Kempelen
Institute of
Intelligent
Technologies

Analysis of selected regulations proposed by the European Commission and technological solutions in relation to the dissemination of disinformation and the behaviour of online platforms.

Analysis of selected regulations proposed by the European Commission and technological solutions in relation to the dissemination of disinformation and the behaviour of online platforms

This study was prepared by the Kempelen Institute of Intelligent Technologies based on an order from Miriam Lexmann, Member of the European Parliament. The contract for services was signed on 1 October 2021.

The authors of the study are Matúš Mesarčík, Róbert Móro, Michal Kompan, Juraj Podroužek, Jakub Šimko and Mária Bieliková from Kempelen Institute of Intelligent Technologies.

Suggested Citation:
Mesarčík, M., Móro, R., Kompan, M., Podroužek, J., Šimko, J., Bieliková, M. Analysis of selected regulations proposed by the European Commission and technological solutions in relation to the dissemination of disinformation and the behaviour of online platforms. March 2022.

Executive summary

Freedom of access to information, freedom of speech and the ability to evaluate the veracity of information and make decisions based on it are among the fundamental pillars of democratic society. New technologies, including social media and other online platforms, have given all of us unprecedented possibilities to express our opinions and reach a potentially large audience. On the other hand, with the addition of artificial intelligence systems, above all personalised recommender systems, online platforms help disseminate false information, including disinformation with significantly negative impacts on society.

With regard to these negative impacts on the one hand and the strong respect for fundamental human rights and freedoms enshrined in European legal systems on the other, disinformation represents one of the greatest challenges for current initiatives for the creation of rules to regulate content and liability on the internet.

The aim of this study is to contribute to the ongoing discussion on possibilities of regulating online platforms and the artificial intelligence (AI) systems they use. The European Commission has presented several proposals for regulation in this regard. As stated in Part 1 of the document (*Introduction*), our objective was to analyse technological solutions for online platforms in relation to dissemination of disinformation and subsequently evaluate the effectiveness of the institutions of these proposed provisions, in particular the proposal for **artificial intelligence act** and proposal for **digital services act**. Specifically, we have focused on the research question:

> *'Are the proposed regulations an adequately effective tool for combating disinformation in the context of the technological solutions the online platforms use?'*

In Part 2 of the document (*Disinformation and its dissemination on online platforms*) we state that, although disinformation does not have a legal definition laid down in EU legislation, several non-binding parts of legal acts, preambles, studies, strategies and recommendations agree that disinformation can be understood as 'verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm.'[1]

Online platforms and particularly very large online platforms have fundamentally changed the way in which people gain information and, for many, have become a main news source. However, the mission of online platforms is not to provide balance of views and objectively inform their users, but they are based on a model of **attention economy**. Their objective is therefore to attract users' maximum attention for the purpose of showing online ads, which

[1] See e.g.: The European Commission. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Tackling online disinformation: a European Approach COM/2018/236 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236.

represent the main source of income for online platforms. For this, they use attractive user interfaces with habit-forming elements and **personalised recommendation methods** for the purpose of providing engaging content and relevant online advertising.

Recommender systems try to reduce the information load on users by filtering relevant information. Several approaches to recommendation exist; personalised approaches must always have some information about user preferences available. We have listed several possible impacts of using these systems both in general and specifically in the context of online platforms.

We have identified the key issues as following: the **level of transparency of the generated recommendations** and the **current setting of the utility function of recommender systems** in the process of training artificial intelligence models. Although it is not always technically possible to provide a satisfactory explanation as to why something has been recommended to the user (this is an ongoing research challenge), the minimum requirement is transparency at the level of inputs used by the recommender system and at the level of the chosen utility function, and the setting of its weighting, since these fundamentally affect the output recommendations. Other problems we have identified are the enclosing of users into information bubbles, bias and fairness in recommendations, and the collection of feedback and deriving user preferences in the context of privacy.

In Part 3 of the document (*Selected regulations proposed by the European Commission in relation to the dissemination of disinformation and the behaviour of online platforms*), we have identified a set of fundamental rights and freedoms that could potentially be affected by the legislative mechanisms to tackle the spread of disinformation on online platforms. From the users' point of view, these may be, in particular, interference with freedom of expression, the right to information, the protection of privacy and personal data, or the right to a fair trial. However, legislative instruments can also affect the online platforms themselves, primarily in terms of the freedom to conduct business or the protection of property rights in the form of the protection of intellectual property rights.

In view of the declared objectives, we have focused on analysing the European Commission's new proposal for legislation on artificial intelligence (proposal for regulation on artificial intelligence) and digital services, including online platforms (proposal for regulation on digital services). We have focused on the legislative instruments in these proposals for regulation that may have a greater impact on the dissemination of disinformation in the online environment.

The proposal for regulation on artificial intelligence (Artificial Intelligence Act, AIA) is conceived as horizontal regulation based on risk analysis. Artificial intelligence systems are divided into four types according to the degree of threat (or potential threat) to the rights and freedoms of individuals. In the context of tackling disinformation, the way the regulation in question defines prohibited 'manipulative practices' will be crucial as, **according to the current wording of the proposal, support for the dissemination of disinformation by online platforms**

would not fall under prohibited practices. Simultaneously, it should be noted that the classification of AI systems does not allow the activities of online platforms to be included in any of the areas for the use of high-risk AI systems. Nevertheless, the proposed regulation arranges a number of interesting mechanisms for assessing the conformity or management of data that can help tackle the dissemination of disinformation.

The proposal for regulation on digital services (Digital Services Act, DSA) defines the set of entities it will apply to. We consider the **requirements for online platforms and very large online platforms** to be key in connection with disseminating disinformation. The vast majority of the requirements in the proposed digital services regulation concern illegal content, a category which disinformation does not always fall into. One of the few exceptions is the institution of risk assessment and the adoption of follow-up measures for very large online platforms. We also consider the provisions concerning the performance of external audits, the transparency of advertising and recommender systems and the obligation to report on transparency to be interesting. The system of liability for third-party content will not change fundamentally for online platforms, but the modified content moderation and the exact process and rights of users regarding the removal or blocking of illegal content by online platforms are stipulated in more detail.

If we take into account the model of operation of online platforms, the methods of recommendation from a technical point of view and the specifics of the examined proposals for European Commission legislation, we believe that the **examined proposed legislative acts can be improved in such a way as to provide more effective tools to tackle dissemination of disinformation**, which we have addressed in Part 4 (*Discussion and proposed solutions*).

In terms of general comments, **we consider it key for the area of attention economy to be considered as an area of high-risk AI systems**, so that the use of algorithms by social media does not escape the legislative requirements presented in the proposal for regulation on artificial intelligence.

Simultaneously, we understand the limits of the regulation of harmful content in terms of interference with freedom of expression and granting too much power to online platforms. On the other hand, legislation can provide a broad palette of tools that have the potential to limit the spread of disinformation on online platforms without directly regulating harmful content. As examples of the introduction of such 'indirect' rules, we recommend:

- Transparency requirements and user choices in recommender systems;
- Labelling on unverified or unverifiable content (*content labelling*);
- Prohibition on promoting certain content or topics;
- Restrictions on the use of certain methods for sensitive content (such as bots or micro-targeting in political ads).

We consider the performance of external audits absolutely essential. Although the DSA directly arranges such a mechanism, we are concerned about its limits regarding the lack of clear rules for identifying suitable entities for carrying out those external audits, the

insufficient binding nature of the results of such audits and the implementation of their conclusions. At the same time, the access of external auditors should not be restricted in terms of protecting the rights of online platforms, such as trade secrets, or of third parties in the form of personal data protection. In technical terms, we recommend a 'sock puppet audit' as a suitable form of audit, but the above comments need to be incorporated in such a way that the use of bots is possible for external audit purposes in terms of legislation and the terms of service of online platforms. We also emphasise the role of scientific research and the importance of access to data by vetted researchers.

Another important area in tackling disinformation is transparency. In this regard, we welcome the proposals presented in the proposal for regulation on digital services concerning the transparency of advertising and recommender systems and publicly available reports. However, these mechanisms must ensure that their results are meaningful and illustrate the real situation for the professional and lay public. **Regarding the transparency of recommender systems, we focus our attention to the fact that the DSA should enshrine a mandatory opt-in for users on first contact with the online platform in any form.** For example, recommendations could consist of two levels: non-personalised and personalised, while the user could, for example, turn on only the non-personalised recommendations, that is, the one that does not take the user's behaviour (i.e., the source for estimating preferences) into account in the recommendation. At the same time, the obligation for keeping of logs for recommender systems should be stipulated in the DSA for cases potentially not covered by the AIA.

Transparency reports should be complemented by case studies to clarify how the online platform behaves in specific situations and what mitigation measures are being taken. Simultaneously, the statistical indicators in these reports should be divided by Member States so it is possible to detect any differences in the dissemination of disinformation content between Member States, and in the approach of platforms and Member States' authorities and the measures taken.

It is no less important for social media to regularly evaluate not only the legal, but also the ethical and societal risks. We consider it essential for platforms to be able to devote time and resources to the continuous evaluation of possible impacts in terms of the moral values and principles involved throughout the cycle, from the design of new functionalities to their deployment. **It is our belief that ethics risk assessment should be considered a binding part of the conformity assessment for AI system providers proposed in the European Commission's artificial intelligence regulation.**

CONTENT

List of abbreviations

AI means artificial intelligence.

AIA means the proposal for Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts.

Charter of Fundamental Rights means the European Union Charter of Fundamental Rights.

Constitution of the SR means 460/1992 Coll. the Constitution of the Slovak Republic.

Convention means the Convention for the Protection of Human Rights and Fundamental Freedoms.

Directive on electronic commerce means Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce).

DSA means the proposal for Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC.

EC means the European Commission.

EU means the European Union.

GDPR means Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation.

1. Introduction

In the past, dissemination of information to the public by the mass media was the domain of newspapers and television broadcasting. However, the situation has changed radically in recent times with the advent of new technologies, in particular the internet. Disseminating any information is extremely easy today, especially on popular social media, with their millions of registered users, becoming the main platform. Today, anyone can set up a blog or fan page on a social network and present his or her contribution to the general public. New technologies have a positive effect on the dissemination of information not only in the form of greater availability of information, but also by highlighting one of the current problems in the information ecosystem – disinformation. Disinformation is very attractive content since it provides apparently simple answers to complex questions. It often operates in the context of a specific story that interests people.[2]

Dissemination of disinformation poses a hybrid threat, which leads to the polarisation of society and can even lead to the internal destabilisation of democratic institutions.[3] The ongoing COVID-19 pandemic also shows that confidence in the disinformation narratives can lead to serious injury or loss of life. According to the GLOBSEC survey, which analysed the attitude of the public in Central and Eastern European countries to disinformation, up to 56% of people in Slovakia trust disinformation content.[4]

Social media and their activities represent a significant variable in the dissemination of disinformation, since they have tools that can be used to inform a large parts of the public. Disinformation can easily reach susceptible population that do not have a clear position regarding a particular topic. A survey by the Slovak Ministry of Defence has shown that the content of disinformation websites reaches users to a very great extent through the Facebook social network.[5] The problem is also emphasised by new techniques and methods for disseminating disinformation, such as setting up troll farms, Potemkin personas,[6] using artificial intelligence, 'deepfakes', moving to private sections of social media and spreading

[2] For details see, for example, the European Parliament. Disinformation and propaganda: impact on the functioning of the rule of law and democratic processes in the EU and its Member States – 2021 update. Available at: https://www.europarl.europa.eu/thinktank/en/document.html?reference=EXPO_STU(2021)653633.

[3] Ministry of Defence of the Slovak Republic. Cena za (dez)informácie. Výnosy z reklám webov s problematickým obsahom. [The price of (dis)information. Revenue from advertisements for websites with problematic content.] Available at: https://www.mosr.sk/data/files/4364_2021-k-01-cena-za-dezinformacie-recenzovane.pdf.

[4] GLOBSEC. Voices of Central and Eastern Europe, 2020. Available at: https://www.globsec.org/publications/voices-of-central-and-eastern-europe/.

[5] Ministry of Defence of the Slovak Republic. Cena za (dez)informácie. Výnosy z reklám webov s problematickým obsahom. [The price of (dis)information. Revenue from advertisements for websites with problematic content.] Available at: https://www.mosr.sk/data/files/4364_2021-k-01-cena-za-dezinformacie-recenzovane.pdf.

[6] Entities that look like credible institutions (such as non-profit organisations) with a large base across the internet. Simultaneously, however, they insert disinformation into true content. See e.g.: DIRESTA, R. - GROSSMAN, S. Potemkin Pages & Personas: Assessing GRU Online Operations, 2014-2019. White paper, Stanford Internet Observatory Cyber Policy Center, Stanford, 2019.

audio messages.[7] On the other hand, artificial intelligence systems can also be used as a tool to tackle disinformation.[8]

The role of online platforms and in particular social media is significant in disseminating disinformation. On the one hand, they represent an ideal space for the dissemination of any information and contribute to fulfilling the ideals of the information society. On the other hand, through their activities they contribute, whether intentionally or unintentionally,[9] to its dissemination by favouring content that has a potentially greater impact (as a rule, more shocking, controversial or emotional content). The parameter settings of the artificial intelligence systems used to personalise the content displayed to individual users also contribute to this issue. Simultaneously, both the professional and the lay public often criticise the inflexibility of social media concerning moderating content and the removal of / failure to remove problematic content. Regulation of social media in the form of measures to tackle disinformation is a topical issue for discussion, and given the global nature of the issue, it is natural that the European Union wants to take a leading role globally in this area of regulation.

However, several countries in the EU have already decided to regulate the dissemination of disinformation in their national legal systems. In 2018, France introduced an act to combat the manipulation of information.[10] This act allows a French court to issue an order to remove content that contains inaccurate or misleading information disseminated on a large scale intentionally, artificially or automatically in order to affect the credibility of an election in the three months prior to it. Simultaneously, the dissemination of fake news in connection with influencing election results is a criminal offence in France.[11] Malta's criminal code also considers dissemination of fake news to be a crime.[12] Similar prohibitions are contained in Lithuanian legislation or the Greek criminal code[13] which directly prohibits the dissemination of disinformation,[14] The criminal consequences of the dissemination of fake news are also set out in the criminal codes of Austria[15] and Poland.[16] Several countries have taken similar measures preventively in the context of the COVID-19 pandemic. Here we can mention

[7] European Parliament. Disinformation and propaganda: impact on the functioning of the rule of law and democratic processes in the EU and its Member States – 2021 update. Available at: https://www.europarl.europa.eu/thinktank/en/document.html?reference=EXPO_STU(2021)653633.

[8] For details see e.g.: GUO, B. - DING, Y. - YAO, L. - LIANG, Y. - YU, Z. The Future of False Information Detection on Social Media: New Perspectives and Trends. ACM Computing Surveys, 53(4), 1–36, 2020. Available at: https://doi.org/10.1145/3393880; or results of projects financed from EU sources, for example WeVerify, Provenance, or EDMO.

[9] *The Wall Street Journal*. The Facebook Files. Available at: https://www.wsj.com/articles/the-facebook-files-11631713039.

[10] Loi 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information [Law 2018-1202 of 22 December 2018 on the fight against the manipulation of information], JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE [J.O.] [OFFICIAL GAZETTE OF FRANCE]. Cited from HOBOKEN VAN, J. - Ó FATHAIGH, R. Regulating Disinformation in Europe: Implications for Speech and Privacy. UC Irvine Journal of International, Transnational, and Comparative Law. Volume 6 Symposium: The Transnational Legal Ordering of Privacy and Speech. Article 3, 2021.

[11] Code électoral [Electoral Code] art. L97. Cited from HOBOKEN VAN, J. - Ó FATHAIGH, R. Regulating Disinformation in Europe: Implications for Speech and Privacy. UC Irvine Journal of International, Transnational, and Comparative Law. Volume 6 Symposium: The Transnational Legal Ordering of Privacy and Speech. Article 3, 2021.

[12] CRIMINAL CODE, § 82 (amended by the Media and Defamation Act, 2018).

[13] POINIKOS KODIKAS [P.K] [CRIMINAL CODE] 4619:191.

[14] Law on the Provision of Information to the Public, No I-1418, Section 2(13) (1996), amended by No XII-2239 of 23 Dec. 2015.

[15] Strafgesetzbuch [StGB] [Penal Code] Section 264 [Verbreitung falscher Nachrichten bei einer Wahl oder Volksabstimmung] [Dissemination of false news in an election or referendum].

[16] Ordynacja wyborcza do rad gmin, rad powiatów i sejmików województw [Law of 16 July 1998 on Elections to Municipalities, District Councils and Regional Assemblies] (Dz.U. 1998 Nr 95 poz. 602), Section 72.

Romania[17] and Hungary.[18] Currently, Slovakia is also proposing the revision of Criminal Code to directly punish spread of disinformation.

Despite the fact that, even in the EU legal environment, there are laws that punish the dissemination of disinformation in various ways, it is necessary to emphasise that European legal culture is founded on strong respect for fundamental human rights and freedoms. The fight against dissemination of disinformation significantly impacts freedom of expression and the right to information. As international documents on the issue conclude, general prohibitions on the dissemination of information should not be based on vague concepts and definitions such as fake news or disinformation.[19] The United Nations Special Rapporteur expressed a similar view when he warned against the powers of public authorities to delete or block content according to their own conception of truth.[20] This approach is also confirmed by case-law of the European Court of Human Rights, where, in the context of dissemination of disinformation, the court confirmed the legality of the restriction of rights to information only in very obvious and confirmed cases, such as denying the holocaust.[21] For these reasons, the approach of the European Commission to regulation of the fight against disinformation is very circumspect.

Given the above reasons and the global nature of the issue, in the present study we focus mainly on legislation tackling disinformation at EU level.

The primary objectives of the study are to:

a) analyse technological solutions of online platforms in relation to the spread of disinformation and
b) subsequently assess the effectiveness of selected institutions of selected legal arrangements proposed by the European Commission, in particular the proposal for regulation on artificial intelligence[22] and the proposal for regulation on digital services.[23]

The above objectives are reflected in the research question:

[17] Decret semnat de Președintele României, domnul Klaus Iohannis, privind instituirea stării de urgență pe teritoriul României, 16 martie 2020 [Decree Signed by the President of Romania, Mr. Klaus Iohannis, Regarding the Establishment of the State of Emergency on the Romanian Territory, Mar. 16, 2020] Section 54.
[18] 2020. évi XII (Act XII of 2020 on the containment of coronavirus), Section 10(2), amending the Criminal Code, Section 337.
[19] Joint Declaration on Freedom of Expression and 'Fake News', Disinformation and Propaganda, FOM.GAL/3/17, (Mar. 3, 2017). Available at: https://www.osce.org/files/f/documents/6/8/302796.pdf.
[20] KAYE, D (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression). Disease Pandemics and the Freedom of Opinion and Expression, U.N. Doc. A/HRC/44/49 (Apr. 23, 2020), p. 13.
[21] Ruling of the European Court of Human Rights in the case of *Garaudy* v *France*. Application No 65831/01, 24 June 2003.
[22] Proposal for Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. Hereinafter we will refer to it as the draft artificial intelligence regulation or by the abbreviation AIA. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.
[23] Proposal for Regulation of the European Parliament and of the Council on A Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. Hereinafter we will refer to it as the draft services regulation or by the abbreviation DSA. Available at: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-parliament-and-council-single-market-digital-services-digital-services.

'Are the proposed regulations an adequately effective tool for combating disinformation in the context of the technological solutions that online platforms use?'

Despite the fact that we have used a rigorous and tailored scientific approach to maximise the validity and reliability of the findings during the elaboration of the present study, we recognise its limits. These limits are chiefly: minimum transparency and a lack of information on the technical functioning of online platforms, the timing for elaboration of the study (October and November 2021) and the timelines of the regulatory proposals and existing legislation; and the focus specifically on issues of the regulation of online platforms in the form of the proposed digital services regulation and the proposed artificial intelligence regulation.

