegal and harmful content that can reach a massive number of users. The sources of harmful biases (data, AI model, human) and content spreading are outlined along with measures to mitigate them.

Author Information

Konstantia Zarkogianni is an Associate Professor of Human-Centered AI at the Department of Advanced Computing Sciences, Faculty of Science and Engineering, Maastricht University. Her research focuses on augmenting human intelligence in decision-making through intelligent user interfaces that enable effective human-AI collaboration.

References

1. Söderlund, K. & Engström, E. & Haresamudram, K. & Larsson, S. & Strimling, P. (2024). *Regulating high-reach AI: On transparency directions in the Digital Services Act*. Internet Policy Review, 13(1). https://doi.org/10.14763/2024.1.1746
2. *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities*: https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf
3. *The Digital Services Act Meets the AI Act: Bridging Platform and AI Governance*: https://www.techpolicy.press/the-digital-services-act-meets-the-ai-act-bridging-platform-and-ai-governance/
4. *Commission sends requests for information on generative AI risks to 6 Very Large Online Platforms and 2 Very Large Online Search Engines under the Digital Services Act*: https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-generative-ai-risks-6-very-large-online-platforms-and-2-very
5. Vosoughi, S., Roy, D., & Aral, S. (2018). *The spread of true and false news online*. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559
6. *Commission opens formal proceedings against Meta under the Digital Services Act related to the protection of minors on Facebook and Instagram*. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2664
7. R. Griffin, E. Stallman. *A Systemic Approach to Implementing the DSA's Human-in-the-Loop Requirement*, VerfBlog, 2024/2/22, https://verfassungsblog.de/a-systemic-approach-to-implementing-the-dsas-human-in-the-loop-requirement/, DOI: 10.59704/b2a7a2ee0ff8bd31.

Policy Brief

Immersion and regulation: extended reality technologies, their impact on innovation and policy recommendations

Prof. Dominik Mahr, Dr Jonas Heller, Dr Tim Hilken

Summary

Extended Reality (XR) technologies, encompassing Augmented Reality (AR) and Virtual Reality (VR), are primed to revolutionise digital interactions across various sectors, from retail and education to entertainment and healthcare. As these immersive technologies rapidly evolve, they present both unprecedented opportunities and novel challenges for citizens, businesses and policy-makers. This policy brief examines the current landscape of XR technologies, their potential impacts on citizens and society and the regulatory implications surrounding their development and implementation.

XR offers significant benefits, including enhanced access to services, improved learning experiences and new forms of creative expression. However, it also raises concerns about privacy, data protection and potential negative psychological effects, such as addiction and difficulties distinguishing between virtual and real experiences. The Digital Services Act (DSA) and Digital Markets Act (DMA) provide a regulatory framework that both supports and potentially hinders XR innovation.

The DSA and DMA are expected to have a mixed but overall positive long-term impact on XR innovation. While compliance efforts may slow down innovation for smaller XR developers due to the complexity of content moderation and data protection, fair competition, enhanced transparency and interoperability are likely to foster innovation, increase user trust and attract more users over time; though challenges around real-time moderation and achieving true interoperability remain.

Policy recommendations to avoid stifling innovation in this emerging field include:

- Establishing balanced XR specific transparency requirements that protect proprietary technologies while ensuring public trust;
- Providing support for SMEs and start-ups to navigate the complex XR regulatory landscape;
- Developing clear and simple to implement data privacy rules specific to XR technologies;
- Creating XR interoperability standards in collaboration with industry stakeholders;
- Penalising companies that do not implement data portability features.

By adopting these recommendations, the EU can position itself as a leader in responsible XR development and deployment. This approach will ensure European companies remain competitive in the global XR market while protecting consumer rights and privacy.

Introduction into extended reality technologies

XR technologies, encompassing augmented reality (AR) and virtual reality (VR) are transforming digital interactions by merging the digital and physical worlds. AR overlays digital information onto the real world, and VR creates fully immersive digital environments (Hilken et al., 2017). These XR technologies offer citizens in their roles as user or consumer enriched, interactive experiences far beyond traditional digital interfaces. In retail, AR applications like Ikea's Place app enable customers to visualise furniture in their homes before making a purchase, enhancing decision comfort and reducing returns (Heller et al., 2019a, b). In education, immersive technologies provide interactive and immersive learning experiences, such as Microsoft's HoloLens, which allows students to engage with 3D models and simulations, making complex subjects like anatomy more accessible and engaging (Hilken et al., 2018).

Artificial Intelligence (AI) significantly advances immersive technologies by enabling personalised and context-aware experiences. AI algorithms analyse citizen behaviour and preferences to deliver tailored content, enhancing relevance and engagement. For example, AI can enhance AR applications by providing real-time recommendations based on previous interactions, further improving the citizen experience. The entertainment industry also benefits from XR; VR gaming offering highly immersive experiences that transport players into entirely virtual worlds, and AR enhances live events with interactive elements that enrich the audience's experience (see Figure 1 for an overview).

The platforms of XR technologies

Devising Digital Services Act (DSA) policies for XR technologies has proven challenging due to the diverse terminology and branding strategies employed by major companies like Meta (metaverse), Apple (spatial computing) and Microsoft (mixed reality), as they seek to capture market share and position themselves in the rapidly evolving technological landscape. This fragmented landscape makes it difficult to gain a comprehensive understanding of the technologies, which is essential for effective policy development. We propose using XR as an umbrella term, as it encompasses the most relevant use cases across augmented reality (AR) and virtual reality (VR). AR and VR serve as the core technologies, while concepts like the metaverse, spatial computing and mixed reality build on these foundations to expand into areas such as social VR environments, innovative interfaces for work and entertainment and the merging of digital and physical spaces (DEXLab, 2024).

Table 1 below showcases the proliferation of terminology and the wide range of potential use cases across various aspects of citizens' lives, from entertainment and education to healthcare and the workplace. This underscores the need to