With regard to the fulfilment of objectives and the answer to the research question, the study is divided into three core parts. Part 2 of the study is devoted to the concept of disinformation and explaining the operation of online platforms in terms of the dissemination of disinformation. We use the term online platform in accordance with the proposed legal definition in the proposal for regulation on digital services. We pay particular attention to models of online platform operation and methods of recommendation from a technical point of view.

Part 3 focuses on legislation and legal instruments tackling disinformation. We begin with an excursion into fundamental human rights and freedoms that may be endangered by the spread of disinformation. Subsequently, we look in detail at the specific institutions of selected regulations proposed by the European Commission, in particular the proposed artificial intelligence regulation and the proposed digital services regulation.

Part 4 presents a discussion and proposed measures based on an analysis of the technical operation of online platforms and deficiencies of the proposed regulations.

2. Disinformation and its dissemination on online platforms

Disinformation represents one of the biggest challenges for current initiatives concerning the development of rules to regulate content and liability on the internet. Current evidence is not just limited to the COVID-19 pandemic and dissemination of information on vaccination. The EU as a 'regulatory superpower' with a significant impact on the legal systems of third countries[24] has taken the lead in regulatory instruments that should contribute to combating disinformation.

There is no legal notion of 'disinformation' in the EU legal system. The definition of the given term may be derived from non-binding parts of legal acts, preambles, studies, strategies and recommendations.

The Communication from the European Commission on a European approach to tackling online disinformation states that:

> 'Disinformation is understood as verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm. Public harm comprises threats to democratic political and policy-making processes as well as public goods such as the protection of EU citizens' health, the environment or security. Disinformation does not include reporting errors, satire and parody, or clearly identified partisan news and commentary.'[25]

This definition is also used by the European Commission Action Plan against Disinformation.[26]

The High-level Group on fake news and online disinformation defines the term disinformation similarly. The Group considers disinformation to be false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit.[27]

[24] BRADFORD, A. The Brussels Effect. How the European Union Rules the World. Oxford University Press Inc, 2020.
[25] European Commission. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Tackling online disinformation: a European Approach COM/2018/236 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236.
[26] European Commission. JOINT COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Action Plan against Disinformation JOIN/2018/36 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018JC0036.
[27] A multi-dimensional approach to disinformation. Report of the independent High level Group on fake news and online disinformation, 2018, p. 10. Original: '…false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit.' Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271.

The Code of Practice on Disinformation defines disinformation positively and negatively. From a positive point of view the definition of disinformation follows the above definition:[28]

> 'verifiably false or misleading information' which, cumulatively,
> (a) 'Is created, presented and disseminated for economic gain or to intentionally deceive the public'; and
> (b) 'May cause public harm', intended as 'threats to democratic political and policymaking processes as well as public goods such as the protection of EU citizens' health, the environment or security.'

The negative definition of disinformation excludes from this concept 'misleading advertising, reporting errors, satire and parody, or clearly identified partisan news and commentary, and is without prejudice to binding legal obligations, self-regulatory advertising codes, and standards regarding misleading advertising.'

In its recital section, the DSA (proposal for regulation on digital services) separates the concepts of disinformation from illegal content[29] or other manipulative and unfair activities.[30]

At the level of the Slovak Republic, similarly to EU law, there is no legal definition of disinformation. However, the interpretation of that term may be based on strategic documents of the Slovak Republic. The coordinated resilience mechanism of the Slovak Republic against information operations includes the following definition:

> 'Disinformation: false or manipulated information that is spread deliberately in order to mislead and cause harm. Disinformation can take the form of false or manipulated text, images, video or sound, and can be used to support conspiracies, disseminate doubt and discredit truths or individuals and organisations. We can also consider true information as disinformation if it is imparted in a manipulative manner. Disinformation does not include unintentional reporting errors, satire and parody, or clearly identified partisan news and commentary.'[31]

The Security Strategy of the Slovak Republic[32] or the Concept of the Slovak Republic for the fight against hybrid threats[33] does explicitly mention dissemination of disinformation in several parts but does not define the term itself.

[28] Code of Practice on Disinformation. 2018. Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=59125.
[29] DSA, Recital 71.
[30] DSA, Recital 68.
[31] Coordinated resilience mechanism of the Slovak Republic against information operations. Available at slov-lex.sk.
[32] Security Strategy of the Slovak Republic. Available at: https://www.nbu.gov.sk/wp-content/uploads/urad/Bezpecnostna-strategia-SR-2021.pdf.
[33] Concept of the Slovak Republic for the fight against hybrid threats. Available at: https://www.nbu.gov.sk/wp-content/uploads/PHHD/Koncepcia-boja-SR-proti-hybridnym-hrozbam.pdf.

Criminal Act of the Slovak Republic

Interestingly, the Slovak Criminal Act contains several terms that are related to disinformation, but it does not define them directly.

The Criminal Act regulates the factual basis of the crime of endangering the safety of an aircraft or ship, within the meaning of which 'a person who reports **false information** which **may threaten the safety or operation of an aircraft in flight or ship under sail** shall be punished with imprisonment of up to three years.'

The term 'false information' is related to the crime of terrorist attack: 'anyone who, with the intention of damaging the constitutional establishment or defence readiness of the state, disrupting or destroying the fundamental political, economic or social structure of the state or international organisation, terrorising the population or forcing the government of the state or other public authority or the international organisation to perform, omit or suffer something..., seizes control of an aircraft, ship, other means of passenger transport or... **reports false information**, thereby jeopardising life or human health, or the safety of such a means of vehicle.'

Another example is the crime of spreading alarming news: 'a person who **intentionally causes serious disorder among at least part of the population of a locality** by **spreading alarming news that is false** or commits another similar act capable of inducing such danger shall be punished with imprisonment of up to two years.'

The term 'false information' is related to the crime of defamation: 'a person who **declares false information about another** that is capable of considerably damaging his or her dignity among fellow citizens, harming him or her in employment or in business, disrupting his or her family relationships or causing him or her other serious harm shall be punished with imprisonment of up to two years.'

Legal literature perceives disinformation from three perspectives:

- content;
- participants; and
- distribution.

From the perspective of content, this concerns a classification of information that is false or misleading. Just this aspect is significantly subject to regulation. The perspective of participants reflects disinformation from the perspective of the entity that disseminates it, such as third countries with the intention of destabilising the democratic system. The third perspective on disinformation concerns the dissemination and various techniques of multiplying the effect of disinformation dissemination.[34]

[34] Cited from HOBOKEN VAN, J. - Ó FATHAIGH, R. Regulating Disinformation in Europe: Implications for Speech and Privacy. UC Irvine Journal of International, Transnational, and Comparative Law. Volume 6 Symposium: The Transnational Legal Ordering of Privacy and Speech. Article 3, 2021, p. 13.

In addition to the term 'disinformation', other related (and sometimes partially overlapping) terms are often used in the literature, such as false information, disinformation, fake news, hoaxes, and similar (for more details on the relationships between these concepts, see e.g., Hrčková et al. (2019)[35]). Gelfert (2018)[36] considers fake news as a type of disinformation and defines it as 'the deliberate presentation of (typically) false or misleading claims *as news*, where the claims are misleading *by design*.' The term began to be used extensively after 2016 in connection with the referendum on the UK's withdrawal from the EU and the US presidential election, when both campaigns were accompanied by a great deal of disinformation in the form of false reports, often spread on major online platforms (Twitter, Facebook).[37,38]

Currently, the use of the term 'fake news' is being abandoned; as the abovementioned High level Group on fake news and online disinformation states, this is due to its inaccuracy and incompleteness, as well as its misuse to discredit political opponents; instead, it favours the term 'disinformation'.[39] For similar reasons, the term 'fake news' is also avoided by a report prepared for the Council of Europe, which distinguishes between *misinformation* (false information that is shared without intent to harm), *disinformation* (false information intended to cause harm) and *malinformation* (true information that is shared with the intention of harm, such as the disclosure of private or secret information, but also hate speech).[40]

A distinction between misinformation and disinformation based on the intention of the person sharing the information is also present in the computer science literature (see, for example, Kumar and Shah (2018)[41]). In practice, however, it is often difficult to identify the intention and attempts to moderate content or detect misinformation or disinformation using automated methods often confuse them.

2.1 Model of operation of online platforms

Online platforms and, above all, very large online platforms (for their definition and distinction according to the DSA, see Section 3.2.1) have, during their relatively short existence, changed how people obtain information (in what form, from what sources, or how they handle it further). This is a global phenomenon; for illustration, Facebook, established in 2004, currently has more than 2.89 billion active users worldwide, YouTube (established in 2005, purchased by Google in 2006) has more than 2.29 billion users, and WhatsApp (established in 2009,

[35] HRČKOVÁ, A. - SRBA, I. - MÓRO, R. - BLAHO, R. - ŠIMKO, J. - NÁVRAT, P. - BIELIKOVÁ, M. Unravelling the basic concepts and intents of misbehavior in post-truth society. Bibliotecas. Anales de Investigación; 15(3), 2019, pp. 421-428. Available at: http://revistas.bnjm.cu/index.php/BAI/article/view/109/110.

[36] GELFERT, A. Fake News: A Definition. Informal Logic, Vol. 38, No 1, 2018, pp. 84-117.

[37] BASTOS, M. T. - MERCEA, D. The Brexit Botnet and User-Generated Hyperpartisan News, Social Science Computer Review, 37(1), 2019, pp. 38–54. Available at: https://doi.org/10.1177/0894439317734157.

[38] BOVET, A - MAKSE, H.A. Influence of fake news in Twitter during the 2016 US presidential election. Nat Commun 10, 7 (2019). Available at: https://doi.org/10.1038/s41467-018-07761-2.

[39] A multi-dimensional approach to disinformation. Report of the independent High level Group on fake news and online disinformation, 2018. Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271.

[40] WARDLE, C. - DERAKHSHAN, H. Information Disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe report DGI, 2017, 09. Available at: https://firstdraftnews.org/articles/coe-report/.

[41] KUMAR, S. - SHAH, N. 2018. False information on web and social media: A survey. arXiv preprint. Available at: arXiv:1804.08559.

purchased by Facebook in 2014) is the third most used online platform worldwide with 2 billion users (see Table 1 for statistics on other widespread online platforms).

The rise of online platforms, above all social media, has also fundamentally changed people's news access. According to a survey by the Reuters Institute,[42] as early as 2012, in many countries people consumed news online (e.g., 86% of respondents in the United States, 82% in the UK, 77% in France and 61% in Germany), but mostly directly from websites of newspapers or (television) broadcasters; 36% of respondents reported social media and blogs as a news source in the United States, but only 18% in Germany and the United Kingdom and 17% in France.

Table 1

Statistics on the use of (very large) online platforms. The online platforms are ranked according to the number of users worldwide, while only platforms that have spread into Slovakia are included (for example, WeChat in 6th place, Sina Weibo in 10th place, etc. are not included). Global user data show the number of active *platform users as of October 2021, with the exception of the WhatsApp and Facebook Messenger platforms, since these platforms have not published updated user figures for the last 12 months. Data on the potential size of the audience in Slovakia (as of January 2021) express the number of people who may be reached by advertising on the given platform and therefore the actual numbers of users may differ from these values. The authors of this study were unable to acquire relevant data for the potential size of the audience in Slovakia for the WhatsApp, TikTok and Telegram platforms.*

Data source: Wikipedia (year of establishment), Statista.com[43] (number of users worldwide), Digital 2021: Slovakia[44] (potential size of audience in Slovakia).

#	Online platform	Year of establishment	No of users worldwide	Potential size of audience in Slovakia
1.	Facebook	2004	2.89 billion	2.70 million
2.	YouTube	2005	2.29 billion	4.03 million
3.	WhatsApp	2009	2.00 billion	-
4.	Instagram	2010	1.39 billion	1.40 million
5.	Facebook Messenger	2011	1.30 billion	2.40 million
7.	TikTok	2016	1.00 billion	-
11.	Telegram	2013	550 million	-
12.	Snapchat	2011	538 million	465 000
15.	Twitter	2006	463 million	140 000

[42] NEWMAN, N. Reuters Institute Digital News Report 2012: Tracking the Future of News. Available at: https://www.digitalnewsreport.org/survey/2012/.

[43] Statista.com. Available at: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[44] Digital 2021. Slovakia. Report. Available at: https://datareportal.com/reports/digital-2021-slovakia.

As opposed to this, in 2016, in a sample of 26 countries, more than half of all respondents (51%; in the EU on average 46%) stated that they use social media as a news source every week and for 12% they were the main source.[45]

In 2021,[46] 66% of respondents from 12 countries[47] had used one or more social media and instant messaging applications for the purposes of consuming, sharing and discussing news. Throughout that time, Facebook was the dominant platform, but in recent years it has weakened in favour of other emerging social media and especially applications for instant messaging (WhatsApp, Telegram). Instant messaging applications are used as a news source mainly in Latin America (40% use WhatsApp) and Africa (61% use WhatsApp and 18% Telegram).[48] Data for Slovakia are given in Table 2.

Table 2

The most used online platforms for news in Slovakia. A total of 56% of respondents use social media as a news source in Slovakia according to the Reuters Institute study of 2021.

Source: Reuters Institute, Digital News Report 2021.

Online platform	Share of respondents using a platform for news	Share of respondents using a platform for any purpose
Facebook	55%	72%
YouTube	26%	77%
Facebook Messenger	19%	51%
Instagram	11%	29%
WhatsApp	9%	29%
Viber	5%	20%

Online platforms are often associated with dissemination of disinformation. In general, 58% of respondents (in European countries 54%) considered disinformation to be a problem. Specifically in relation to disinformation about COVID-19, most respondents were concerned in connection with information on Facebook (28%), on the pages and in the applications of newspapers (17%), in instant messaging applications (15%), in search results (7%) or on YouTube (6%).[49]

[45] NEWMAN, N. et al. Reuters Institute Digital News Report 2016. Available at: https://www.digitalnewsreport.org/survey/2016/.
[46] NEWMAN, N. et al. Reuters Institute Digital News Report 2021. Available at: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021.
[47] Ibid. The authors of the Reuters Institute chose 12 countries (the UK, the United States, Germany, France, Spain, Italy, Ireland, Denmark, Finland, Japan, Australia and Brazil) for which they had data available from 2014 or 2015, so they could compare changes in that period.
[48] Ibid.
[49] Ibid.

Given the fundamental importance of impartial news and free access to information for society and democracy, the statistics above show why so much attention is devoted to online platforms at a time when we are witnessing a growing polarisation in society and the dissemination of disinformation.

The problem remains that the **mission of online platforms is not balance of views and to objectively inform their users**.

Most of their revenue (USD 85.9 billion in 2020 in the case of Facebook[50], USD 3.7 billion in the case of Twitter, of which USD 3.2 billion from advertising,[51] and USD 1.9 billion in the case of TikTok[52]) comes from the **sale of online advertising**. The more time users spend on an online platform, the more adverts the platform can potentially show them.

Very large online platforms compete for the **attention** of their users, which is a limited resource. According to the Encyclopaedia Britannica, in psychology, attention is defined as 'the concentration of awareness on some phenomenon to the exclusion of other stimuli'.[53] In regard to the increasing importance of attention in economic relationships in society, Goldhaber and, after him, others talk about **attention economy**[54] - although our age is sometimes termed the age of information economics, according to Goldhaber this is incorrect, since in the economy what is important is how society divides up limited resources, which information is not, but attention is.

Online platforms therefore try to gain as much as possible of this limited resource – the attention of their users – and convert it into their revenue in the form of online advertising.

They use three approaches for this:

1. Providing **attractive user interfaces with habit-forming elements** using **persuasion techniques** based on knowledge from behavioural psychology, which includes, for example, push notifications (of a new post by a friend, of acceptance of friendship, etc.), information on a message being read by a recipient, infinite scrolling and the like.[55]

2. Providing **interesting, engaging content** – there is more content on online platforms than can be consumed in a limited (finite) time period. Therefore, it is necessary to filter the information shown using **personalised recommendation**, i.e., information (posts on the online platform) is listed according to its expected relevance/interest

[50] Facebook Revenue and Usage Statistics (2021). Available at: https://www.businessofapps.com/data/facebook-statistics/.
[51] Twitter Revenue and Usage Statistics (2021). Available at: https://www.businessofapps.com/data/twitter-statistics/.
[52] TikTok Revenue and Usage Statistics (2021). Available at: https://www.businessofapps.com/data/tik-tok-statistics/.
[53] MCCALLUM, W.C. 'Attention'. Encyclopaedia Britannica, 9 June 2015, Available at: https://www.britannica.com/science/attention
[54] GOLDHABER, M.H. Attention Shoppers! 1997, WIRED. Available at: https://www.wired.com/1997/12/es-attention/.
[55] See e.g. the article on Tristan Harris, former employee of Google responsible for product philosophy, BOSKER, B. The Binge Breaker. The Atlantic, November 2016. Available at: https://www.theatlantic.com/magazine/archive/2016/11/the-binge-breaker/501122/.

to the specific user. The online platform selects (in the case of social media) posts from users/channels that the user follows (which they subscribe to, like, etc.), but also posts from other users/channels that it thinks could be of interest to the user. The order of posts is also influenced by their sponsorship or the inclusion of paid advertising.

3. Providing **relevant adverts** – the goal of the online platform is not just *impressions*, i.e., displays of an advert, but since the user's attention is limited, they try to select those where the likelihood of a user response is higher. Therefore, the platforms use **personalised recommendation** methods also for online advertising.

The harmfulness of the mechanisms on which the attention economy is based have been a subject of debate for several years; for example, the effect on the ability of people to concentrate and do deep work,[56] but also on the psyche of children and adolescents.[57] Despite this, it does not seem that a major change in this regard will occur in the near future, although some platforms (e.g. Spotify or YouTube Music) allow users to decide whether to 'pay' for their service with their attention by watching/listening to ads or pay a premium subscription.

Simultaneously, there is debate about the **risks of online platforms in connection with the extensive collection of user data** which takes place for the purposes of personalising recommendations. This does not mean just recording what posts the user clicked on or interacted with by sharing, commenting, etc., but often how long the user looked at a particular post (to which the interface is adapted so that just one post is visible for simpler measurement), what other pages the user visited on the website (using cookies and website plugins, such as the 'like' button), data from the device on which the user accesses the platform, etc.

According to Shoshana Zuboff, this collection is the basis of **surveillance capitalism**, in which the traditional relationships of industrial capitalism are replaced or supplemented by new practices based on the unrestricted 'mining' of behavioural data and the consequent monetisation of user behaviour prediction and manipulation of their behaviour.[58]

Attention economy may also operate outside online platforms (after all, when Goldhaber wrote the article about it in 1997, the very large platforms and social media we know today did not exist), but these seem to significantly streamline the sharing of the limited resource of user attention. Conversely, the surveillance capitalism referred to by Zuboff presupposes the concentration of a large amount of data on a large number of users in the hands of a few large companies, which in some places enjoy a market monopoly. She does not offer solutions to

[56] Ibid.

[57] According to recently leaked internal documents from Facebook (the 'Facebook Files'), the company has known for some time about the negative impact of Instagram on the mental health of adolescent girls, for example, in the form of an increased risk of eating disorders, depression, etc. WELLS, G. - HORWITZ, J. - SEETHARAMAN, D. Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show. *The Wall Street Journal*, 2021. Available at: https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739.

[58] ZUBOFF, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* PublicAffairs, 1st issue, 2019.

the attention economy but focuses on steps to reduce the impact of very large online platforms, e.g., with regulations.

The risks from the collection of large amounts of personal data are already comparatively well-known and regulated within the EU (by the GDPR). However, the recommendation methods that online platforms use to gain users' attention (and for which they collect user data), and the benefits and potential risks arising from them are much less known outside the professional community.

2.2 Recommendation methods

Recommender systems were designed in response to the information overload of users, especially in the corporate environment in the late 1980s in order to reduce the amount of information. As technology gradually developed, the research went through three developmental phases.[59]

In the first developmental phase the user received only part of the arriving information: in an analogy from the corporate environment, a portion of the incoming emails, e.g., 20%, would be redirected to them. Even though the amount of information is reduced in this way, a lot is still irrelevant for the user. In the next developmental phase, a filter was deployed, which filtered the output from the first phase by keywords and removed some of the irrelevant information (e.g., spam or communication irrelevant to the company's business). The last developmental phase represents the personalised recommendations themselves, which recommend only information relevant to a specific user from the whole quantity. Again, in the analogy from the corporate environment: users only receive emails that are relevant to their position and to them as a person.