Figure 1 Illustration of XR Technologies

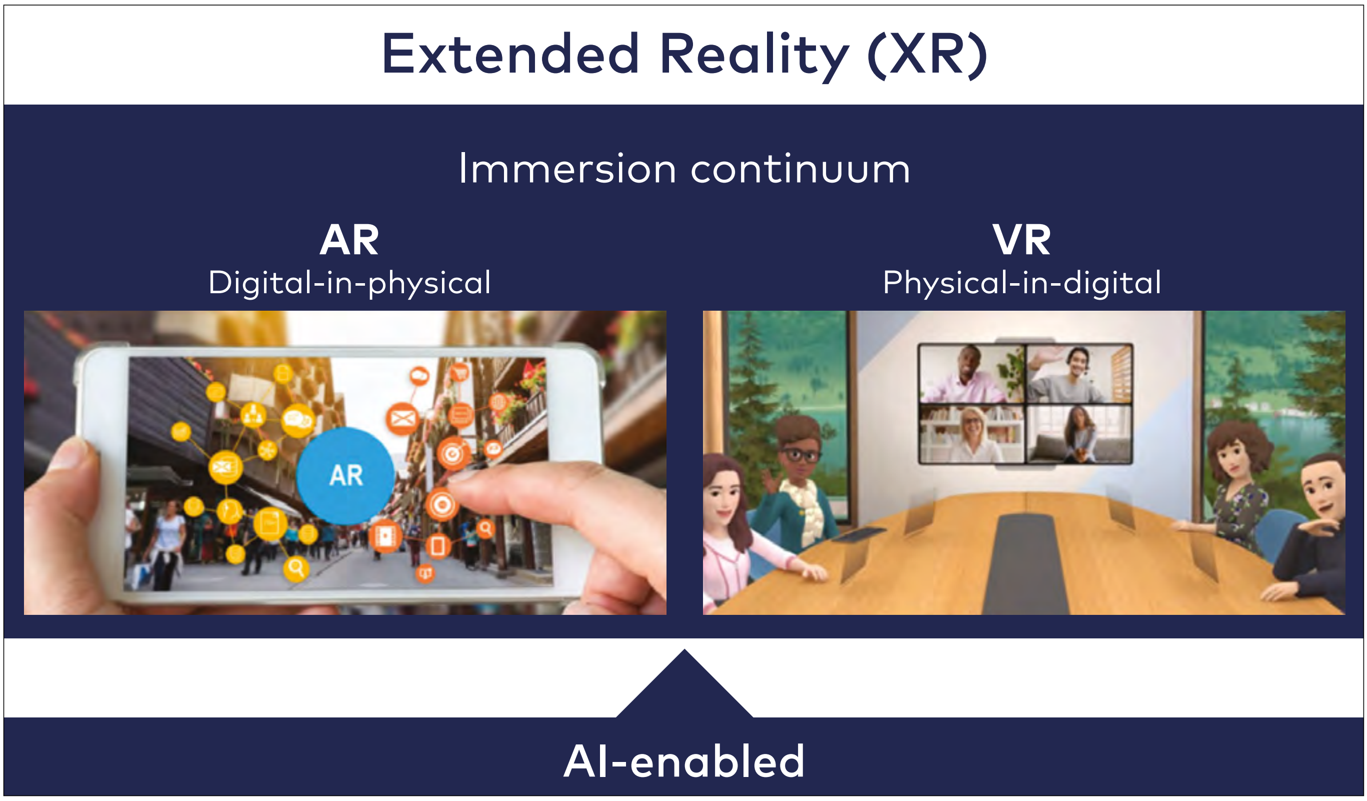


Table 1 Illustrative XR Use cases and platforms

Example	Platform/Applications	What it means for citizens	Description
Virtual Concerts	Roblox, Decentraland	Provides accessible service experiences	Hosting massive interactive live events where all citizens have premium access
Creator Tools	Unity, Unreal Engine	Empowers citizen to create and monetize content, promotes entrepreneurship	Provides tools and marketplaces for citizen-generated content creation and monetization
Spatial Computing	Microsoft HoloLens, Snapchat Landmarker	Blends virtual and physical worlds for natural citizen experiences	Uses AR to create seamless interactions between digital and physical environments
AR Shopping Apps	IKEA Place, Amazon AR View	Allows citizens to visualize products in their real environment before purchase	AR apllications that help citizens see how furniture or products will look in their home
VR Education Tools	Google Expeditions, ENGAGE	Offers immersive learning & training experiences	VR tools that enable students to explore virtual field trips and interactive lessons
XR Fitness Apps	Supernatural,FitXR	Provides immersive workout & physical health experiences	Fitness apps tha use VR to create engaging and interactive workout environments

better understand not only the commonalities and differences among these XR technologies, but also the benefits and concerns they present. A clearer understanding of the potential societal impact—such as privacy, security, accessibility, and ethical considerations—is vital for developing nuanced, future-proof policies that maximise the benefits of XR while addressing possible risks and challenges.

Despite the rapid advancements in XR technologies, the market remains fluid with no clear leaders emerging. This presents significant opportunities for new entrants and innovations to disrupt the status quo and establish leadership. As such, a more informed policy approach will be key in fostering innovation while ensuring that the adoption of these technologies serves the broader public interest.

Benefits and concerns for citizens using XR

These use cases engender many benefits to citizens but also present novel risks. The benefits include facilitating access to services that enhance citizen well-being, such as education, cultural events or healthcare. Often these are inaccessible to certain groups of the population due to physical distance, personal or resource limitations. XR also enriches learning experiences, enabling students and workers to better develop their knowledge and skills, for instance, through novel types of virtual simulations and trainings (Won et al., 2023).