In general, we define recommender systems as software tools that generate item proposals for a specific user.[60]

With the increase and widespread of user-generated content (e.g., social networks, blogs), recommender systems have also fluidly spread into non-corporate applications – in particular online services – and today recommender systems are practically the gold standard for user interaction with web applications. With the gradual expansion, the reasons why the recommender systems were deployed have also expanded: from the primary function of reducing the user information overload, to increasing customer loyalty, offering diverse content and increasing product sales.

[59] GOLDBERG, D.- NICHOLS, D. - OKI, B.M. - TERRY, D. Using collaborative filtering to weave an information tapestry. Commun. ACM 35, 12, December 1992, pp. 61-70. Available at: https://doi.org/10.1145/138859.138867.

[60] RICCI, F. - ROKACH, L - SHAPIRA, B. Introduction to recommender systems handbook. In Recommender systems handbook, Springer, 2011, pp. 1-35. Available at: https://doi.org/10.1007/978-0-387-85820-3_1.

Regardless of the type of recommendation and the method used for generating item proposals, there is a necessary condition that must be met – the recommendation method must make use of information on user preferences. When interacting with the system or web application, the user leaves traces" feedback") in various forms (e.g., views, ratings, comments and purchases). This feedback is used as an indicator of user preference (e.g., product evaluation).

2.2.1 Personalisation and adaptation

Looking at the recommendation methods, we can distinguish two basic concepts:

- *Personalised recommendations* – suggestions are generated directly for a specific user based on his or her previous behaviour (or preferences).
- *Non-personalised recommendations* – generally valid recommendations for users of a service or specific product (for example, a recommended method of food preparation, recommended sequence of educational materials, or the most read reports).

Non-personalised recommendations are found on practically all websites, though users are rarely aware of this. For example, the list of the most frequently visited sites or the most read articles are a form of a non-personalised recommendation.

On the boundary between a personalised and a non-personalised recommendation is *adaptation* – *tailoring* to a specific group of users, for example, showing a tutorial to new users or notifying fans who prefer a specific film genre of a new film. In other words, the term 'adaptation' means tailoring content or an interface. While, when making recommendations to the user, we present a set of proposals or recommendations, in the case of adaptation, the content or interface which the user interacts with is directly modified. For example, the navigation (menu) items are rearranged.

In this study, we will look more closely at personalised recommendations, with regard to their greatest impact on individuals, options and dissemination on online platforms.

2.2.2 User modelling

A necessary condition for generating a personalised recommendation is the knowledge of user preferences (or previous user actions). User models are used for the purposes of storage and modelling user preferences on online platforms. In its basic form, a user model consists of a pair [key, value][61], where the key represents a given preference (e.g., news category – sport) and the value quantifies the user's preference. In other words, a user model characterises a given user at a specific time based on information acquired in a given system (web application).

[61] SENOT, CH. - KOSTADINOV, D. - BOUZID, M. - PICAULT, J. - AGHASARZAN, A. - BERNIER, C. Analysis of strategies for building group profiles. User Modeling, Adaptation, and Personalization, Springer Berlin Heidelberg., 2010, pp. 40-51. Available at: https://doi.org/10.1007/978-3-642-13470-8_6.

Despite many limitations, the behavioural approach is used in online platforms, where we assume that the user's behaviour is conditioned by his or her preferences. Based on this assumption, user modelling is based on monitoring his or her behaviour – the actions the user performed in the system (or did not perform). It is a kind of simplification which allows generation of recommendations even without real knowledge of the given user's preferences.

We differentiate between two basic sources of information:

- *Implicit feedback* – information derived from user actions. For example, based on the number of articles read about sport, we can derive the user's preference for a given category.
- *Explicit feedback* – direct expression of the evaluation for a specific element or preference. Users evaluate the given element (e.g., they evaluate whether they like the article about sport they have just read, or whether it is relevant to them), or explicitly indicate that they want to receive news about sport.

Depending on the complexity of the user model used, the amount and type of information represented in the model grows:

- Demographic Information
- Objectives and tasks
- Interests
- Skills and competences
- Characteristics – personality, cognitive style and similar.
- Mood.

Given the different domains and methods of application, one system can use several user models, which also take the time aspect of the user's preferences into account –- e.g., short-term and long-term. Regardless of the type of user model, modelling is a continuous process made up of three basic steps:

- *Data collection* – evaluations, interaction records, context;
- *Inference (reasoning)* – derivation of knowledge, preferences, goals, etc. based on the information collected;
- *Adaptation or personalisation* – application of a recommendation method based on the user model.

The importance of the user model varies in various application domains. For example, in teaching systems, modelling of the knowledge of specific students is essential. In the case of e-commerce, the basic recommender gets by with information about previous purchases. In general, however, the more complex the user model, the better the quality of the recommendations generated.

2.2.3 Personalised recommendation

Personalised recommendation approaches generally use machine learning methods and statistical approaches. Current research trends are developing machine learning methods further, particularly neural networks.

The basic division of personalised recommendation approaches is based on an analogy from everyday life and the behaviour of people who tend to consider the advice of those around them:

- *Collaborative filtering* – generates recommendations based on similarity of users. It is based on the assumption that if the preferences (or history) of two users coincided or were similar in the past, it is highly likely that their preferences will also coincide or be similar in the future (Figure 1). The core of collaborative filtering is information about user preferences. In order to achieve satisfactory results, it is essential that a large number of users use the system, which increases the likelihood of finding users with similar preferences. The advantage of collaborative filtering is abstraction from the content of recommended items, which we do not need to know anything about (unambiguous identification of elements is sufficient for the methods to work).

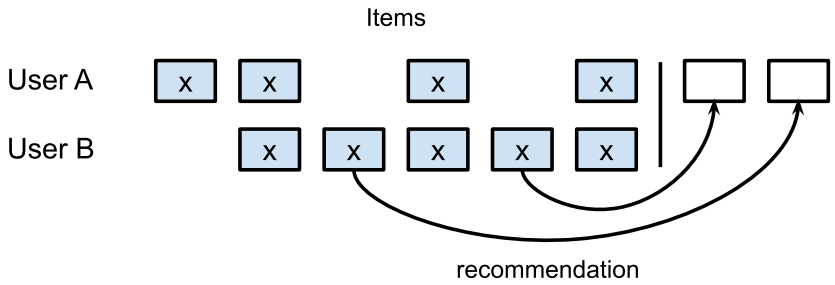

Figure 1
Functional principle of collaborative filtering.
Source: Authors' own work.

- *Content-based recommendation* – generates recommendations based on analysis of the content of items. It recommends items similar to those the user liked in the past (Figure 2). Similarity between items has been historically defined as textual similarity (e.g., news articles), or textual similarity of metadata (director, genre, etc.). These days, no special requirements are placed on content due to the shift in image or sound processing. Content recommendation can also be applied to new domains, while it is sufficient to know the preferences of the given user and process available content for its generation. Therefore, in comparison with collaborative recommendation, there is no need to analyse the behaviour of other users.

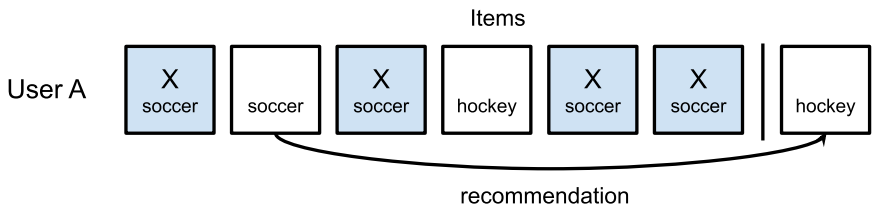

Figure 2
Functional principle of content recommendation.
Source: Authors' own work.

On the other hand, it is clear that, by its essence, content recommendation converges to enclose the user in a filter bubble and subsequently degrades. Therefore, mechanisms ensuring the diversification of the recommended elements must be implemented. Another option is to combine several types of recommendation on a given online platform (this is the gold standard for modern web applications).

- *Hybrid approaches* – based on the assumption that we get 'better' outputs by combining several methods (not necessarily different approaches). Improvements can relate to various aspects (e.g., more precise recommendations, greater diversity). The hybrid approach usually aims to reduce the limitations and problems of the particular methods involved. There are a relatively large number of combinations of two or more methods. The simplest include switching hybrid systems that switch between several recommendation methods based on an explicit rule (e.g., the number of rated items). Another example is the cascade hybrid system, where the first method generates possible recommendations and the second rearranges and modifies this list.

Aside from the basic division described above, we recognise other dimensions, based on which we also differentiate other types of recommenders. If the recommender system also considers contextual information (e.g., user location, the device from which it is accessed, weather or seasons), we refer to this as *context-based recommendation*.

With the introduction of GDPR (and the increase in users for whom we have no information available about their preferences), the *session-based recommendation* is becoming more and more popular. By session, we mean a sequence of actions performed by a user during a single visit to a website or service. With this approach, the methods mainly consider the user's current actions without taking account of historical preferences. In this case, it is possible to generate recommendations even for unidentified users about whom we have no previous information.

2.2.4 Evaluating success

Metrics taken from the information retrieval field and machine learning are typically used to evaluate the success of personalised recommendation methods. It should be noted that these metrics address the quality of the generated recommendations in technical terms, while

ignoring other important aspects that mean better quality, or address business issues related to the deployment of recommenders.

The basic metrics for the characteristic of recommendation 'success' are:

- *Precision* – characterised as the ratio of correctly recommended items to all recommended items. For example, if we recommend a list of 10 items, from which the user marks 2 as relevant, this recommendation will achieve an accuracy of 2/10 = 20%.
- *Recall* – a complementary metric to the precision metric. It expresses the ratio of relevant recommendations to all relevant items for a given user. Therefore, if we know that there are 5 relevant items for a given user on the online platform, and the user identifies 2 of them from the recommendations, the recall will represent 2/5 = 40%.
- *Accuracy* – reflects accurately recommended (relevant) items and accurately not recommended (irrelevant) items. Given the large imbalance between the number of relevant and irrelevant items in real systems (there are fewer relevant than irrelevant items), the metric reports high values even with qualitatively bad recommendations. For example, let have a total of 100 items on the online platform, but only 5 are relevant to the user. We generate 10 recommendations, which will include 1 correctly recommended (relevant) item and 9 incorrectly recommended (irrelevant). Of the remaining 90 items, we accurately did not recommend 86 irrelevant items and inaccurately 4 relevant items. In this case, we achieve an accuracy of (1+86)/100 = 87%, despite the fact that 90% of the recommended items were inaccurate (precision 10%) and we recall 1/5 = 20% of all relevant items on the platform with recommendations.

Since the abovementioned metrics ignore the order of the items (their order is not important), which does not reflect the established user behaviour (users go through the lists from top to bottom by default, and assume that the relevant items are at the top of the ordered list), the next group of metrics also takes the position of the recommendations into account:

- *Normalised Discounted Cumulative Gain (nDCG)* – takes into account the order of the items during recommendation and compares them with the ideal order. The ideal order represents a list in which the most relevant item is in first place, followed by the second most relevant, and so on.
- *Average precision (aP)* – is used to compare two ordered lists of recommendations, while we consider the better recommendation to be one that has generated the relevant items at the top of the list (Figure 3).

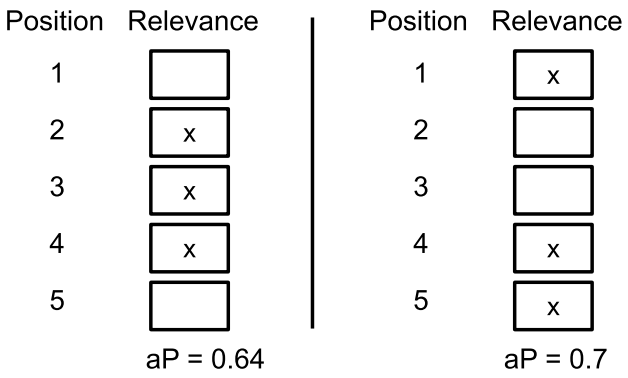

Figure 3
Comparison of two ordered lists of personalised recommendations using the average precision metric – the slightly higher aP rating was caused by the relevant first position (on the right).
Source: Authors' own work.

The third group of metrics are metrics expressing the deviation from the user's rating. If a user rates a recommended item (e.g., the number of stars), these metrics reflect a recommendation estimate error:

- *Average absolute error* – is expressed as the average difference between the forecast and the actual rating by the user. It is clear that the error is defined only for the items that the user has rated, which is why it does not give a complete picture of the success of the system.
- *Mean squared error* – compared to the average absolute error, this takes more into account larger errors in the estimate (thanks to their exponentiation).

Generally, not just one metric is used to evaluate the quality of a recommendation method, but a group of metrics able to describe the system from several perspectives. We can also describe the quality of recommendations as such from the perspective of other 'soft' metrics:

- *Confidence* – the level of a user's confidence in a recommendation method. There is a stronger level of confidence in the recommender when relevant recommendations are generated.
- *Novelty* – represents recommendations that are new to the user. For example, it is unlikely that a user will evaluate a currently mass-promoted blockbuster film as novel.
- *Diversity* – expresses the level of diversity of the generated recommendations. This metric helps detect filter bubbles when the recommender generates thematically identical recommendations.
- *Tail* – this describes the system's ability to recommend items from a 'long tail'. There are a small number of highly requested items and a large number of items that show a low level of interaction (e.g., users do not know these items exist) in the system. The ability to recommend these items often brings with it economic benefits while contributing to greater diversity.

We can also monitor the indirect effects of the recommender system on users (e.g., on user satisfaction with the online platform) depending on the reason for its deployment. Over the long term, we can track the rate of customer loss.

Typical business metrics also directly expressing the economic side of the recommender method include:

- *Turnover* – usually monitored in e-shops. The recommender system can be deployed with the aim of reflecting a selected issue – for example, to engage new users. Over the medium and long term, we can evaluate the contribution of recommendations to increasing turnover.
- *Click-through rate (CTR)* – a standard metric in online platforms, regardless of the domain. It represents the rate of 'clicks' and impressions of, in our case, the recommendation of a given item. It is built on the assumption that a user's click implicitly represents positive feedback.
- *Time spent* – allows evaluation of the user's interest based on the time they spend on a given platform or a specific item. For example, in a news domain, consideration of the time metric can filter out articles that a user has visited, but not read, from those that they have actually read and been interested by. When considering the platform overall, strict optimisation of the time spent by users may cause undesirable behaviour in the model – its optimisation only to this metric.

The chosen metric or group of metrics allows the reduction of possible undesirable effects caused by a machine approach based on the processing of a large amount of data. By monitoring several criteria, it is possible to prevent the generation of uninteresting, overly specialised or unreliable recommendations.

2.2.5 Typical problems of recommender systems

From the perspective of recommendation methods, we encounter typical problems that arise directly from the essence of the recommendation as such, or from the very concept of a given type of recommendation method.

A cold start is a situation where the method is unable to generate recommendations for a user whose preferences are unknown (a new user). Obviously, the same problem will occur with a new item (new article, product) which has not yet been rated, meaning that the system cannot recommend it. Different methods react to new users/items with different sensitivity. While with content recommendation it is possible to generate recommendations after the initial user interaction, with collaborative filtering, the cold start issue persists longer.

Black and grey sheep represent problematic users in terms of the recommendation method. Users who behave differently (or have different preferences) from other users and for whom collaborative recommendation approaches cannot be used are referred to as black sheep. Users whose preferences are constantly changing or do not correlate with any of the user

groups are referred to as grey sheep. It is again challenging to generate recommendations through collaborative approaches for such users.

Over specialisation is the state of a recommendation model which is closely focused on the user's specific preference. For example, only news articles from one subdomain are offered to them and their other preferences are ignored. Thus, the user gets into an information bubble, from which he or she cannot leave by interacting with the recommender system only.

Data sparsity is a problem typical of recommendation methods. With regard to application domains, recommender systems are deployed in systems with many users and many elements needing recommendation. However, from the nature of user behaviour, the system only recognises (or can estimate) user preferences for a very limited set of items. For example, there are thousands of films in a film domain, but one user usually rates a maximum of dozens of them. As a result, the data sparsity normally reaches 99+%.

2.2.6 Examples of use on online platforms

Recommender systems are an integral part of modern web applications. Given the amount of user-generated content, they allow reduction of the information overload and, depending on the application, bring opportunities for both providers and users.

For example, the LinkedIn professional network uses personalised recommendations based on candidate profiles for practically all the products it provides on its online platform – recommendations of contacts, subscription channels, companies, and so on. However, recommendations for the job offer domain are not generated unilaterally, but this is a special case of bilateral recommendation. Like recommendations on dating platforms, it is not enough that the candidate is suitable for a given offer; the offer must also attract a specific candidate.[62] The LinkedIn network model uses two-tier architecture where in the first step the candidates matching the given offer are ordered by relevance and in the second step the *top-N* relevant candidates are reordered on the basis of current dynamic preferences using the multi-armed bandit method.

Figure 4 shows the architecture of the system using the multi-armed bandit method. We can easily explain the multi-armed bandit method with the example of a real-life problem. Imagine a player standing in front of *K* slot machines ('one-armed bandits'). The player is faced with the question of which slot machine to use, while logically trying to optimise winnings. During the game, the player must decide how long to try one slot machine, when to try another, and so on. He or she must find the optimum strategy for exploring other options and testing a particular machine.

In the language of online platforms, the problem is analogous – we have a lot of content (e.g., posts) and we need to ascertain which item to display to a given user who we have no

[62] AI Behind LinkedIn Recruiter Search and Recommendation Systems. Available at: https://engineering.linkedin.com/blog/2019/04/ai-behind-linkedin-recruiter-search-and-recommendation-systems.

information about, in what order and time. This is a complex optimisation problem, since the system only receives information about the 'machines' chosen by the user and nothing about the others.

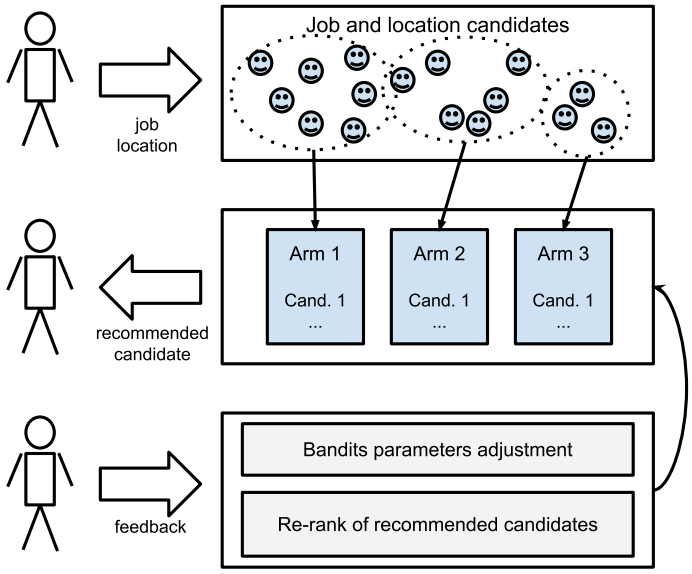

Figure 4
Architecture of personalised recommendation using the multi-armed bandit algorithm – LinkedIn.
Source: Inspired.[63]

Depending on the size of the online platform, the architecture of the recommender system must be designed with regard to the number of elements and users using the system. For example, the YouTube platform uses a two-tier approach (Figure 5).[64]

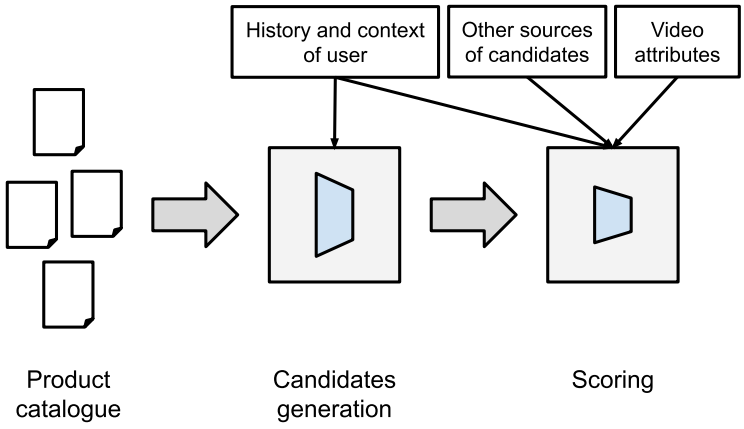

Figure 5
YouTube online platform recommendation model architecture.
Source: Inspired.[65]

[63] Ibid.
[64] COVINGTON, P. - ADAMS, J. - SARGIN, E. Deep Neural Networks for YouTube Recommendations. In Proc. of the 10th ACM Conf. on Recommender Systems (RecSys '16). Association for Computing Machinery, New York, NY, USA, 2016, pp. 191-198. Available at: https://doi.org/10.1145/2959100.2959190.
[65] Ibid.