Furthermore, by supporting positive forms of digital escapism, XR allows individuals to gain new perspectives and foster creativity and inspiration (Jessen et al., 2019). XR offers enhanced opportunities for self-expression in the digital sphere, for instance, through the design of a lifelike (or fantastical) avatar (Golf-Papez et al., 2022). Social XR platforms can also promote human connection and support. For instance, AR enables consumers to better exchange advice about purchase decisions, fostering a sense of social empowerment (Hilken et al., 2020). While VR enhances empathy by allowing people to take the perspective of the recipients of charitable donations, thereby fostering a deeper connection with distant others (Kandaurova & Lee, 2019). Additionally, XR technologies facilitate virtual travel experiences, enabling citizens to explore distant locations without the need for physical transportation, potentially reducing carbon emissions associated with traditional travel.

By the same virtue, however, XR also introduces new risks for citizens' rights and well-being. Due to the high degree of realism offered in many XR experiences, it can become difficult for people to distinguish between actual events and those that occurred in XR (Slater et al., 2020). This has the potential to skew perceptions of actual events and is vulnerable to unethical use in marketing and especially in persuasion attempts, whether intentional or unintentional (Mahr et al., 2020). XR can also bias perceptions of reality, leading to unrealistic expectations. For example, in the context of beauty ideals, sharing AR face filters (e.g., of makeup styles) on social media can amplify body dysmorphia effects, creating gaps between a person's perceived actual versus ideal self and resulting in decreased self-compassion (Javornik et al., 2021).

Next, there are concerns that XR can make it more difficult to disengage from immersive experiences, resulting in addiction and broader social isolation (Merkx & Nawijn, 2021). Conversely, XR may also support cognitive closure through overstimulation effects, where people believe they have already seen it all (Pala et al,. 2022). Current echo chamber effects on social media might be further enhanced in shared, private VR spaces.

Finally, XR poses novel concerns for privacy and personal data protection (Lammerding et al., 2021). These include a lack of information about what novel types of data are captured (e.g., location or movement data in virtual spaces, contextual data when using AR cameras) and how organisations might use this data to provide contextualised and hyper-personalised communications (e.g., 3D images of promoted products appearing in one's home in AR). Moreover, the energy consumption associated with powering XR devices and data centres supporting these immersive experiences raises sustainability concerns, potentially offsetting some of the environmental benefits of well-being and reduced physical travel.

DSA and DMA fostering and hindering XR innovation

This section demonstrates how the DSA (Digital Services Act) and DMA (Digital Market Act) work together to empower businesses within the European XR ecosystem and foster a thriving digital market. While the DSA focuses on protecting consumer rights, the DMA tackles the dominance of gatekeepers (i.e., large online platforms) to create a fairer playing field for European SMEs and start-ups, especially critical for emerging XR technologies and developing ecosystems. Aligning XR development with regulatory principles will be vital in ensuring positive contributions of companies and non-commercial organisations to the digital ecosystem.

Table 2 below provides an overview of how DSA and DMA goals and their regulatory fields might support or hinder XR innovation of organisations developing and using XR solutions.