In the first step, hundreds of candidates are generated from the available items (videos) based on the user's history. In this step, the collaborative filtering approach is used, extended by user demographics or, currently, probably by deep learning. In the second step, the set of generated candidates is ordered based on an evaluation function defined by the platform. In this step, elements from other sources may appear among the generated candidates. Similarly, other aspects of the individual elements are taken into account in this step, e.g., their popularity and development over time, the relevance of the given channel and history of interaction with it, language, and so forth.

The personalised post-sorting algorithm on the Facebook online platform[66] uses a similar two-tier architecture (Figure 6). In the first step, all relevant elements for the given user are collected (based on his or her social network and active subscriptions). Not only items created since the last login are selected as candidates, but also items that the user has not yet seen. Furthermore, items which the user has interacted with, but which have been active since that interaction (e.g., there has been a discussion about them), are also included among the candidates.

In the following step, each element is scored on the basis of deep learning models, while the individual characteristics are combined linearly (first from evaluations, comments, etc.). Rules are applied in the first transition – for example, to ensure diversity of content (reduction of several repetitive videos etc.). The model selects approximately 500 final candidates from this adjusted set, which are finally ordered in the final transition (taking account of the current context).

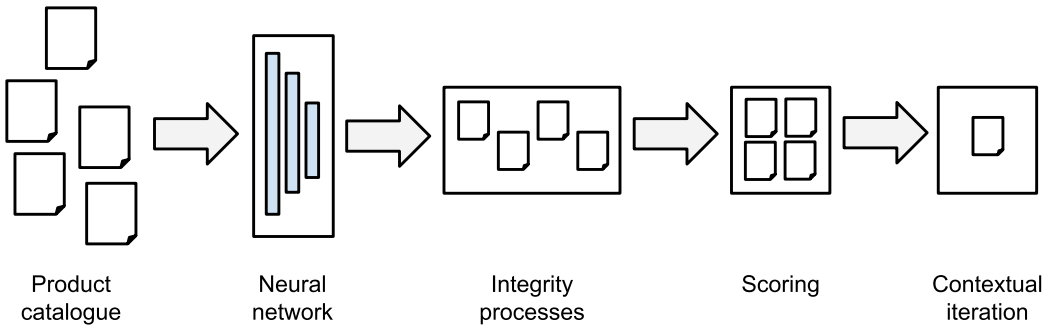

Figure 6
Architecture of the personalised order of posts on the Facebook online platform.
Source: Inspired.[67]

2.3 Impacts of the use of recommender systems

In terms of the impact of the technology used on the individual, two main aspects and other derived ones are important in the context of recommender systems:

[66] News Feed Ranking. Available at: https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/.
[67] Ibid.

Transparency of generated recommendations. Artificial intelligence and especially machine-learning methods often do not allow an unambiguous and understandable explanation of how a given recommendation was generated (or why item A is recommended and not item B); it is an open research problem.[68] Generation of explainable recommendations is in the interest of the recommender systems' operators themselves.

The use of deep learning and neural networks complicates the issue of the transparency of the generated recommendations further, since we do not have explicitly represented user preferences or a user model in this approach. In recent years, increasing scientific attention has been paid to the generation of human comprehensible explanations for end users (explanation of recommendations has been shown to improve users' trust and experience during interactions with the system[69]).

Utility function optimised in the model training process. Like other technologies, the trained recommendation model optimises or reflects the criteria set in the training process. It is the utility function that means the difference between recommendations that maximise user satisfaction and recommendations that maximise the time spent on a given online platform. However, in several real cases, the utility function is supplemented with explicitly defined rules. In other words, the list of recommendations generated by the utility function optimising the user's needs is supplemented with recommendations optimising other criteria (which could negatively affect the user).

For example, according to a 2019 research paper (the current situation may be different), the YouTube platform uses several utility functions based on user interaction metrics (click-through rate with respect to the recommended videos, time spent, etc.) and user satisfaction metrics.[70] Metric weights are set manually in order (in the words of the authors of the paper) "to achieve the best performance in terms of user interactions and satisfaction".[71]

Unique documents leaked from Facebook show that, in 2018, the company changed the utility function to prevent the decline in user interactions on the platform and support "Meaningful Social Interactions" (MSIs); this resulted in an increase of interactions on the platform, but user satisfaction declined.[72] Within the scope of these changes, a higher weighting was also given to user interactions in which they expressed their emotions (e.g., anger, laughter, etc.) compared to the neutral "like".[73] The consequence of both of these changes is that polarising

[68] See e.g.: ZHANG, Y. - CHEN, X. Explainable Recommendation: A Survey and New Perspectives, Foundations and Trends® in Information Retrieval, vol. 14, No 1, 2020, pp. 1-101. Available at: https://doi.org/10.1561/1500000066.
[69] TINTAREV, N. - MASTHOFF, J. A Survey of Explanations in Recommender Systems, 2007 IEEE 23rd International Conference on Data Engineering Workshop, 2007, pp. 801-810. Available at: https://doi.org/10.1109/ICDEW.2007.4401070.
[70] ZHAO, Z. et al. Recommending what video to watch next: A multitask ranking system. RecSys 2019 - 13th ACM Conference on Recommender Systems, 2019, pp. 43-51. Available at: https://doi.org/10.1145/3298689.3346997.
[71] Ibid.
[72] HAGEY, K. - HORWITZ, J. Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead. *The Wall Street Journal,* 2021. Available at: https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215.
[73] MERRILL, J.B. - OREMUS, W. Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation. *The Washington Post*, 2021. Available at: https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/.

content has become recommended more and users are more likely to comment on, share and express (often negative) emotions on it.

These cases both illustrate the importance of the utility function and the setting of weighting for output recommendations; without knowledge of them, it is not possible to talk about real transparency and the possibility of controlling possible negative impacts. Simultaneously, it has been shown that a change in the utility function or its weightings can have unintended impacts on the recommended content and behaviour of users, which are difficult to detect without systematic auditing or testing that currently seems to be absent for online platforms.

Enclosing users in information bubbles. If the design of the recommender system (e.g., type of method, utility function) ignores the limitations of the technology, its use may lead to users being enclosed in *information bubbles*.[74] On the one hand, in real life, users are surrounded by people with similar views or interests. In this respect, it is clear that the recommender system will recommend posts from such similar users (who are, for example, friends) on the online platform. On the other hand, some methods are likely to recommend similar content by their very nature (e.g., content recommendations).

Echo chambers are also referred to in connection with information bubbles, which mean a state of intellectual isolation in which the same thoughts and opinions are repeated, mutually affirmed and reinforced in relatively homogeneous groups.[75] These chambers are the result of the natural behaviour of humans described above, but algorithms contribute to this, for example by enclosing users in information bubbles.

Approaches exist to reduce the tendency of recommendation algorithms to enclose users in information bubbles, e.g., diversification,[76] or presentation of different perspectives in news reporting,[77] though these also have their limitations (e.g., the identification of different perspectives and positions requires advanced natural language methods and is still an open research problem) and they are not always consistent with the main objective of the platforms based on attention economy.

Bias and fairness in recommendations. As we have already indicated in a previous point, there are various biases in the data used to generate recommendations, which often arise from the prejudices and stereotypes present in society. Since a recommender system is based on such data, it is almost certain that this bias will also be reflected in the generated recommendations. For example, if there is a bias in the data that women get lower-paid work, from the point of view of the utility function (generation of a successful recommendation) recommending lower-paid work is optimal (because this is an observation from everyday life).

[74] The term 'filter bubble' is also used. See e.g.: PARISER, E. The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think, 2012.
[75] BESSI, A. Personality traits and echo chambers on Facebook. Computers in Human Behavior 65, 2016, pp. 319-324. Available at: https://doi.org/10.1016/j.chb.2016.08.016.
[76] KUNAVER, M. - POŽRL, T. Diversity in recommender systems – A survey, Knowledge-Based Systems, vol. 123, 2017, pp. 154-162. Available at: https://doi.org/10.1016/j.knosys.2017.02.009.
[77] DRAWS, T. - TINTAREV, N. - GADIRAJU, U. Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics. SIGKDD Explor. Newsl. 23, 1, June 2021, pp. 50-58. Available at: https://doi.org/10.1145/3468507.3468515.

In this way, we inadvertently encode such biases and various stereotypes into decision support systems.

At the same time, biases are caused not only by the data and the learning algorithm, which picks them up from the data, but also by the interface and user interaction with it (for example, *position bias*, where items placed higher in the list are more likely to be clicked on by users, plays a significant role in presenting recommendations, but also *selection bias* caused by the choices of users who rate the items they like more often than those they are not interested in); simultaneously, this is a cycle of reinforcement, where one bias generates and reinforces another.[78]

Biases and guaranteeing fairness are a problem for artificial intelligence methods in general, not only recommender systems, and it manifests itself in different domains, such as evaluation of candidates for jobs or promotions, assessment of financial situation in order to provide a loan, and so forth, but also on social media and other online platforms.

Collection of feedback and deriving user preferences in the context of privacy protection. Although collection of feedback or user actions is conceptually separate from the methods of personalised recommendation, it is clear that information about the user preferences is a necessity in order to generate personalised recommendations. Just as in other application domains, such information poses a potential risk of abuse. Therefore, steps must be taken to minimise it, which can take several forms – from transparently communicating what is recorded about the user, to allowing the setting of the scope of data collected, to storing data and derived preferences (user model) on the user side (which is being researched, for example, in the area of distributed user models[79] or differentiable privacy approaches[80]).

In the conclusion to this section, we can state that recommender systems are a useful technology aiming to reduce the information overload on users and give them more effective access to information relevant to them. On the other hand, like other AI system deployments, they bring several risks of negative impacts on users, in general and specifically in the context of the dissemination of disinformation. At the same time, recommender systems are widely used by online platforms as part of their operation model based on attention economy, in which users are used as a profit-increasing product (through online advertising sales). Simultaneously, today, very large online platforms (especially social media) are a major source of information and news for a large part of the population, which increases the risk of possible negative impacts. Therefore, given the economic power of online platforms and their

[78] BAEZA-YATES, R. Bias on the web. *Communications of the ACM*, 61(6), 2018, pp. 54-61. Available at: https://doi.org/10.1145/3209581.

[79] For example, GUO, H. - CHEN, J. - WU, W. - WANG, W. Personalization as a service: The architecture and a case study. In Proceedings of the 1st International Workshop on Cloud data management (CloudDB '09). ACM, New York, NY, USA, 2009, pp. 1-8. Available at: http://pages.cs.wisc.edu/~wentaowu/papers/cikm09-clouddb-workshop.pdf.

[80] See e.g.: CHEU, A. - SMITH, A. - ULLMAN, J. - ZEBER, D. - ZHILYAEV, M. Distributed Differential Privacy via Shuffling. In: ISHAI, Y. - RIJMEN, V. (eds) Advances in Cryptology – EUROCRYPT 2019. Lecture Notes in Computer Science, vol. 11476. Springer, Cham. Available at: https://doi.org/10.1007/978-3-030-17653-2_13.

impact on society, we see a clear need to legislatively regulate the operation of online platforms and the algorithms they use.

3. Selected regulations proposed by the European Commission in relation to the dissemination of disinformation and the behaviour of online platforms

The dissemination of disinformation and measures to limit or moderate speech on the internet inevitably clash with the boundaries of fundamental rights and freedoms. In the Slovak Republic the system for the protection of fundamental rights and freedoms is three-tiered. The first tier is represented by legal provisions in the Second Part of the Constitution of the Slovak Republic, which cover fundamental rights and freedoms. As a Member State of the EU, the Slovak Republic is bound by provisions on human rights in EU primary law, including the Charter of Fundamental Rights of the European Union (the second tier). Simultaneously, within the scope of the third tier of protection, the Slovak Republic is a signatory to several international legal obligations, including the Council of Europe Convention for the Protection of Human Rights and Fundamental Freedoms.[81]

During the regulation of disinformation on online platforms, it should be stressed that the fundamental rights and freedoms of both users (individuals) and providers of platforms may be infringed. From the individuals' point of view, we may mention interference with freedom of expression, the right to information, the protection of privacy and personal data, and the right to a fair trial. The activities of online platforms are, on the other hand, supported by the right to disseminate information, the right to own property and the right to conduct business (Figure 7).

[81] See e.g. HUSOVEC, M. - MESARČÍK, M. Ľudské práva ako všeobecný limit technológií [Human rights as a general limit on technology]. In HUSOVEC, M. - MESARČÍK, M. - ANDRAŠKO, J. Právo informačných a komunikačných technológií I. [ICT Law I] TINCT, 2021, p. 66 et seq.

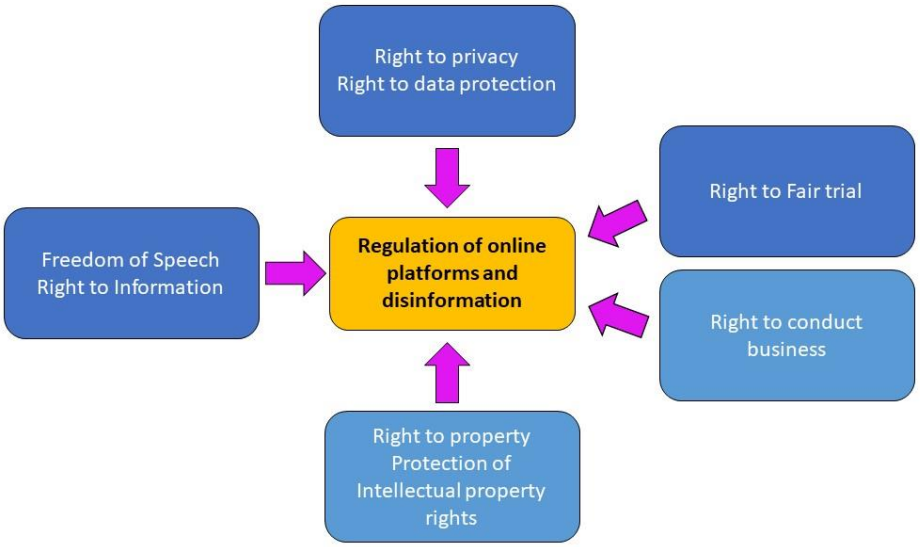

Figure 7
Dissemination of disinformation on online platforms and potential infringements of fundamental human rights and freedoms. Rights and freedoms involving individuals are indicated by a darker colour. Online platforms are indicated by a lighter colour.
Source: Authors' own work.

All the above-mentioned legal acts contain a list of fundamental rights and freedoms, and an intervention must be balanced, necessary and proportionate. One of the fundamental rights most mentioned in relation to disinformation is **freedom of expression**. Freedom of expression is a pillar of democratic societies and essential for social dialogue.[82] However, at the current time, public debate and democratic processes are often exposed to the deliberate dissemination of disinformation,[83] so it is appropriate to consider how to regulate this issue and restrict freedom of expression.

Freedom of expression protects not only individuals, but also service providers. Typical interventions by authorities on this right in the context of new technologies are removal of content or the blocking of websites, especially through the court´s orders.

Freedom of expression is not the only fundamental right affected by the dissemination of disinformation. Interventions on the **right to privacy** and the **right to personal data**

[82] European Commission. JOINT COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Action Plan against Disinformation JOIN/2018/36 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018JC0036.
[83] Ibid.

protection[84] also need to be discussed, since disinformation can significantly affect an individual's privacy and, simultaneously, its disclosure may involve automated analysis of different types of personal data. Legal requirements for the processing of personal data are unified in EU law in the form of the General Data Protection Regulation.

The **European Court of Human Rights** often handles cases related to the dissemination of information and freedom of expression. The decisions against Poland concerning the dissemination of disinformation in election campaigns, where the ECtHR criticised Polish election laws, are particularly well known. In the case of *Brzeziński* v *Poland*, the court ruled that freedom of expression had been infringed on the grounds that the courts had classified political statements as lies without further examination. The court presented a similar statement when examining the veracity of the information in the case of *Kwiecień* v *Poland*. Last but not least, in the case of *Kita* v *Poland* the court again ruled that freedom of expression had been infringed, since the Polish courts did not proceed according to objective criteria and classified all the complainant's statements as false without examining the facts.[85]

Similarly, in an analysis of the provisions of the Criminal Code, the **Constitutional Court of the Slovak Republic** referred to the problem of classifying information as false: 'Deciding on verbal hate crimes is a very delicate matter. With theft, the matter is straightforward. Constitutionally, there is essentially nothing to balance. However, with hate crimes, the values at stake are as much as freedom of speech, on the one hand, and the social cohesion or sociality of the state as such, on the other. Hate crimes are the only legal norms based on which an opinion may be punished that is often meaningless, even stupid, but still an opinion (…) The specific environment of the internet complicates the current situation further. Any ruling of the Constitutional Court may not be socially ideal if it weakens one of the valuable constitutional values.'[86]

From the perspective of users, the right to a fair trial is also worth mentioning. In terms of human rights, a fair trial does not only encompass the right of access to a court, or resolution of a dispute in due time. Increasingly, case-law, but also practice, construes the right in question through the requirement of *due process*, in the broadest possible sense. For this reason, it is essential to guarantee users transparency and give them tools to protect their rights in automated decision-making, including the use of artificial intelligence on online platforms. In the context of the right to a fair trial and algorithmic decision-making, the literature emphasises, above all, the transparency, accountability and contestability of such decisions.[87]

[84] The Charter of Fundamental Rights of the EU and the Constitution of the Slovak Republic make an explicit distinction between the right to privacy and the right to protection of personal data and arrange them separately. For details, see FUSTER, G. - HIJMANS, H. The EU rights to privacy and personal data protection: 20 years in 10 questions. Available at: https://brusselsprivacyhub.eu/events/20190513.Working_Paper_Gonza%CC%81lez_Fuster_Hijmans.pdf.

[85] For details on the case-law of the European Court of Human Rights and on freedom of expression see the European Parliament study. The fight against disinformation and the right to freedom of expression. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/695445/IPOL_STU(2021)695445_EN.pdf.

[86] Finding of the Constitutional Court of the Slovak Republic. PL. ÚS 5/2017-117. Point, 71.2.

[87] For details, see, for example, FROSIO, G. - GEIGER, CH. Taking Fundamental Rights Seriously in the Digital Services Act's Platform Liability Regime. European Law Journal (October 2021, Forthcoming), pp. 27-28. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3747756.

As EU-level strategy papers tackling disinformation recognise simultaneously, its dissemination through online platforms, including social media, is a key element. The regulation of the rules of business and the activities of private entities necessarily conflicts with the right to conduct business or the right to own property, including the protection of intellectual property rights. For this reason, legislation concerning the **accountability (and liability) of online platforms** should be mentioned.

The general regulation of liability for third-party content in information society services is presented in the **Directive on electronic commerce**, which regulates the conditions under which the provider is not liable for the content on its platform according to its type ('safe harbour'). Under the current wording of the law, the *notice-to-takedown* principle applies, which, in practice, means that if the platform is not aware of the problematic content, it is not accountable for it.[88] If the platform cannot rely on safe harbour (for example, if someone reports the illegal content to it), it may be held accountable through national law. In the Slovak Republic, this is primarily through civil (tort) liability.[89]

Simultaneously, doing business is linked to the ability to promote your products or services through advertising. As mentioned in Part 2 of this study, automated tools for profiling and the use of artificial intelligence for personalised advertising have come to the fore. These are the issues that are regulated by area of personal data protection, advertising or consumer protection.

In view of the economic power and market position of online platforms, another set of tools to regulate and restrict their activities is represented by **competition law**, which restricts the right to do business in a specific manner. This area will be regulated by the 'Digital Markets Act' in relation to large online platforms.[90]

In connection with the dissemination of disinformation, the EU has currently presented a package of legislative acts to significantly help tackle disinformation. One of the fundamental changes is the new rules for regulation of online platforms in the form of the Digital Services Act (**DSA)**, which simultaneously revises the regime of accountability in the Directive on electronic commerce.

Since the legislative process takes years for objective reasons, in the context of tackling disinformation, the EU presented the **Code of Practice on Disinformation** as a kind of 'vanguard' before adopting the DSA. Simultaneously, it is an instrument that most online

[88] For more details see HUSOVEC, M. Zodpovednosť na internete podľa českého a slovenského práva. [Accountability on the internet under Czech and Slovak law.] CZ-NIC, 2014. Available at: https://knihy.nic.cz/files/edice/zodpovednost_na_internete.pdf.

[89] The theoretical justification for using a given accountability model is explained in more detail in the literature, e.g. FROSIO, G. - HUSOVEC, M. Accountability and Responsibility of Online Intermediaries in FROSIO, G. (ed). The Oxford Handbook of Online Intermediary Liability, pp. 613-630. Specific regime applies to copyright protected works on online content platforms.

[90] Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act). COM/2020/842 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A842%3AFIN#.

platforms have undertaken to respect. The Code arranges specific rules for controlling the placement of advertising; political advertising and current issues advertising; integrity of services; empowering users and the research community. Compliance with the rules is subject to surveillance by the online platforms themselves and the European Commission. The Commission itself presented guidelines for reinforcing the Code in summer 2021.[91] An updated version of the Code should be drawn up in accordance with those guidelines. Monitoring of compliance with the Code shows that not all online platforms effectively put the requirements into practice.[92]

In spring 2021, the proposal for artificial intelligence regulation was presented in the form of the Artificial Intelligence Act (**AIA**). In the context of the conclusions of Part 2 of the study it is therefore essential to consider the rules for the development and use of artificial intelligence as they are also important in the fight against dissemination of disinformation.

Given the timeliness of the proposals and given that dissemination of disinformation is today enhanced by the use of artificial intelligence, we will look more closely at the provisions contained in the DSA and AIA.

3.1 Proposal for artificial intelligence regulation (AIA)

In April 2021, the European Commission presented a proposal of the first comprehensive regulation of artificial intelligence in the EU – the draft artificial intelligence regulation, known by the acronym AIA (**A**rtificial **I**ntelligence **A**ct). The Commission justifies the adoption of the legislation with several reasons, which are named and analysed in the impact assessment.[93] The reasons the Commission gave for adopting the artificial intelligence regulation included increased risk for the safety of individuals, risks for fundamental rights and freedoms and a lack of powers for public authorities for surveillance of compliance with the rules. The Commission also highlighted the legal uncertainty with the complex and fragmented legal provisions, lack of trust in society towards AI applications, which could eventually slow innovation and affect the competitiveness of the EU on the global market, and the danger of fragmented national legal provisions that could jeopardise the EU single market.[94] The above reasons are subsequently projected into specific legal requirements in the AIA proposal.

3.1.1 General considerations

The AIA is *horizontal product legislation* on a risk-based approach. The horizontality reflects the fact that the AIA will apply to AI systems in general, not just to narrowly defined sectors.

[91] European Commission. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS European Commission Guidance on Strengthening the Code of Practice on Disinformation. COM/2021/262 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021DC0262#.

[92] For example, ERGA. ERGA Report on disinformation: Assessment of the implementation of the Code of Practice. Available at: http://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf.

[93] European Commission. IMPACT ASSESSMENT Accompanying the Proposal for a Regulation of the European Parliament and of the Council LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. SWD(2021) 84 final.

[94] Ibid. pp. 13-28.

Article 3(1) of the AIA defines an AI system as:

> 'software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.'

Annex I gives a list of techniques and approaches[95] that the European Commission can dynamically amend with regard to the newest developments.

The definition of artificial intelligence in the proposed regulation is problematic, because it is conceived very broadly and will cover a great number of AI-based applications.[96] In our publicly accessible comments, we have emphasised that the definition in question should be adapted in such a way that the techniques and approaches set out in Annex I of the AIA shall represent an integral part of an AI system and not just be 'used' in its development. Simultaneously, their impact on the generation of outputs should be taken into account. Likewise, we consider that the techniques and procedures in Annex I of the AIA should not be interpreted in isolation, but only under the condition of 'intelligent' behaviour, whose criteria (e.g., learning ability) should be defined by the expert community.[97]

The AIA is conceived using a risk-based approach. Practically speaking, this means that it categorises AI systems according to the degree of risk to the fundamental rights and freedoms or security of individuals. The AIA differentiates (see Figure 8):

1. unacceptable risk (prohibited practices)
2. high risk
3. low risk
4. minimum risk.

A no less important factor is that the AIA is a product regulation. That means that artificial intelligence systems are regarded as a product and in terms of the conditions they must fulfil in order to be developed and placed on the EU's single market. It does not contain a list of the specific rights of individuals, as for example, the GDPR does for the rights of the data subjects.

The AIA largely stipulates the requirements for high-risk AI systems. These will have to undergo conformity assessment, where the AIA provisions several requirements, for example,

[95] Specifically, it lists: '(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;
(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
(c) Statistical approaches, Bayesian estimation, search and optimization methods.'

[96] See, e.g. the Kempelen Institute of Intelligent Technologies. Stance on the Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Available at: https://kinit.sk/wp-content/uploads/2021/09/KINIT_Stance-of-AIA_Paper_2021_09.pdf. MULLER, C. - DIGNUM, V. ARTIFICIAL INTELLIGENCE ACT: ALLAI ANALYSIS & RECOMMENDATIONS. Available at: https://allai.nl/draft-ai-act-allai-analysis-and-recommendations/.

[97] Kempelen Institute of Intelligent Technologies. Stance on the Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Available at: https://kinit.sk/wp-content/uploads/2021/09/KINIT_Stance-of-AIA_Paper_2021_09.pdf.

in terms of data management, detection of potential biases, technical documentation and other areas. In most cases, conformity assessment will be performed by AI system providers at their own expense. Subsequently, AI system providers may have an EU declaration of conformity at their disposal, which entitles them, after registration in the EU database for standalone high-risk artificial intelligence systems, to place the given system on the market, where it is subject to system monitoring requirements. In the event that AI system operators are already subject to a conformity assessment according to special laws, they will extend the initial conformity assessment to include the AIA requirements.

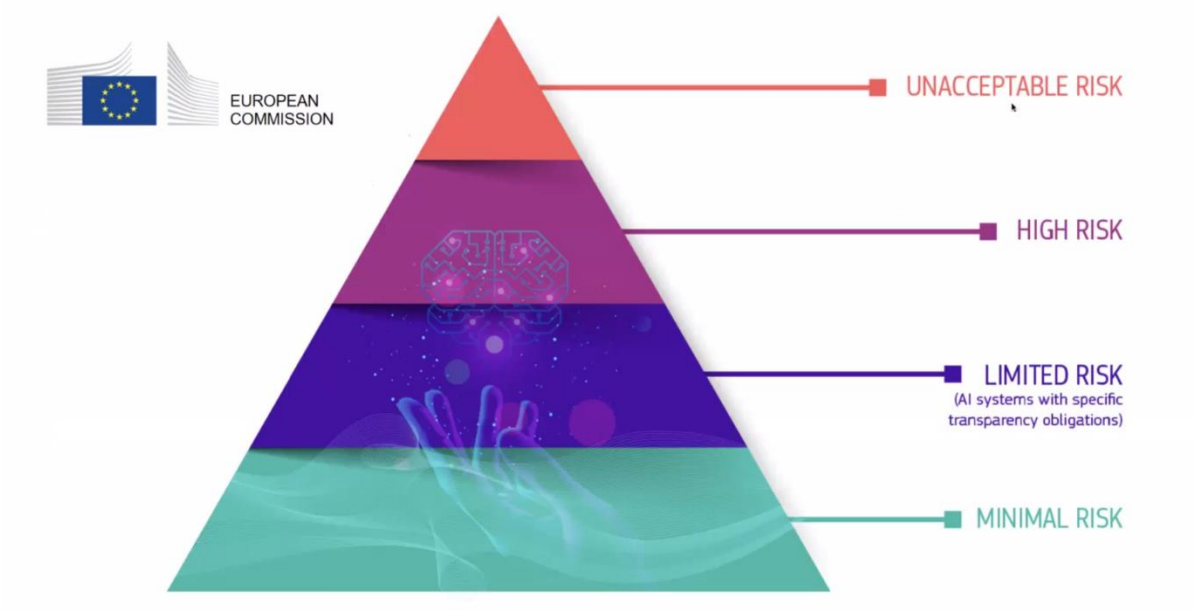

Figure 8
Proposed regulation of artificial intelligence and risk assessment of AI systems.
Source: European Commission website.[98]

The AIA envisages oversight through national authorities, which Member States will have to set up. Simultaneously, like the GDPR, the AIA contains heavy penalties for non-compliance, and the largest administrative fine that can be imposed is up to EUR 30 000 000 or, if a company is the infringer, up to 6% of its total global annual turnover for the previous financial year, whichever is higher.[99]

3.1.2 Combating disinformation.

The text of the AIA does not contain a direct reference to disinformation and the same conclusion applies to the impact assessment. However, this does not automatically mean that the AIA does not affect the dissemination of disinformation using AI systems. On the contrary. Since it is a horizontal general regulation, the provisions can also affect online

[98] European Commission. Regulatory framework proposal on artificial intelligence. Available at: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.
[99] See Article 71 of the AIA.

platforms and the dissemination of disinformation. We will focus on issues of prohibited practices, high-risk AI systems and specific provisions in the framework of conformity assessment and audits.

3.1.2.1 Prohibited practices

Article 5 of the AIA prohibits three artificial intelligence practices and significantly limits another.

Specifically, it prohibits:

> '(a) the placing on the market, putting into service or use of an AI system that **deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour** in a manner that causes or is likely to cause that person or another person physical or psychological harm;
>
> (b) the placing on the market, putting into service or use of an AI system that **exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group** in a manner that causes or is likely to cause that person or another person physical or psychological harm;
>
> (c) the placing on the market, putting into service or use of AI systems **by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons** over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both of the following:
>
> I. detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected;
> II. detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity.'

Simultaneously, Article 5(1)(d) of the AIA in connection with paragraph 2 significantly restricts the use of **'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement**. The AIA stipulates precise derogations and room for manoeuvre for EU Member States to stipulate such derogations.

In the context of AI system use and the dissemination of disinformation, interpretation of prohibited practices concerning limiting of manipulation,[100] whether exploiting subliminal techniques or the vulnerabilities of a specific group of people, will be the key. From the perspective of interpretations of the term manipulation presented in the literature, the AIA proposal reflects most of the requirements. It requires intent ('with purpose'), abuse of vulnerabilities (age, mental or physical disorder) or subliminal effect. However, the essential

[100] Recital 15 of the AIA: 'Aside from the many beneficial uses of artificial intelligence, that technology can also be misused and provide novel and powerful tools for **manipulative, exploitative** and social control practices.'

feature of manipulation is missing, i.e., the requirement to support the manipulator's own objectives.[101] Rather, the AIA requires real or potential mental or physical harm.

Furthermore, recital 16 of the AIA limits the use of these prohibitions even more significantly and does not reflect the dynamics of relationships during use of AI systems:

> 'The intention **may not be presumed if the distortion** of human behaviour results from **factors external to** the AI system which are **outside of the control of the provider or the user.**'

Practically speaking, this could mean that recommender systems on online platforms that favour disinformation will not fall under the prohibited practice. We present this conclusion because the current proposal ofAIA does not require fulfilment of the requirement of manipulation to support own objectives of the provider. It would be very difficult to prove that all disinformation *en bloc* could potentially harm individuals. Simultaneously, the problem is that these prohibited practices do not take context into account, and at this point we do not know what the exact interpretation of the practices will be in the future.

3.1.2.2 High-risk AI systems

As already mentioned, most of the provisions in the AIA proposal will apply to high-risk AI systems. An AI system can be classified as a high-risk system by two methods. The first method is a reference to specific product legislation that already contains certain requirements for AI systems as safety components.[102] The second method is cases of standalone AI systems referred to in Annex III of the AIA.[103]

In Annex III the European Commission lists areas and applications of AI systems that it considers high-risk AI systems:

- Biometric identification and categorisation of natural persons;
- Management and operation of critical infrastructure;
- Education and vocational training;
- Employment, workers management and access to self-employment;
- Access to and enjoyment of essential private services and public services and benefits;
- Law enforcement;
- Migration, asylum and border control management.

A closer look at the areas and specific applications of high-risk AI systems reveals that the classic attention economy models of online platforms characterised in Part 2 of this study are not included in the high-risk AI systems of Annex III. These systems cannot be subsumed under any of the areas or applications.

[101] VAELE, M. - BORGESIUS, F.Z. Demystifying the Draft EU Artificial Intelligence Act. Forthcoming in 2021 22(4) Computer Law Review International.
[102] Article 6(1) of the AIA.
[103] Article 6(2) of the AIA.

We consider this fact to be a fundamental shortcoming of the AIA, which we have also drawn attention to publicly elsewhere.[104] We propose inclusion of the area of attention economics among high-risk artificial intelligence systems, similar to the case of biometrics, where a part is prohibited (restricted) and a part is subject to the requirements for high-risk AI systems.

3.1.2.3 Requirements for high-risk AI systems and post-market monitoring

In the event of classification of AI systems that use online platforms that could be involved in the dissemination of disinformation among high-risk AI systems, there would be several essential requirements. In connection with high-risk AI systems, the AIA sets out a number of key requirements that the AI system provider must take into account when assessing conformity.

One of them is the risk management system pursuant to Article 9 of the AIA, which includes 'continuous iterative process run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic updating.' Simultaneously, the AIA directly states the minimum requirements for the risk management system, which are, however, relatively general and must be adapted to the specific case.

In terms of data management, the AIA contains an important requirement for training, validation and testing data sets, namely that appropriate data management and control procedures must be in place that require examination for possible bias (*bias detection*).[105]

Equally, key requirements concern transparency towards business users[106] and human oversight.[107]

After placing high-risk AI system on the market, there is a requirement to implement and document a monitoring system in a manner that is appropriate to the nature of the artificial intelligence technology and the risks of the high-risk artificial intelligence system.[108] However, this requirement is formulated very generally, and it is not possible to ascertain the specific parameters of potential publicly available reports or audits from it.

3.2 Proposal for digital services regulation (DSA)

At the end of 2020, the European Commission presented a proposal for a regulation on digital services known by the acronym DSA (**D**igital **S**ervices **A**ct). The principal motive was the revision of the 20-year-old legal framework for online platform accountability in the form of the Directive on electronic commerce. Simultaneously, the legislator reflected the

[104] Kempelen Institute of Intelligent Technologies. Stance on the Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Available at: https://kinit.sk/wp-content/uploads/2021/09/KINIT_Stance-of-AIA_Paper_2021_09.pdf.
[105] Article 10(2)(f) of the AIA.
[106] Article 13 of the AIA.
[107] Article 14 of the AIA.
[108] Article 61 of the AIA.

development of the digital market and the growing economic power of online platforms, which has led to debate on possible risks and regulation.

The Commission has identified three specific reasons leading to the proposed legislation:

- First, there has been an increase in the social and economic risks associated with the use and behaviour of online platforms and risk of potential harm to individuals and their fundamental rights and freedoms.
- Second, there is a lack of supervision and cooperation in digital services, which is eroding the EU single market.
- The third reason is to remove barriers for smaller companies to provide digital services due to the position of large online platforms.[109] These reasons are subsequently projected into specific legal requirements in the DSA.

3.2.1 General considerations

In terms of personal scope, the draft DSA makes a distinction between four actors, to whom specific legal requirements apply. Essentially, there are a number of sets of relationships.
The DSA differentiates (see Figure 9):

- Intermediary services;
- Hosting services;
- Online platforms; and
- Very large online platforms. [110]

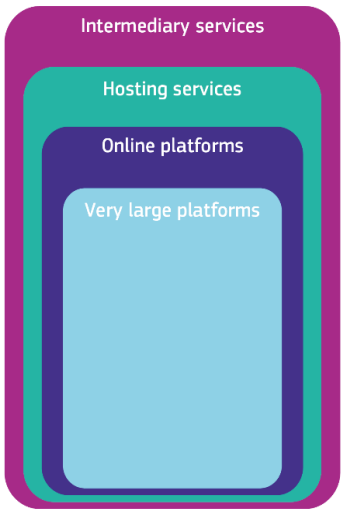

Figure 9
Number of entities governed by the DSA.
Source: European Commission website.[110]

The greatest number of actors are **intermediary services**, which the draft DSA divides into mere conduit, caching and hosting services.[111] Mere conduit type services include internet connection providers or open WIFI network operators. The essence of caching services is the intermediate and temporary storage of information, for the purpose of making the service more efficient.

[109] European Commission. IMPACT ASSESSMENT Accompanying the document PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. SWD(2020) 348 final.
[110] European Commission. The Digital Services Act: ensuring a safe and accountable online environment. Available at: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en.
[111] Article 10(2)(f) of the DSA:
'intermediary service' means one of the following services:
–a '**mere conduit**' service that consists of the transmission in a communication network of information provided by a recipient of the service, or the provision of access to a communication network;

–a '**caching**' service that consists of the transmission in a communication network of information provided by a recipient of the service, involving the automatic, intermediate and temporary storage of that information, for the sole purpose of making more efficient the information's onward transmission to other recipients upon their request;

–a 'hosting' service that consists of the storage of information provided by, and at the request of, a recipient of the service.'

In terms of disseminating disinformation, **hosting** services are an important set. In principle, these are web hosting services, cloud services, repositories, social networks or online advertising.

A subset of hosting services are **online platforms**, which the DSA defines as

> 'a provider of a hosting service which, at the request of a recipient of the service, stores and disseminates to the public information, unless that activity is a minor and purely ancillary feature of another service and, for objective and technical reasons cannot be used without that other service, and the integration of the feature into the other service is not a means to circumvent the applicability of this Regulation.'[112]

Among online platforms we may typically include social media, online marketplaces and app stores. At first glance, it seems that the given definition will not include messaging services such as WhatsApp or Telegram. However, this conclusion will depend on the interpretation of the term 'public dissemination of information' and may change during the process of fine-tuning legislation.

Specific attention and legal requirements also focus on the last set of actors – **very large online platforms**. These are not directly defined in the DSA, although the proposal for the regulation contains a guide for classification. Under Article 25(1) of the DSA very large online platforms shall be understood as those intermediaries 'which provide their services to a number of average monthly active recipients of the service in the Union equal to or higher than 45 million, calculated in accordance with the methodology' set out in the DSA.

The requirements placed on the abovementioned actors by the DSA vary and thus it is always important to correctly classify the service provider. Below we provide a brief overview of the actors and selected obligations under the DSA (Table 3).[113]

[112] Article 10(2)(h) of the DSA.
[113] European Commission. The Digital Services Act: ensuring a safe and accountable online environment. Available at: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en.

Table 3.
Obligations for various entities under the DSA.
Source: DSA.