Table 2 Innovation affected by DSA and DMA

DSA and DMA Goals	Regulatory fields	XR Innovation Support	XR Innovation Hindrance
Enhanced Transparency	**Content Moderation** The DSA requires platforms to be transparent about their content moderation policies and algorithms.	XR developers can better align their content creation with platform standards, reducing the risk of content being removed or demoted without clear reasons.	XR environments often involve citizen-generated content that is highly interactive and dynamic. Ensuring real-time moderation can be extremely challenging and resource-intensive. Forcing XR companies to reveal proprietary algorithms and methods potentially reduces their competitive edge.
	Data Access The DSA mandates that platforms provide data to researchers and authorities.	XR developers and researchers can benefit from greater access to data, which can be used to improve VR experiences and innovate new solutions.	XR applications often collect extensive personal data to function effectively (e.g., spatial data, citizen interactions). The requirement to share this data may raise significant privacy concerns, leading to citizen reluctance to engage deeply with these technologies.
Stronger Consumer Protection	**Safety and Trust** By emphasising the protection of citizens from illegal content and ensuring their safety online, the DSA helps build trust in XR platforms.	Increased citizen trust can lead to higher adoption rates of XR technologies, encouraging more investment and innovation.	The obligation to share sensitive citizen data increases the risk of data breaches and slows down the adoption of XR technologies.
	Liability Framework Clearer rules on the liability of digital service providers help reduce legal uncertainties for XR companies.	Such a framework allows developers to focus more on innovation rather than worrying about compliance complexities and legal issues.	Stricter liability rules can increase the risk and cost associated with developing and launching new AR/VR products. Companies might become more cautious and avoid potential legal repercussions.
Fair Competition	**Gatekeeper Regulation** The DMA targets large platforms that act as gatekeepers to prevent anti-competitive practices.	This regulation ensures that smaller XR companies can compete on a level playing field, fostering a more dynamic and innovative market.	XR companies that develop unique ecosystems might be forced to open up their platforms, leading to less incentive to innovate.
	Interoperability The DMA promotes interoperability between different platforms and services.	For XR, this can mean more seamless integration of different XR systems and applications, fostering innovation through collaboration and compatibility.	Achieving true interoperability in XR systems can be technically complex. Companies might be forced to conform to specific standards, potentially limiting the unique features and advancements they can offer.
Market Access	**Equal Treatment** The DMA requires gatekeepers to treat all business citizens fairly.	XR developers benefit from fair access to essential platforms and services, which is crucial for bringing innovative products to market.	Compliance with equal treatment provisions may impose additional operational burdens on XR companies, particularly smaller ones.
	Data Portability Enhanced data portability rules allow citizens to transfer their data between different XR systems easily.	This can drive innovation by enabling developers to create more personalised and citizen-centric experiences across XR systems.	Ensuring compliance with data portability requirements can be resource-intensive, potentially slowing down the pace of innovation as companies focus on regulatory adherence.

To summarise, the DSA and the DMA support XR innovation by bringing clarity and fairness, which attracts investment into these high-cost, R&D-intensive fields. Clear rules encourage investors to fund projects, while a focus on interoperability and fair competition promotes collaboration among companies, leading to innovative solutions. Increased transparency, safety and data portability enable XR companies to create more citizen-centric innovations, enhancing personalisation and citizen satisfaction.

However, compliance with these regulations can be burdensome, particularly for smaller companies. The fear of ongoing regulatory adjustments creates uncertainty and complicates long-term planning. Privacy concerns from data sharing requirements may reduce citizen engagement, limiting the data needed for XR innovation. Efforts to comply with content moderation and interoperability requirements might compromise citizen experience and diminish the appeal of XR applications. For example, a user with a high-end VR headset might experience a less immersive environment because the platform needs to accommodate users with simpler devices like mobile phones or basic VR glasses. As a result, the overall experience might feel compromised which can reduce user satisfaction and potentially diminish the appeal of the XR application. Additionally, large tech companies may hesitate to invest in high-risk projects due to the stringent regulatory environment, potentially slowing overall industry innovation.

Assessment of DSA and DMA with respect to XR

Drawing from the above-mentioned aspects and our expertise, we expect that the DSA and DMA are likely to have a mixed but generally positive long-term impact on XR innovation.

- Compliance efforts will slow down innovation, particularly for smaller developers who may struggle with the cost and complexity of moderating content and protecting data in immersive 3D environments and managing new types of data. This might involve implementing AI-powered content filters, real-time monitoring of virtual interactions or restrictions on user-generated content.

- Fair competition will foster innovation from smaller XR companies rather than allowing only the big players to dominate the market. The DMA's efforts to reduce platform dominance—by limiting gatekeeper control over app stores, payments, or user data—will enable start-ups and independent developers to benefit from fairer access to distribution channels and reduce their dependence on major platforms like Meta and Apple.

- Enhanced transparency will increase user trust in XR platforms, especially given the highly sensitive nature of interactions within these environments. Combined with shared interoperability requirements that enable more seamless experiences across platforms, this is expected to attract more users in the long run.

- Difficult to predict how content moderation will be implemented in real-time, dynamic XR environments as this is technically challenging. Additionally, achieving true interoperability across different operating systems and 3D modeling standards remains uncertain.

Policy recommendations

Below, we describe specific suggestions for EU and national policy-makers involved in implementing DSA and DMA. The suggestions include a brief reasoning and point for action

- **Balanced Transparency Requirements**: Protecting proprietary technologies while ensuring transparency can foster innovation and maintain competitive advantages. Keep focus on providing high-level explanations and anonymised data. ***Ensure transparency requirements do not force XR companies to reveal proprietary algorithms and methods.***

- **Support for SMEs and Start-ups**: Smaller companies often lack the resources to navigate regulatory landscapes. ***Provide financial and technical support to XR SMEs and start-ups to help them comply with DSA and DMA regulations.***

- **Clear Data Privacy and Security Rules**: XR applications often collect extensive personal data. Clear, consistent guidelines can help protect users while allowing developers to use data to improve and innovate their technologies. ***Establish clear and simple to implement guidelines for privacy and data security that balance user protection with the need for data-driven innovation.***

- **Interoperability Standards**: Interoperability drives innovation if these standards are developed in collaboration with industry stakeholders to ensure they are practical and do not limit technological advancements. ***Policy-makers take active role in the creation of (industry-specific) interoperability standards for XR.***

- **Incentives for Data Portability**: Enhanced data portability can drive innovation by enabling users to move their data across platforms easily, encouraging developers to create more personalised and user-centric experiences. ***Penalise companies that do not implement data portability features.***

- **Gradual Implementation of Liability Rules**: Reducing the immediate burden of compliance (e.g., accountability for harm) can help companies focus on developing new technologies without excessive risk. ***Phase in liability rules gradually and provide clear guidelines to help companies.***

- **Consumer Education and Awareness**: Education initiatives can enhance user adoption and provide valuable feedback for developers. ***Implement programmes to educate users about their rights and the benefits of XR technologies.***

- **Regulatory Sandboxes**: So-called sandboxes allow companies to test new innovations in a controlled environment, helping regulators understand the implications of new technologies and adjust policies accordingly. ***Establish regulatory sandboxes for XR technologies.***

By adopting these recommendations, policy-makers can create a regulatory environment that fosters growth in XR technologies and regulate how users interact with the digital and physical worlds. These innovations enhance consumer experiences across various sectors from retail to education by providing interactive and immersive environments that traditional interfaces cannot match. The Digital Services Act (DSA) and Digital Markets Act (DMA) offer frameworks to ensure fair and transparent digital ecosystems, but they must evolve to address the unique aspects of fostering XR innovation and protecting consumers effectively.

Author Information

Dominik Mahr is a Professor in Digital Innovation & Marketing specialising in the impact of digital data and technologies on individuals, organisations and society at large.

Tim Hilken is an Assistant Professor researching the impact of new technologies such as augmented and virtual reality on user experiences, decision-making, and well-being in consumer and industrial markets.

Jonas Heller is an Assistant Professor focusing on Augmented and Virtual Reality applications in retailing and services, digital marketing, and customer experience with emerging technologies.