#	INTERMEDIARY SERVICE	HOSTING	ONLINE PLATFORM	VERY LARGE ONLINE PLATFORM
Transparency and accountability	✓	✓	✓	✓
Requirements for general terms and conditions	✓	✓	✓	✓
Cooperation with national authorities	✓	✓	✓	✓
Designation of contact point (representative)	✓	✓	✓	✓
Information obligations		✓	✓	✓
Mechanisms for handling complaints and out-of-court settlement of disputes			✓	✓
Trusted flaggers			✓	✓
Measures to prevent abusive notifications			✓	✓
Traceability of traders			✓	✓
Transparency of advertising			✓	✓
Risk management and identification of the responsible person				✓
External audits and public responsibility (accountability)				✓
Transparency of recommender systems				✓
Cooperation in crisis management				✓

Like the AIA, the DSA foresees the designation of a competent authority in each Member State to supervise compliance with the requirements stipulated in the DSA. Member States will be able to determine the specific sanctionsby themselves. However, the DSA assumes that the maximum amount of penalties imposed for a failure to comply with the obligations laid down in the DSA Regulation will not exceed 6% of the annual income or turnover of the provider of intermediary services concerned.[114] The European Commission itself will be able to act against very large online platforms.[115]

3.2.2 Tackling disinformation

The European Commission, in line with the conclusions presented in the document 'Tackling online disinformation: a European Approach', decided not to adopt general regulation aimed exclusively at tackling disinformation.[116] Instead, it presented a 'softer' mechanism in the form of the abovementioned Code of Practice on Disinformation. Simultaneously, the DSA itself partly reflects the need for greater accountability of online platforms for disseminated content in order to prevent the intentional manipulation of services by their users through fake accounts or bots.[117]

In the context of the DSA, tackling disinformation is explicitly mentioned in the impact assessment as well as in recitals of the regulation. The impact assessment identifies the dissemination of disinformation as one of the activities, which may not strictly fall under the definition of illegal content, but it also poses a huge risk in terms of informed decision-making and the openness of political processes. It is the dissemination of disinformation that can be emphasised and multiplied through online platform algorithms[118] or organised bot activities which artificially increase the interest or reach of content and contribute to the dissemination of disinformation.[119] This conclusion applies particularly in relation to recommender systems.[120]

The non-binding text of DSA recitals also highlights the abovementioned risks in relation to the advertising systems of large online platforms and the access to their archives.[121] The DSA emphasises the role of **codes of conduct** in disseminating disinformation, while large online platforms should be bound by these codes:

[114] Article 42 of the DSA.
[115] Article 51 et seq. of the DSA.
[116] European Commission. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Tackling online disinformation: a European Approach COM/2018/236 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236.
[117] HOBOKEN VAN, J. - Ó FATHAIGH, R. Regulating Disinformation in Europe: Implications for Speech and Privacy. UC Irvine Journal of International, Transnational, and Comparative Law. Volume 6 Symposium: The Transnational Legal Ordering of Privacy and Speech. Article 3, 2021, pp. 15-16.
[118] European Commission. IMPACT ASSESSMENT Accompanying the document PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. SWD(2020)348 final, p. 16.
[119] Ibid. p. 17.
[120] Ibid. p. 28.
[121] Recital 63 of the DSA.

> ‘Another area for consideration is the possible negative impacts of systemic risks on society and democracy, such as **disinformation** or manipulative and abusive activities. This includes coordinated operations aimed at amplifying information, including **disinformation, such as the use of bots or fake accounts for the creation of fake or misleading information, sometimes with a purpose of obtaining economic gain, which are particularly harmful for vulnerable recipients of the service,** such as children. In relation to such areas, adherence to and compliance with a given **code of conduct** by a very large online platform may be considered as an appropriate risk mitigating measure. The refusal without proper explanations by an online platform of the Commission’s invitation to participate in the application of such a code of conduct could be taken into account, where relevant, when determining whether the online platform has infringed the obligations laid down by this Regulation.’[122]

The issue of disinformation is also mentioned in connection with drafting crisis protocols in the event of emergencies (for example, a pandemic or a natural disaster), when large online platforms are expected to cooperate with the competent authorities.[123]

To conclude this passage, it should be emphasised that the DSA significantly regulates the actions of online platforms in relation to **illegal content**. The DSA defines ‘illegal content’ as:

> ‘any information, which, in itself or by its reference to an activity, including the sale of products or provision of services is **not in compliance with Union law** or the law of a Member State, irrespective of the precise subject matter or nature of that law.’

Semantically, the dissemination of disinformation does not necessarily meet the definition of illegal content, such as child pornography or advertising for psychotropic substances. This statement is also affirmed by recital 22 of the DSA: ‘The removal or disabling of access should be undertaken in the observance of the principle of freedom of expression.’

3.2.2.1 Obligations of online platforms

A basic issue in combating disinformation is the responsibility of key actors in the process – the platforms through which disinformation spreads the fastest. In this regard, there is a need to discuss options for liability for content posted on the platforms by third parties (users). From the perspective of Slovak law, this is a situation consisting of cooperation in committing offences by way of participation, the use of a person to commit an offence or failure to intervene in order to protect the rights of another party.[124] The question of whether a person is responsible for third-party content published on an online platform is today covered by the Directive on electronic commerce and the Slovak Act on electronic commerce. This regime is also the subject of provisions in the DSA.

[122] Recital 68 of the DSA.
[123] Recital 71 of the DSA.
[124] HUSOVEC, M. Digitálny trh EÚ a zodpovednosť poskytovateľov služieb [The EU digital market and the responsibility of service providers]. In HUSOVEC, M. - MESARČÍK, M. - ANDRAŠKO, J. Právo informačných a komunikačných technológií I. [ICT Law I] TINCT, 2021, p. 128 et seq.

Given the scope of this study, we will consider the issues of liability for third-party content of 'hosting' services, since within this type of service online platforms and very large online platforms, including social networks, fall under. Like the Directive on electronic commerce, the DSA maintains the concept of 'safe harbour', and thus the conditions that a hosting service should meet in order not to be liable for third-party content. The DSA arranges these conditions in Article 5(1):

> 'the service provider **shall not be liable** for the information stored at the request of a recipient of the service on condition that the provider:
> a) **does not have actual knowledge** of illegal activity or illegal content and, as regards claims for damages, is not aware of facts or circumstances from which the illegal activity or illegal content is apparent; or
> b) **upon obtaining such knowledge or awareness, acts expeditiously** to remove or to disable access to the illegal content.'

The requirement for meeting the conditions of a safe harbour and avoiding liability is passivity of the hosting service operator. The filtering and displaying of specific content to specific users is considered a technical and passive matter of platform operation.[125] However, this statement is dubitable when the activities of platforms include maintaining attention and algorithmic decision-making about displayed content.[126]

In a case submitted to the Court of Justice of the European Union, the **Austrian Supreme Court** also considered the question of whether online platforms remain neutral when optimising the display of their content. The Austrian Supreme Court favoured the interpretation that such conduct is a traditional business model and that structuring the search results should still be considered the passive behaviour of the platform.[127]

Simultaneously, a provider must not directly or indirectly incite a crime, otherwise it loses the safe harbour and becomes liable.[128] If a provider acquires actual (for example, based on notification) or constructive knowledge (for example, based on its activity)[129] about illegal activity or content, it is obliged to remove or disable access to such content. The DSA gives a relatively fairly accurate description of the mechanism by which a user or other entity can report illegal content ('*notice and action mechanisms*'). If these notifications fulfil the requirements laid down by the regulation, they are considered to be sufficiently accurate to

[125] BUITEN, M. The Digital Services Act: From Intermediary Liability to Platform Regulation. Working Paper. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3876328, p. 15. The European Commission identically. IMPACT ASSESSMENT Accompanying the document PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. SWD(2020) 348 final, p. 159.

[126] BUITEN, M. The Digital Services Act: From Intermediary Liability to Platform Regulation. Working Paper. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3876328, p. 15.

[127] Handelsgericht Wien [Higher Regional Court of Vienna] *Puls 4 TV GmbH & Co. KG* v *YouTube LLC and Google Austria GmbH*, Case No 4 R 119/18a, p. 10.

[128] Article 5(2) of the DSA. 'Paragraph 1 shall not apply where the recipient of the service is acting under the authority or the control of the provider.'

[129] See, for example, ruling of the Court of Justice C-324/09 of 12 July 2011 in *L'Oréal SA and Others* v *eBay International AG and Others* or Joined Cases C-236/08 to C-238/08 of 23 March 2010 – Google.

meet the requirement of 'actual knowledge' of illegal content and providers must act swiftly and take action against that content.[130]

In the case of online platforms and very large online platforms, the DSA also stipulates the new institution of trusted flaggers,[131] whose notices must be processed and decided upon with priority and without delay. An entity (not an individual) can be awarded the status of trusted flagger by the Digital Services Coordinator after informing the European Commission.

Article 6 of the DSA simultaneously explicitly supports voluntary own-initiative investigations by platforms. During activities aimed at detecting, identifying and removing, or disabling access to, illegal content, or taking the necessary measures to comply with the legal requirements, they can rely on the exemption from liability mentioned above.[132] In other words, if a platform voluntarily investigates its own content, this may not automatically mean, in terms of the law, acquiring objective knowledge of illegal content. The platform can still rely on the conditions of the safe harbour.

Article 7 of the DSA also prohibits obligations of own general monitoring of content:

> '**No general obligation to monitor the information** which providers of intermediary services transmit or store, nor **actively to seek** facts or circumstances indicating illegal activity shall be imposed on those providers.'

In its case-law, the Court of Justice of the EU has repeatedly stated that a general obligation for providers to monitor content would mean a significant violation of the right to conduct business and, in the long term, limits the innovation.[133] Perhaps this was put best by the Advocate General in the case of YouTube/Cyando in the context of copyright protection.[134] However, his conclusions can be applied by analogy to the obligation to check any content before publishing on platforms:

> 'Requiring online platform operators to **check, in a general and abstract manner, all the files** which their users intend to publish before they are uploaded in search of any copyright infringement would **introduce a serious risk of undermining these different fundamental rights**. Given the potentially considerable volume of hosted content, it would be impossible to carry out such a check in advance manually and, furthermore, the risk in terms of liability for those operators would be excessive. In practice, the smallest of them would be at risk of not surviving that liability and **those with sufficient resources would be forced to carry out general filtering of their users' content, without judicial review, which would result in a substantial risk of "over-removal" of that content.'**

[130] Article 14 of the DSA.
[131] Article 19 of the DSA.
[132] This is the 'Good Samaritan clause'. For details see KUCZERAWY, A. The Good Samaritan that wasn't: voluntary monitoring under the (draft) Digital Services Act. Available at: https://verfassungsblog.de/good-samaritan-dsa/.
[133] See, for example, judgment of the Court of Justice C-360/10 of 16 February 2012 in *SABAM* v *Netlog*.
[134] Judgment of the Court of Justice C-682/18 of 22 June 2021 in *Frank Peterson* v *Google LLC and Others* and *Elsevier Inc.* v *Cyando AG*.

3.2.2.2 Content moderation

Linking to the illegal content reporting system, the DSA contains additional obligations regarding the justification for removing content and internal and external dispute resolution mechanisms. Where, after investigation, an online platform decides to remove or disable access to illegal content, the DSA stipulates precise requirements for justification of that decision to the recipients of the service. Among others, the given justification must include an explanation of the removal or disabling of access to the illegal content and whether this was performed using automated means.[135] The DSA thus indirectly reflects the requirement of human oversight in the mechanism for reviewing decisions to remove or disable access to content. All decisions by online platforms will be published in a register maintained by the European Commission.

The DSA provides several mechanisms of legal protection against decisions by online platforms. Above all, this means an internal complaint-handling system. Recipients of the service have the option of filing a complaint if the platform has judged their activity to involve dissemination of illegal content or be non-compliant with the general business terms, which results in a decision to remove the content, partially or entirely suspend the provision of services to the recipient or block or cancel their account.[136] Online platforms must have an internal complaint-handling mechanism that such recipients can use. The DSA specifically stipulates that complaint handling shall not be based purely on output from automated means.[137]

If a complaint has been handled, the recipients have the option of using an out-of-court dispute resolution mechanism in the event of dissatisfaction with the result.[138]

Online platforms have the option of suspending, for a reasonable period of time, the provision of their services to users that provide manifestly illegal content. Likewise, online platforms have the right to temporarily suspend the processing of notices or complaints by users who abuse the notification system and submit notices or complaints that are manifestly unfounded.[139] Such steps are conditional on having given the affected users prior warning. The platform must communicate the criteria for assessing the validity of notices and complaints transparently in its general business terms.

Furthermore, in terms of moderating content, very large online platforms are required under Article 26 of the DSA to **regularly assess risk** and take measures to mitigate it (Article 27). They are required to carry out assessments once a year and it must be emphasised that this requirement concerns not only illegal content, but also assessment of:

[135] For details, see Article 15 of the DSA.
[136] Article 17(1) of the DSA.
[137] Article 17(5) of the DSA.
[138] Article 18 of the DSA.
[139] Article 20 of the DSA.

'**any negative effects for the exercise of the fundamental rights** to respect for private and family life, freedom of expression and information, the prohibition of discrimination and the rights of the child, as enshrined in Articles 7, 11, 21 and 24 of the Charter respectively;' and

'**intentional manipulation of their service**, including by means of inauthentic use or automated exploitation of the service, with an actual or foreseeable negative effect on the protection of public health, minors, civic discourse, or actual or foreseeable effects related to electoral processes and public security.'[140]

Therefore, the assessment of risk will necessarily have to include the dissemination of disinformation on very large online platforms. This is specifically indicated by the provision of Article 26(2) of the DSA, which requires specific examination of 'how their content moderation systems, **recommender systems and systems for selecting and displaying advertisement** influence any of the systemic risks referred to in paragraph 1, including the potentially rapid and wide dissemination of illegal content and of information that is incompatible with their terms and conditions.'[141]

3.2.2.3 Transparency, notification and accountability obligations

The DSA arranges several transparency requirements, whether in the form of publicly accessible reporting, advertising or recommender systems. The content and quality of legal transparency requirements vary depending on whether they are general intermediary services, online platforms or very large platforms. Transparency issues have been identified by the expert community as key for gaining knowledge of the methods and reasons for the effective dissemination of disinformation.[142]

The most general obligation of transparency is found in Article 12 of the DSA concerning the **general terms and conditions** of intermediary services. That inter alia must include information on 'any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review.'[143]

The requirement to regularly submit **reports on transparency** attracts our attention. Such a report must be published by intermediary services that are not micro or small enterprises at least once a year.[144] The same obligation applies to online platforms[145] but, in terms of content, the DSA adds several particulars for reports (see Table 4). Very large platforms must also publish transparency reports supplemented with information about audits or risk analyses performed every six months.[146]

[140] Article 26(1) of the DSA.
[141] Zhodne SAVIN, A. The EU Digital Services Act: Towards a More Responsible Internet. Copenhagen Business School Law Research Paper Series No 21-04. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3786792.
[142] Report of the independent High level Group on fake news and online disinformation. A Multi-Dimensional Approach to Disinformation, p. 22. Available at: https://op.europa.eu/en/publication-detail/-/publication/6ef4df8b-4cea-11e8-be1d-01aa75ed71a1.
[143] Article 12(1) of the DSA.
[144] Article 13 of the DSA.
[145] Article 23 of the DSA.
[146] Article 33 of the DSA.

Table 4
Mandatory particulars for transparency reports according to the size of the platforms
Source: DSA.

Intermediary services (generally) provide information on	**Online platforms provide information on**	**Very large online platforms provide information on**
the number of orders received from Member States' authorities, categorised by the type of illegal content concerned, including orders to act or provide information, and the average time needed for taking the action specified in those orders	the number of orders received from Member States' authorities, categorised by the type of illegal content concerned, including orders to act or provide information, and the average time needed for taking the action specified in those orders	the number of orders received from Member States' authorities, categorised by the type of illegal content concerned, including orders to act or provide information, and the average time needed for taking the action specified in those orders
the number of notices of illegal content, categorised by the type of alleged illegal content concerned, any action taken pursuant to the notices differentiated according to whether the action was taken on the basis of the law or the terms and conditions of the provider, and the average time needed for taking the action	the number of notices of illegal content, categorised by the type of alleged illegal content concerned, any action taken pursuant to the notices differentiated according to whether the action was taken on the basis of the law or the terms and conditions of the provider, and the average time needed for taking the action	the number of notices of illegal content, categorised by the type of alleged illegal content concerned, any action taken pursuant to the notices differentiated according to whether the action was taken on the basis of the law or the terms and conditions of the provider, and the average time needed for taking the action
the content moderation engaged in at the providers' own initiative, including the number and type of measures taken that affect the availability, visibility and accessibility of information provided by the recipients of the service and the recipients' ability to provide information, categorised by the type of reason and basis for taking those measures;	the content moderation engaged in at the providers' own initiative, including the number and type of measures taken that affect the availability, visibility and accessibility of information provided by the recipients of the service and the recipients' ability to provide information, categorised by the type of reason and basis for taking those measures;	the content moderation engaged in at the providers' own initiative, including the number and type of measures taken that affect the availability, visibility and accessibility of information provided by the recipients of the service and the recipients' ability to provide information, categorised by the type of reason and basis for taking those measures;
the number of complaints received through the internal complaint-handling system, the basis for those complaints, decisions taken in respect of those complaints, the average time needed for taking those decisions and the number of instances where those decisions were reversed	the number of complaints received through the internal complaint-handling system, the basis for those complaints, decisions taken in respect of those complaints, the average time needed for taking those decisions and the number of instances where those decisions were reversed	the number of complaints received through the internal complaint-handling system, the basis for those complaints, decisions taken in respect of those complaints, the average time needed for taking those decisions and the number of instances where those decisions were reversed
	the number of disputes submitted to bodies for the out-of-court settlement of disputes, the results of dispute settlement and the average time needed for completing dispute settlement procedures	the number of disputes submitted to bodies for the out-of-court settlement of disputes, the results of dispute settlement and the average time needed for completing dispute settlement procedures

Intermediary services (generally) provide information on	Online platforms provide information on	Very large online platforms provide information on
	the number of suspensions, distinguishing between suspensions enacted for the provision of illegal content, the submission of manifestly unfounded notices and the submission of manifestly unfounded complaints	the number of suspensions, distinguishing between suspensions enacted for the provision of illegal content, the submission of manifestly unfounded notices and the submission of manifestly unfounded complaints
	any use made of automatic means for the purpose of content moderation, including a specification of the precise purposes, indicators of the accuracy of the automated means in fulfilling those purposes and any safeguards applied	any use made of automatic means for the purpose of content moderation, including a specification of the precise purposes, indicators of the accuracy of the automated means in fulfilling those purposes and any safeguards applied
		reports giving the results of the risk assessment;
		associated mitigation measures
		audit reports
		a report on the implementation of the audit recommendations under the article

Another institution of transparency is the **publication of advertising information**, which applies to online platforms and, with modifications, to very large platforms. Article 24 of the DSA states for online platforms:

> 'Online platforms that display advertising on their online interfaces shall ensure that the recipients of the service can identify, for each specific advertisement displayed to each individual recipient, in a clear and unambiguous manner and in real time:
>
> a) that the information displayed is an advertisement;
> b) the natural or legal person on whose behalf the advertisement is displayed;
> c) meaningful information about the main parameters used to determine the recipient to whom the advertisement is displayed.'

The basic idea of this requirement is for users to understand who is showing them a particular advert and why. The general ban on personalised advertising, as communicated by the European Parliament in response to the draft DSA, seems idealistic.[147]

Furthermore, very large online platforms are required to make an advertising archive available through the API, showing specific information stored until one year from the last ad impression. Such an archive must not contain any personal data of the recipients. Article 30(2) stipulates that:

> 'The repository shall include at least all of the following information:
> a) the content of the advertisement;
> b) the natural or legal person on whose behalf the advertisement is displayed;
> c) the period during which the advertisement was displayed;
> d) whether the advertisement was intended to be displayed specifically to one or more particular groups of recipients of the service and if so, the main parameters used for that purpose;
> e) the total number of recipients of the service reached and, where applicable, aggregate numbers for the group or groups of recipients to whom the advertisement was targeted specifically.'

Last but not least, in the context of very large online platforms, the DSA requires **recommender system transparency**. The legal requirement primarily concerns disclosure of information on the parameters that the recommender systems use.