All three authors are part of the Department of Marketing & Supply Chain Management at Maastricht University's School of Business and Economics. They are also co-founders of the Digital Experience Lab (DEXLab) that drives research on emerging technologies such as XR

References

1. DEXLab (2024). *Navigating the Immersive Terminology Forest: What is AR, VR, MR, XR, & co.?* Retrieved 06/10/2024 from https://www.sbe-dexlab.com/post/navigating-the-immersive-terminology-forest-what-is-ar-vr-mr-xr-co.
2. Golf-Papez, M., Heller, J., Hilken, T., Chylinski, M., de Ruyter, K., Keeling, D. I., & Mahr, D. (2022). *Embracing falsity through the metaverse: The case of synthetic customer experiences*. Business Horizons, 65(6), 739-749.
3. Heller, J., Chylinski, M., de Ruyter, K., Mahr, D., & Keeling, D. I. (2019a). *Let me imagine that for you: Transforming the retail frontline through augmenting customer mental imagery ability*. Journal of Retailing, 95(2), 94-114.
4. Heller, J., Chylinski, M., de Ruyter, K., Mahr, D., & Keeling, D. I. (2019b). *Touching the untouchable: exploring multi-sensory augmented reality in the context of online retailing*. Journal of Retailing, 95(4), 219-234.
5. Hilken, T., Heller, J., Chylinski, M., Keeling, D. I., Mahr, D., & de Ruyter, K. (2018). *Making omnichannel an augmented reality: the current and future state of the art*. Journal of Research in Interactive Marketing, 12(4), 509-523.
6. Hilken, T., Keeling, D. I., de Ruyter, K., Mahr, D., & Chylinski, M. (2020). *Seeing eye to eye: social augmented reality and shared decision making in the marketplace*. Journal of the Academy of Marketing Science, 48, 143-164.
7. Javornik, A., Marder, B., Pizzetti, M., & Warlop, L. (2021). *Augmented self-The effects of virtual face augmentation on consumers' self-concept*. Journal of Business research, 130, 170-187.
8. Jessen, A., Hilken, T., Chylinski, M., Mahr, D., Heller, J., Keeling, D. I., & de Ruyter, K. (2020). *The playground effect: How augmented reality drives creative customer engagement*. Journal of Business Research, 116, 85-98.
9. Kandaurova, M., & Lee, S. H. M. (2019). *The effects of Virtual Reality (VR) on charitable giving: The role of empathy, guilt, responsibility, and social exclusion*. Journal of Business Research, 100, 571-580.
10. Lammerding, L., Hilken, T., Mahr, D., & Heller, J. (2021). *Too real for comfort: Measuring consumers' augmented reality information privacy concerns. In Augmented reality and virtual reality: New trends in immersive technology* (pp. 95-108). Cham: Springer International Publishing.
11. Mahr, D., Caic, M., & Odekerken-Schröder, G. (2020). *An interdisciplinary view of marketing ethics*. In The SAGE handbook of marketing ethics (pp. 58-73). London: Sage.
12. Merkx, C., & Nawijn, J. (2021). *Virtual reality tourism experiences: Addiction and isolation*. Tourism Management, 87, 104394.
13. Slater, M., Gonzalez-Liencres, C., Haggard, P., Vinkers, C., Gregory-Clarke, R., Jelley, S., ... & Silver, J. (2020). *The ethics of realism in virtual and augmented reality*. Frontiers in Virtual Reality, 1, 512449.
14. Won, M., Ungu, D. A. K., Matovu, H., Treagust, D. F., Tsai, C. C., Park, J., ... & Tasker, R. (2023). *Diverse approaches to learning with immersive Virtual Reality identified from a systematic review*. Computers & Education, 195, 104701.

Editor in chief:
Dr Philippe Verduyn, Faculty of Psychology and Neuroscience, Maastricht University.

Design:
Oburon Grafisch Ontwerp.

Acknowledgements:
The authors of this Policy Brief collection express their gratitude to the editor and reviewers for their valuable feedback and constructive suggestions, which have contributed significantly to the quality of this work. We also thank colleagues who contributed their expertise in other capacities.

Date of publication:
March 2025

DOI:
10.26481/mup.rep.sem.2501

Images used in this publication were generated using artificial intelligence and are intended for illustrative purposes only.

Publisher:
Maastricht University Press | Maastricht University, The Netherlands, on behalf of Studio Europa Maastricht

STUDIO EUROPA MAASTRICHT

Maastricht University