> 'Very large online platforms that use recommender systems shall set out in their terms and conditions, in a **clear, accessible and easily comprehensible manner, the main parameters** used in their recommender systems, **as well as any options for the recipients of the service to modify or influence those main parameters that they may have made available, including at least one option which is not based on profiling**, within the meaning of Article 4(4) of Regulation (EU) 2016/679.'[148]

Where a very large platform provides several options for users to use recommender systems, the platform interface must give users an easy option 'to select and to modify at any time their preferred option for each of the recommender systems that determines the relative order of information presented to them.'[149] It is assumed that these requirements for recommender systems will still be the subject of discussions at the level of the EU institutions as there is strong opposition from platforms in terms of the protection of intellectual property rights and trade secrets.[150]

---

[147] European Parliament. Report with recommendations to the Commission on a Digital Services Act: adapting commercial and civil law rules for commercial entities operating online (2020/2019(INL)). Available at: https://www.europarl.europa.eu/doceo/document/A-9-2020-0177_EN.html.

[148] Article 29(1) of the DSA.

[149] Article 29(2) of the DSA.

[150] BUITEN, M. The Digital Services Act: From Intermediary Liability to Platform Regulation. Working Paper. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3876328, p. 22. Cited from BERBERICH, M. - SEIP, F. Der Entwurf des Digital Services Act. (2021) 1 Gewerblicher Rechtsschutz und Urheberrecht, Praxis im Immaterialgüter- und Wettbewerbsrecht (GRUR-Prax) 4.

3.2.2.4 Audits

The DSA sets out specific requirements for the performance of an external audit at least once a year for very large online platforms. The audit's objective should be to assess compliance with the requirements for very large online platforms prescribed in the DSA or with the code of conduct that the platform has undertaken to comply with.[151] Audits are conceived as external, since they can only be performed by independent organisations that have sufficient professional capacity and have proven objectivity and ethics.[152]

The result of the audit should be an audit report which must include, in addition to the formal requirements, 'an audit opinion on whether the very large online platform subject to the audit complied with the obligations and with the commitments referred to in DSA, either positive, positive with comments or negative.' Where the audit opinion is negative, it must also include 'operational recommendations on specific measures to achieve compliance' with the DSA.[153]

Very large online platforms are required to take due account of the audit report and take the measures proposed therein. They are required, within one month of implementing those measures, to adopt an audit implementation report. 'Where they do not implement the operational recommendations, they shall justify in the audit implementation report the reasons for not doing so and set out any alternative measures they may have taken to address any instances of non-compliance identified.'[154]

[151] Article 28(1) of the DSA.
[152] Article 28(2) of the DSA.
[153] Article 28(3) of the DSA.
[154] Article 28(4) of the DSA.

4. Discussion and proposal of solutions

Part 2 of the study presented the operation of online platforms in terms of an economic model and the technical possibilities of optimising the displayed content. The conclusions to this section clearly indicate the need for the regulator to control the operation of online platforms in view of their economic strength, impact on society and the use of users as a product for increasing profits. Simultaneously, we have outlined and analysed the use of recommender systems, together with examples of their utilization by specific social media. As the key issues we have identified transparency of the generated recommendations and the setting of the utility function of recommender systems in the process of training artificial intelligence models. Other problems represent the enclosing of users in information bubbles, bias and fairness in recommendations, the collection of feedback, and deriving user privacy preferences.

In Part 3, we focused on analysing the legal framework proposed by the European Commission, which is devoted to the regulation of artificial intelligence (AIA) and online platforms (DSA). We specifically focused on institutions that have the potential to contribute to more effective tackling of disinformation on the internet in the context of the use of artificial intelligence.

In this part of the study, we put the conclusions from Part 2 in the legal framework presented in Part 3. We add our own recommendations for regulation to the discussion. We present our ideas and suggestions at a general level and specifically on the provisions of online platform audits, the transparency of artificial intelligence, reports on the activities of online platforms and the need for ethics risk assessment.

4.1 General comments on regulation

4.1.1 Attention economy as an area of high-risk AI systems

As we stated in the section concerning analysis of the AIA, the use of artificial intelligence by online platforms is currently generally not in any way covered by the proposed legislation. In the case of prohibited practices, we believe it will not be possible to ban the use of AI systems in the context of the proposed exceptions, since their interpretation remains very unclear and there is no examination of promotion of the manipulator's own goals in the definition of manipulative practices. Rather, the AIA requires real or potential harm to individuals for application of a ban.

In addition, the areas for high-risk AI systems do not cover the area of use of artificial intelligence by online platforms. The economic model of online platforms in the form of efforts to maintain user activity and attention may be legitimate, but it is by no means risk-free. The exposure of an individual to a large quantity of disinformation is also a major risk, which could result in very specific decisions with negative consequences for the individual's health and life.

Given the above, **we recommend that Annex III of the AIA arrange a separate area of 'attention economy' or define the use of AI systems for acquiring attention as a high-risk AI system application** under point 5 of Annex III 'Access to and enjoyment of essential private services and public services and benefits'.

We welcome the compromised text of AIA presented by the Slovenian presidency (partially) mitigating some of the issues mentioned above. General ban on private social scoring may soften the spread of disinformation online. However, if ban on private social scoring would be part of the legislation, AIA should also contain regulatory provisions on high-risk AI systems in case that scoring would not fulfil all requirements as a banned practice.

4.1.2 Regulation of illegal and harmful content

One of the most frequent comments about the DSA in connection with the dissemination of disinformation is that the vast majority of the DSA's provisions apply to illegal content. Simultaneously, it is clear that disinformation will not automatically always fall under the notion of illegal content. Exceptions are situations where content directly prohibited by law is disseminated (such as extremist material, denial of the Holocaust) or hate speech. The rest of the disinformation can 'only' be characterised as harmful content. We understand that analysis of the harmfulness of content can be problematic and complicated in certain cases.

At the same time, the proposed wording of the DSA counts on an analysis of hazardous, harmful content only in connection with the requirement of a risk assessment for very large online platforms, inter alia by assessing the settings of recommender systems. As we have already mentioned in the first part, the setting of the utility functionality and highlighting disinformation is one of the key problems. If a very large online platform identifies a risk of dissemination of disinformation, it should take mitigation measures, whether through content moderation mechanisms or through recommender systems. Simultaneously, these platforms are required to publish reports on the results of the risk analysis.

The position of the DSA's scope against illegal and harmful content is not even uniform at European Parliament level. The following examples can illustrate this:

- In its report, the Committee on the Internal Market and Consumer Protection (IMCO) defends the DSA's narrow scope over illegal content;
- On the contrary, the Committee on Civil Liberties, Justice and Home Affairs proposes a broader scope to include harmful content;
- In its report, the Committee on Legal Affairs (JURI) is against online platforms determining what is legal and what is not.[155]

[155] WINGFIELD, R. The Digital Services Act and Online Content Regulation: A Slippery Slope for Human Rights? Available at: https://medium.com/global-network-initiative-collection/the-digital-services-act-and-online-content-regulation-a-slippery-slope-for-human-rights-eb3454e4285d.

We understand that the direct competence of very large online platforms to assess and remove harmful content is very strong and can be exploited easily. Such an activity can very easily limit the exercise of freedom of expression and the right to information. Simultaneously, we reflect the state of the art, where even artificial intelligence is still unable to adequately recognise harmful content.[156] Despite this, it is too risky to pretend that the dissemination of disinformation is not harmful, since it can have a significant social impact. We feel that some kind of legislative intervention is justified.

For these reasons, **we consider that, as a bare minimum, the legislation should contain strict rules for cases where it is clear that an online platform emphasises and facilitates the dissemination of disinformation, either directly or indirectly**. Since it is inappropriate to directly order platforms to delete harmful content in terms of its identification and potential violation on freedom of expression, we believe that mechanisms that significantly limit the dissemination of disinformation shall be enshrined.

We consider as appropriate means:

- Transparency requirements and user choices in recommender systems (see Section 4.3 below);
- Labelling on unverified or unverifiable content (content labelling);
- Prohibition on promoting certain content or topics (such as promoting political messages or medicaments);
- Restrictions on the use of certain methods for sensitive content (such as bots or micro-targeting in political ads).

We welcome the explicit recognition of disinformation as systemic risk in recitals of the compromised text of DSA. However, tailored, and specific measures as outlined above are needed to target the spread of disinformation.

4.2 Audits and independent controls

4.2.1 Technological aspects of audits

As we showed in Section 2.3, changes in the algorithms of online platforms (e.g., in the setting of the utility function and its weighting) can often result in unintended negative impacts on users. Further risks in connection with the development and deployment of artificial intelligence are in the training data themselves, which can contain various biases and can result in unfair behaviour of AI systems towards individuals or user groups, but also in the feedback the system receives and uses to improve future predictions.

[156] MARSDEN, CH. - BROWN, I. - VEALE, M. Responding to Disinformation. Ten Recommendations for Regulatory Action and Forbearance. In Martin Moore & Damian Tambini (eds.) Dealing with Digital Dominance, OUP 2021, p. 205.

Therefore, it is a complex problem requiring complex solutions that include **testing** and **auditing**. Testing is a normal practice in software engineering, which occurs in software development at several levels, above all as:

- *Unit testing*, i.e., testing at unit level (e.g., of individual modules) of the software solution,
- *Integration testing*, which tests the interaction between individual modules.

The problem is that the development and deployment of AI systems includes training machine learning models using learning algorithms, which are often a **black box** (i.e., we know both their input and output format and their internal architecture, but their behaviour is based, particularly in the case of deep neural networks, on complex mathematical function with millions of parameters, which often cannot be expressed using rules that could be easily interpreted by humans).

Simultaneously, such solutions often suffer from '**technical debt**', when the individual phases of AI system development and operation are not sufficiently specified, tested and documented; therefore, methods have been proposed in the literature for evaluating the production readiness of an AI system.[157] It is also important to define test examples[158] that are easy to generate to allow detection of potential problems and biases in trained models outside the scope of a single, often distorting, performance measure for a given algorithm (such as accuracy, coverage, etc.; See Section 2.2.4).

The second approach to resolving the negative impacts of online platform algorithms is audits, which can also be performed at several levels. We can audit online platform **processes** during the development and monitoring of algorithms (e.g., at what level they are technically ready, what the processes are for identifying negative impacts and their solutions, etc.), but also (and above all) **data used for training, online platform algorithms and their behaviour during interaction with users**.

While the former most resembles a financial audit, the latter is rooted in social science audit studies. In the traditional sense, these were field experiments; the first such experiments were carried out by economists from the US Department of Housing and Urban Development to detect racial discrimination in access to housing.[159]

[157] BRECK, E. – CAI, S. - NIELSEN, E. – SALIB M.- SCULLEY, D. The ML test score: A rubric for ML production readiness and technical debt reduction, 2017 IEEE International Conference on Big Data (Big Data), pp. 1123-1132. Available at: https://doi.org/10.1109/BigData.2017.8258038.

[158] We can present, as an example, an approach designed to test the behaviour of natural language processing models: RIBEIRO, M.T. - WU, T. - GUESTRIN, C. - SINGH, S. Beyond Accuracy: Behavioral Testing of NLP models with CheckList. Association for Computational Linguistics (ACL), 2020. Available at: https://homes.cs.washington.edu/~marcotcr/acl20_checklist.pdf.

[159] SANDVIG, C. - HAMILTON, K. - KARAHALIOS, K. - LANGBORT, C. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry, a Preconference at the 64th Annual Meeting of the International Communication Association. May 22, 2014; Seattle, WA, USA., pp. 1-23. Available at: https://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf.

In the context of auditing algorithms and their behaviour, we can distinguish between several possible approaches:[160,161]

- **Source code audit** – the basic type of audit when the source code is examined. As we have argued above, this is not enough in itself, because the behaviour of online platform algorithms is determined by trained models. Simultaneously, this type of audit is problematic from the viewpoint of the intellectual property rights of online platforms. However, transparency at the level of the type of input data, utility functions and so forth is the necessary minimum (see Section 4.3).

- **Non-invasive auditing of users** – is based on interviews with platform users about their experiences and any problems during its use. However, this type of audit is only complementary, since it cannot reveal the causes of the observed behaviour of the platform (causal links).

- **Audit involving content scraping** – in this type of audit, the researcher enters queries to the platform algorithm (e.g., search queries) either using the programming interface provided (API) or by direct scraping of content from the platform's website, although this may violate the terms of service of online platforms if it happens without the knowledge of the platform owner. Another limitation is the fact that, in this case, the researcher does not simulate user behaviour, which limits the possibilities of auditing personalised systems, such as recommender systems on online platforms.[162] On the other hand, this approach allows content that would otherwise be provided by the platform owner in a non-transparent manner to be obtained independently.

- **Sock puppet audit** – in contrast to the previous type, in this case, the user behaviour is simulated using purposefully created profiles that the researcher can control (like a sock puppet master controls a sock puppet, from which it gets its name). The great advantage of this type of audit is the level of control the researchers (auditors) have over (simulated) user behaviour. This allows them to examine how the system responds to various types of user feedback, or how it behaves with various types of content. However, similarly, there could be a problem with the terms of service of online platforms, which generally prohibit creation of inauthentic profiles (bots) and more or less actively prevent it.

- **Collaborative (crowdsourcing) audit** – this involves the use of real users who agree to the recording of their behaviour for auditing purposes and are willing to perform assigned tasks that are the subject of the audit. This type most faithfully records user

[160] Ibid.

[161] BANDY, J. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 74 (April 2021), 34 pages. Available at: https://doi.org/10.1145/3449148.

[162] However, it is suitable, for example, for auditing non-personalised algorithms, e.g., for face recognition and analysis: RAJI, I.D. – BUOLAMWINI, J. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). Association for Computing Machinery, New York, NY, USA, pp. 429-435. Available at: https://doi.org/10.1145/3306618.3314244.

behaviour. On the other hand, it requires a sufficiently large representative sample of users and their active involvement during the required audit period (which is problematic in the case of audits of longer duration). Since real users are involved, this audit is also associated with greater risks to the personal data of the audit participants. None of these problems arise if the audit is performed as an internal audit by the online platform itself.

In addition, the audits can be either *internal*, i.e., performed by the provider of the audited algorithm or by an auditor having access to audited algorithm and data, or *external*, performed usually by independent researchers who can only observe the outputs, i.e., the behaviour of the audited algorithm.[163]

There are several examples of recent audits of online platforms in the literature.[164] We consider a sock puppet audit to be the best solution from the point of view of the possibility of controlling the audit conditions, its reproducibility and repeatability. From a technological point of view, the implementation of such an audit requires the following:

- **Programme bots (scripts) that will simulate user behaviour on the platform.** The interaction could be simple (e.g., watching videos on the YouTube platform, as we did in our audit study[165]) or more complicated, i.e., such that more faithfully simulates user behaviour (e.g., on the YouTube platform, it could be giving feedback on videos in the form of a 'like', clicking on one of the recommended videos, subscribing to the channel that is the author of the video, etc.). A new user on the platform (i.e., without previous history) or, alternately, a user with a certain history of behaviour on the platform can be simulated. Likewise, it is possible to compare the influence of a user's geographical location, gender and other attributes on the resulting recommendations.[166]

- **Record the content displayed (recommended) by the platform.** A link to the content (e.g., a video on YouTube) and linked metadata (e.g., number of positive ratings, number of comments, author, title and description of the video, etc.) can be recorded.

- **Evaluate the recorded content.** In the case of disinformation, the evaluation concerns whether or not the recorded content item supports a disinformation narrative. Most

[163] KNOTT, A. et al. Responsible AI for Social Media Governance: A Proposed Collaborative Method for Studying the Effects of Social Media Recommender Systems on Users. 2021. Available at: https://gpai.ai/projects/responsible-ai/social-media-governance/responsible-ai-for-social-media-governance.pdf.

[164] See, e.g., the collaborative audit of Facebook: SILVA, M. et al. Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook. In Proc. of The Web Conference (WWW '20). ACM, New York, NY, USA, 2020, pp. 224-234. Available at: https://doi.org/10.1145/3366423.3380109. An example of a sock puppet audit on the YouTube platform: HUSSEIN, E. - JUNEJA, P. - MITRA, T. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. Proc. ACM Hum.-Comput. Interact. 4, CSCW1, Article 048, May 2020, 27 s. Available at: https://doi.org/10.1145/3392854.

[165] TOMLEIN, M. - PECHER, B. - SIMKO, J. - SRBA, I. - MORO, R. - STEFANCOVA, E. - KOMPAN, M. - HRCKOVA, A. - PODROUZEK, J. - BIELIKOVA, M. An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes. RecSys '21: Fifteenth ACM Conference on Recommender Systems. Available at: https://doi.org/10.1145/3460231.3474241.

[166] HUSSEIN, E. - JUNEJA, P. - MITRA, T. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. Proc. ACM Hum.-Comput. Interact. 4, CSCW1, Article 048, 2020. Available at: https://doi.org/10.1145/3392854.

often, evaluation using trained annotators or crowdsourcing is used. Attempts at automatic annotation also exist.[167]

Such an audit can ascertain the prevalence of disinformation in recommendations on an online platform and compare how it changes over time, or what other factors affect it. However, comparison is only possible when the audit is repeated, which is currently rare due to the effort involved. It consists mainly of manual evaluation (annotation) of the recorded content; to illustrate, in our audit study cited above, we collected more than 17 000 unique videos, of which we manually annotated more than 2 900.

Thus, if we want it to be possible for similar audits to be performed on a regular basis (continuously, e.g., monthly), it is necessary to allocate considerable human resources or use automatic annotation methods. This entails several technical (and currently also research) problems, including guaranteeing and checking the fairness of the trained model for automatic annotation, guaranteeing and checking that the method will generalise for new disinformation narratives and its accuracy will not decline over time and, last but not least, more faithful simulation of user behaviour and maintenance of scripts to acquire content from the online platform. Another challenge is to differentiate *exogenous* influences (changes in platform content caused by external events, new content, etc.) from *endogenous* ones (changes caused by changes in the online platform algorithm).[168]

Technologically, audits of online platform algorithms and their behaviour are therefore feasible, even if they would benefit from greater automation and reduction of the human effort needed. Simultaneously, it seems that they are useful for analysing the behaviour of online platforms and detecting biases in algorithms, but also for independently checking the actual situation on online platforms and the obligations which the platforms themselves have committed to.

The current unclear legal status is the problem; researchers conducting the audit studies often go beyond the terms of service of online platforms. Furthermore, in the recent past, we have witnessed platforms punishing researchers who have pointed out their problematic behaviour by blocking their user profiles.[169]

---

[167] PAPADAMOU, K. - ZANNETTOU, S. - BLACKBURN, J. - DE CRISTOFARO, E. – STRINGHINI, G. - SIRIVIANOS, M. 'It is just a flu': Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations, 2020. Available at: https://arxiv.org/abs/2010.11638.

[168] SIMKO, J. - TOMLEIN, M. - PECHER, B. - MORO, R. - SRBA, I. - STEFANCOVA, E. - HRCKOVA, A. - KOMPAN, M. - PODROUZEK, J. - BIELIKOVA, M. Towards Continuous Automatic Audits of Social Media Adaptive Behavior and its Role in Misinformation Spreading. In Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21). Association for Computing Machinery, New York, NY, USA, 2021, pp. 411-414. Available at: https://doi.org/10.1145/3450614.3463353.

[169] These were NYU researchers who performed a collaborative audit of the advertising displayed to users (using an installed browser extension). Facebook blocked their accounts on the grounds that they had collected data outside the official platform tools and, simultaneously, that they had also compromised the privacy of users who did not consent to such recording. Available at: https://about.fb.com/news/2021/08/research-cannot-be-the-justification-for-compromising-peoples-privacy/. The reaction of the affected researchers is available here: https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html and here: https://www.wired.com/story/facebooks-reason-banning-researchers-doesnt-hold-up/. They protested that they had not collected any personal data, but only data about advertising and anonymised data about the users

We are in favour of **increased adoption of audits to monitor the behaviour of online platform algorithms**; both internal audits, although they may have their limitations, which are recognised by some online platform researchers,[170] and **independent external audits** performed by auditors according to clear rules (see Section 4.2.2). However, given the research nature of several problems associated with online platform algorithms, we believe that **there is a clear public interest in researchers being also able to use audit tools** under clearly defined conditions (see Section 4.2.3).

4.2.2 Legislation and recommendations

Given the technological aspects of conducting audits mentioned above and our previous research, we would like to make a number of recommendations for external audits to be included in the legislative proposals.

First, **we consider it appropriate for the European Commission to adjust the requirements for organisations that may carry out audits under the DSA**. In our opinion, the set criteria of independence, expertise and ethics of auditing are too broad and vague. We feel the process of certification of auditors directly by the European Commission through accredited bodies would be beneficial. EU law already regulates a similar process in the area of cyber security and personal data protection. The list of certified auditors should be made publicly available.

Second, **we consider it essential for the DSA to directly legitimise interventions on the rights and interests of online platforms during audits**. That means that an online platform could not deny auditors access to certain assets protected by personal data protection or intellectual property rights. For external audits to be truly effective and serve their purpose, authorised auditors should have the widest possible access to information. Simultaneously, it is appropriate for the terms of service not to prohibit the use of bots solely for the purpose of the sock puppet audits mentioned above.

We consider the **inadequate binding effect and enforceability of audit findings** to be a further shortcoming of the DSA proposal. Very large online platforms can substitute the recommendations of auditors with their own alternative measures. In our view, the formulation of the legal arrangements allows platforms not to take the findings of expert auditors into account and, ultimately, to ignore those findings.

Simultaneously, **we believe that the DSA should arrange the precise mechanism for examining the implementation of external audit findings**, for example through an automatic

involved (with their consent), and every one of them could check the functioning of their extension, because it had an open source code. Furthermore, Facebook did not technically or legally block their extension, but prevented their access to data that is accessible to other users and researchers.

[170] RAJI, I.D. - SMART, A. - WHITE, R.N. - MITCHELLI, M. - GEBRU, T. - HUTCHINSON, B. - SMITH-LOUD, J. - THERON, D. - BARNES, P. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, pp. 33-44. Available at: https://doi.org/10.1145/3351095.3372873.

check by the digital services coordinator or the competent authority or expert group of the European Board for Digital Services.

Requirements for auditing as stipulated in original draft of DSA are mostly intact in the compromised text. Additionally, we welcome the inclusion of broad access to data during auditing and authorization of auditors proposed by the European Parliament.

4.2.3 Status of scientific research

At this point, we consider it appropriate to emphasise the status of scientific research in the framework of auditing online platforms. The DSA emphasises the status of the academic sector through the institution of access to data, which can also be used by vetted researchers associated with academic institutions. **Again, we consider the vague criteria for identifying vetted researchers to be problematic and recommend a mechanism similar to certification with an approval process at EU level.** In general, we welcome provisions on process of award of status of vetted researcher in the compromised text of DSA. However, the requirement of "independency of commercial interests" may prove inadequate concerning privately funding research institutions without further explanation.

Simultaneously, **we are following with concern attempts to introduce restrictive measures on access to data for scientific institutions** by supplementing provisions that would allow information to be made inaccessible **for reasons of protection of intellectual property rights**. In our opinion, this is a situation that is resolvable through specific contractual arrangements, including contractual penalties and confidentiality obligations for researchers, instead of legal authorization to prevent access to data that may be important for research and understanding of systemic risks. Further requirements for research purposes may be part of codes of conduct.

4.3 Transparency requirements

Transparency is one of the key attributes for trustworthy artificial intelligence.[171] The given value can take several forms in relation to AI systems. Specific public awareness of the use of artificial intelligence, record keeping and traceability of processes (logging) that led to a decision and the explainability of specific decisions in specific cases. In Part 2 we have already outlined the fundamental problems of the transparency of the generated results and the use of the utility function in recommender systems in connection with the dissemination of disinformation. We consider that all three forms of transparency mentioned above have an important role to play in the context of the outlined problems.

From the legal perspective, awareness of the use of artificial intelligence is increasing, above all through various information obligations. Directly in relation to artificial intelligence, the

[171] High-Level Expert Group on Artificial Intelligence (AI HLEG). Ethics Guidelines for Trustworthy AI. Available at: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-en/format-PDF.

GDPR regulates such an obligation, where controllers must inform the data subjects of the **existence** of 'automated decision-making, including profiling, referred to in Article 22(1) and (4) [of the GDPR] and, at least in those cases, provide meaningful information about the **logic involved, as well as the significance and the envisaged consequences** of such processing for the data subject.'[172] The problem, however, represents Article 22 of the GDPR itself, to which the provision refers. Its application requires the fulfilment of several very unclear requirements (like the existence of a legal or similar effect in automated individual decision-making) and its very essence is still debated.[173] Therefore, it is common practice for controllers to assess the situation in such a way that, when using AI systems, Article 22 of the GDPR does not apply to them, so they do not even provide the relevant details.

For this reason, we welcome the information obligations concerning the display of advertising and the transparency of recommender systems for the very large online platforms presented in the draft DSA. Nevertheless, it is essential that the information on the parameters of the recommender systems used shall be provided in a meaningful way and not only generally. It is absolutely necessary to evaluate which parameters can lead to more effective dissemination of disinformation.

Simultaneously, we consider it necessary that, when the online platform is accessed for the first time, recommender systems are turned off by default. If users want to use this function, they will have to turn it on (opt-in). Alternatively, the recommendation function may be divided into smaller components (recommendation levels) defined by regulations, which users could turn on or off at their discretion, according to the preferred level of recommendation and personalisation. For example, a recommendation could consist of two levels: non-personalised and personalised, while users could, for example, turn on only the non-personalised.

In our view, this relatively simple technical solution will give users more autonomy from the first contact with the online platform and can reduce the display of harmful content. However, this rule should apply in such a way that from the moment when DSA will come into force, the platforms will be obliged to disable the recommender systems for existing users as well. We are aware of the inclusion of a new provision in the provisions governing recommender systems in the DSA in the compromised text. However, we are not convinced that obligation "not to subvert or impair the autonomy, decision-making, or choice of the recipient of the service through the design, structure, function or manner of operating of their online interface" is sufficient from the reasons stated above.

Log keeping will be required by the AIA in the scope of the data management provisions. It is the traceability of high-risk AI system decisions that is explicitly given in the draft AIA as the reason for keeping logs. Besides this reason, we consider it necessary for the AIA to classify the attention economy as a high-risk area.

[172] Article 13(2)(f) and Article 14(2)(g) of the GDPR:
[173] For example, TOSONI, L. The right to object to automated individual decisions: resolving the ambiguity of Article 22(1) of the General Data Protection Regulation. In International Data Privacy Law, Volume 11, Issue 2, April 2021, pp. 145-162.

If the original proposal in this matter is maintained, the traceability of AI system decisions in the private sector would, in principle, have no legal basis. In such case, we recommend that the final wording of the DSA include an obligation to keep logs related to recommender systems.

Apparently, the right to explanation of specific decisions by AI systems is not currently regulated in the legal system.[174] In our opinion, the current state of science and technology would be unable to cover such a legal obligation in many cases. **However, we should aim towards a situation that the current state of the art can cover, at least through information about the data processed, the parameters used and their values.** This information can be very valuable, particularly in tackling disinformation and setting up mechanisms to reduce its dissemination. We consider it right for users to know, at the very minimum, why specific content is displayed to them and to be able to adjust the parameters of displaying that content.

4.4 Reports on platform activities

Reports on platform activities reflect the principle of transparency and the public right to information about the activities of the online platforms. However, that information should be relevant and precisely specified by the DSA proposal. In addition, the DSA foresees that very large platforms will commit to comply with a code of conduct; in the context of disinformation, this concerns a code that arises by strengthening the Code of Practice on Disinformation. That code also imposes obligations regarding the reporting of activities connected with disinformation on very large platforms. The European Commission Guidance on Strengthening the Code of Practice on Disinformation emphasises the need for signatories to 'provide the information and data for the monitoring in standardised formats, with Member States breakdowns and on a timely basis'.[175] **We support reporting broken down by Member State.** In our opinion, the same principle should be introduced when providing information in the framework of platform activity reports provided for by the DSA. We welcome inclusion of the requirement in proposals of the European Parliament.

Simultaneously, we believe that the vast majority of requirements for platform activity reports relate to statistical indicators. In our opinion, **it would be beneficial for online platform reports to also provide information on specific solutions in specific cases**. To illustrate, we can mention reports on the state of personal data protection, which are published annually by the Office for the Protection of Personal Data of the Slovak Republic.[176] The given reports contain

[174] WACHTER, S. - BITTELSTATD, B. - FLORIDI, L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law, Volume 7, Issue 2, May 2017, pp. 76-99.

[175] European Commission. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS European Commission Guidance on Strengthening the Code of Practice on Disinformation. COM/2021/262 final. Available at: https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX%3A52021DC0262#.

[176] See e.g. the Office of the Slovak Republic for the Protection of Personal Data. Report on the state of personal data protection for 2020. Available at: https://dataprotection.gov.sk/uoou/sk/content/vyrocne-spravy.

specific cases and their analysis from the factual and legal perspective. Thus, they provide relatively accurate instructions for dealing with certain situations for personal data controllers. Simultaneously, the public has information on the steps taken by supervisory authorities in some cases.

In our opinion, the publishing of solutions to specific cases by online platforms would further strengthen the requirements of transparency and, simultaneously, present the strengths and weaknesses of content moderation on social networks.

4.5 Assessment from the ethical perspective

In the previous sections, we have referred to the reason why the operation of online platforms and activity of their algorithms should be subject to a certain level of intervention by the regulator. However, such regulation should come not only from states or international organisations, but also from the online platforms themselves, and its aim should be proactive identification of potential risks with regard to key societal values and ethical principles.

In the impact assessment of online platforms, it should be of primary importance to identify the various groups of persons or organisations that the platform's activities may affect. These need not only be the users themselves, or stakeholders who come into direct contact with the platform. They could also be indirect stakeholders, who often represent very vulnerable groups requiring our attention.[177] It should be emphasised that each of these stakeholder groups deserves the attention of online platform operators, regardless of whether they are direct users of the platform or even generate any profit for that platform.

The correct and timely identification of the affected stakeholders, and any social and ethical risks is not trivial and is also the subject of the 'Collingridge dilemma'.[178] Thus, a situation where, at the start of development of a technology, when we still have a great deal of control over its final form, our knowledge of its possible impacts on humans is very limited but, after its deployment, our ability to control this technology is limited.

Furthermore, the affected stakeholders, the extent of the risks, and also their level of impact may change over time, whether under the influence of technological development or the level of our knowledge of these risks.

It is therefore important for platforms to be able to devote **time and resources to the continuous evaluation** of possible ethical and societal impacts in terms of the **moral values and principles involved throughout the cycle**, from the design of new functionalities to their deployment.

[177] FRIEDMAN, B. - KAHN J, P. – BORNING, A. Value Sensitive Design and Information Systems. In: DOOR, N. et al. (eds). Early engagement and new technologies: Opening up the laboratory. Philosophy of Engineering and Technology, vol. 16. Springer, Dordrecht. Available at: https://doi.org/10.1007/978-94-007-7844-3_4.

[178] COLLINGRIDGE, D. The Social Control of Technology. New York: St. Martin's Press, 1980.

The question arises of whether it is in any way possible to define any set of universal moral principles and values that we should, as a society, follow.[179] As shown by recent research that has analysed dozens of major initiatives in the field of ethical and trustworthy artificial intelligence,[180] we can propose a set of certain universal values or principles in them which all these initiatives repeatedly address and emphasise. These values and principles include, for example, the previously mentioned transparency, but also fairness or privacy.

We can find similar concepts, for example, among the requirements for trustworthy artificial intelligence, as formulated by the High-Level Expert Group on Artificial Intelligence (AI HLEG) in the Ethics Guidelines for Trustworthy Artificial Intelligence (see Figure 10).

We suggest that AI ethics have undergone the phase of clarifying basic principles and values. Therefore, expert attention in this area is shifting to 'operationalisation', i.e. their translation into specific requirements that must be observed in the design, development, deployment and use of AI systems.

Currently, there are several different tools and methods for assessing artificial intelligence systems from an AI ethics perspective.[181] The most comprehensive include, for example, the **Assessment List for Trustworthy AI** (ALTAI), which is also the result of the activities of AI HLEG. This assessment list and some other tools for assessing the social impact of AI[182] are already attempting the abovementioned operationalisation of the values and principles we respect in our society and whose observance we consider key to the further functioning of democratic society.

[179] This is one of the major debates deep at the heart of philosophical research in metaethics. Although there are several strong arguments in favour of ethical relativism, we agree with some experts in a view that in principle we can determine such universal moral principles and values. See RACHELS, S. - RACHELS, J. The Elements of Moral Philosophy. McGraw-Hill Education, 2015.

[180] FIELD, J. et al. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center for Internet & Society, 2020.

[181] AYLING, J. - CHAPMAN, A. Putting AI ethics to work: are the tools fit for purpose? AI Ethics, 2021. Available at: https://doi.org/10.1007/s43681-021-00084-x offers a very good overview of dozens of instruments for assessment purposes from the perspective of AI ethics. Apart from summarising and categorising the most important AI system assessment tools, it also adds historical context to these tools from other areas (the environment, financial sector, etc.)

[182] Also worth mentioning are the procedures introduces by IEEE, which deal with the impact assessment of autonomous and intelligent systems for the human good, e.g. the IEEE 7010-2020 standard, available at: https://standards.ieee.org/standard/7010-2020.html or the data ethics impact assessment for public institutions in the UK. Available at: https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-2020.

Figure 10
Requirements for trustworthy artificial intelligence.
Source: Ethics Guidelines for Trustworthy AI (AI HLEG).

The presence of moral considerations in artificial intelligence and the effort at self-regulation of its systems can be of great help in view of the situation in which we find ourselves in the regulation of new technologies. In regard to the regulation of artificial intelligence we are currently in a state that the British philosopher James Moor has called a '**policy vacuum**'.[183] This is a situation where policies for the regulation of digital technology are still being developed and, often, even the key concepts necessary for discourse are yet to be fully established. It is understandable that in such unstable situation there will be calls for more moderate regulation of artificial intelligence and stronger self-regulation in order to bridge these intermediate policy vacuums and periods of uncertainty.

However, regulators are not making sufficient use of this opportunity. As we have also stated in our position on the draft Artificial Intelligence Act (AIA),[184] the text of the draft covers ethical implications only in a non-binding form of substantiation and in the preamble. We feel this is a wasted opportunity in view of the work done by AI HLEG in this area. **It is our belief that ethics risk assessment should be considered as a binding part of the conformity assessment proposed in the AIA for AI system providers.** Well-chosen tools for ethics assessment of AI

[183] MOOR, J. What is Computer Ethics? Metaphilosophy 16(4), 1985, pp. 266-275.
[184] Kempelen Institute of Intelligent Technologies. Stance on the Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Available at: https://kinit.sk/wp-content/uploads/2021/09/KINIT_Stance-of-AIA_Paper_2021_09.pdf.

systems, which already work in practice,[185] can thus help to bridge the previously mentioned policy vacuums and the periods during which the draft regulations are not yet fully in effect.

On the other hand, one of the risks of over-reliance on self-regulation is that it will only be carried out formally, as we can also see in the case of internal audits of online platforms and their activity reports. Such affected self-regulation of digital technologies is termed *ethics bluewashing*.[186] Therefore, as in the case of greenwashing, when a company only pretends to act in the interests of environmental protection, a situation may arise where companies will only pay lip service to ethical and social reflection in order to avoid direct regulation by the state.

Then, the model of '**hybrid regulation**' seems ideal, where several regulatory instruments act simultaneously on the entities and persons concerned.[187]

4.6 Conclusions

If we take into account the partial conclusions presented in Part 2 of the present study and the specific examined legislative proposals from the European Commission (mainly the proposal for regulation on artificial intelligence and the proposal for regulation on digital services) presented in Part 3, we consider that these legal acts can be improved in such a way as to provide more effective tools to tackle dissemination of disinformation.

In terms of general comments, **we consider it key for the area of attention economy to be considered as an area of high-risk AI systems**, so that the use of algorithms by social media does not escape the legislative requirements presented in the AIA.

Simultaneously, we understand the limits of the regulation of harmful content in terms of interference with the freedom of expression and granting too much power to online platforms. On the other hand, legislation can provide a broad palette of tools that have the potential to limit the dissemination of disinformation on online platforms without directly regulating harmful content.

As examples of the introduction of such 'indirect' rules, we recommend:

- Transparency requirements and user choices in recommender systems;
- Labelling on unverified or unverifiable content (*content labelling*);
- Prohibition on promoting certain content or topics;
- Restrictions on the use of certain methods for sensitive content (such as bots or micro-targeting in political ads).

[185] The interactive version of the ALTAI tool is available publicly and free of charge, and it already provides relatively decent user comfort https://altai.insight-centre.org/.
[186] FLORIDI, L. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. Philos. Technol. 32, 2019, pp. 185-193.
[187] For example, we can mention the four modalities of the regulation of social relations from L. Lessig, where all modalities are mutually complementary and work side by side (law, ethics, the market and architecture/code). LESSIG, L. Laws of Cyberspace. Available at: https://cyber.harvard.edu/works/lessig/laws_cyberspace.pdf.

We consider the performance of external audits absolutely essential. Although the DSA directly arranges such a mechanism, we are concerned about its limits regarding the lack of clear rules for identifying suitable entities for carrying out those external audits, the insufficient binding nature of the results of such audits and the implementation of their conclusions. At the same time, the access of external auditors should not be restricted in terms of protecting the rights of online platforms, such as trade secrets, or of third parties in the form of personal data protection. In technical terms, we recommend a 'sock puppet audit' as a suitable form of audit, but the above comments need to be incorporated in such a way that the use of bots is possible for external audit purposes in terms of legislation and the terms of service of online platforms. We also emphasise the role of scientific research and access to data by vetted researchers.

Another important area in tackling disinformation is transparency. In this regard, we welcome the proposals presented in the DSA concerning the transparency of advertising and recommender systems or publicly available reports. However, these mechanisms must ensure that their results are meaningful and illustrate the real situation to the professional and lay public. **Regarding the transparency of recommender systems, we direct our attention to the fact that the DSA should enshrine a mandatory opt-in for users on first contact with the online platform in any form.** For example, recommendations could consist of two levels: non-personalised and personalised, while the user could, for example, turn on only the non-personalised recommendations, which does not take the user's behaviour, i.e., the source for estimating preferences, into account in the recommendation. At the same time, the keeping of logs for recommender systems should be legally stipulated in the DSA for cases potentially not covered by the AIA.

Transparency reports should be complemented by case studies to clarify how the online platform behaves in specific situations and what mitigation measures are being taken. Simultaneously, the statistical indicators in these reports should be divided by Member States so it is possible to detect any differences in the dissemination of disinformation content between Member States, and in the approach of platforms and Member States' authorities and the measures taken.

It is no less important for social media to regularly evaluate not only legal, but also ethical and societal risks with respect to different groups of direct and indirect stakeholders. We consider it essential for platforms to be able to devote time and resources to the continuous evaluation of possible impacts in terms of the moral values and principles involved throughout the cycle, from the design of new functionalities to their deployment.

It is our belief that ethics risk assessment should be considered a binding part of the conformity assessment proposed in the AIA for AI system providers.

Authors

Matúš Mesarčík

Specialist in ethics and law
contact: matus.mesarcik@kinit.sk

Róbert Móro

Researcher in artificial intelligence and machine learning
contact: robert.moro@kinit.sk

Michal Kompan

Researcher in artificial intelligence and machine learning
contact: michal.kompan@kinit.sk

Juraj Podroužek

Researcher in the ethics of digital technologies
contact: juraj.podrouzek@kinit.sk

Jakub Šimko

Researcher in artificial intelligence and machine learning
contact: jakub.simko@kinit.sk

Mária Bieliková

Researcher in artificial intelligence and machine learning
contact: maria.bielikova@kinit.sk

Kempelen Institute of Intelligent Technologies
Mlynské Nivy II. 18890/5, 821 09 Bratislava,
Entity reg. no.: 53290046, contact: info@kinit.sk