EDMO
European Digital Media Observatory

EUI SCHOOL OF TRANSNATIONAL GOVERNANCE

European Digital Media Observatory

Defining Disinformation across EU and VLOP Policies

Konrad Bleyer-Simon
Urbano Reviglio

October 2024

Defining Disinformation across EU and VLOP Policies

Abstract 3

Introduction 4

1. An EU approach to disinformation 5

1.1 EU National Legislations: When Disinformation Overlaps with Illegal Content 9

1.2 The Digital Services Act 10

1.3 The Code of Practice on Disinformation 11

2. (Very Large) Online Platforms 12

3. Discussion 18

4. Conclusion 22

References 24

Bibliography 25

Annex 1: Platforms' terms, definitions, criteria, categories and related concepts (as of 31 May 2024). 29

VLOP signatories of the CoP 30

Non-VLOP/VLOSE signatories of the CoP 36

VLOP, not signatory of CoP 40

Annex 2: Platform policies that refer to disinformation and related phenomena. 43

Annex 3: Peukert' categorisation of disinformation and related concepts in the 2022 Strengthened Code of Practice on Disinformation 45

The European Digital Media Observatory has received funding from the European Union under contract number LC-01935415

Defining Disinformation across EU and VLOP Policies

Abstract

Defining disinformation is essential to propose concrete measures to identify and, possibly, moderate harmful online content. In this article, we underline the ways in which disinformation and related concepts are defined in relevant European policies, focusing, among other things, on the Digital Services Act and the Code of Practice on Disinformation at the EU-level, national legislation, as well as the policies and community guidelines of major Very Large Online Platforms (VLOPs). By analysing these different, yet intertwined, policies, we identify the approaches that guide VLOPs' actions related to information manipulations, as well as the challenges and opportunities of defining disinformation and thus operationalising its governance at the European level. We observe that the dominant definitions of disinformation in the EU stipulate that, in order to be considered disinformation, content needs to contain (1) verifiably false or misleading information, (2) have a potential to cause harm to society, (3) must be intentionally spread (4) for possible economic or political gain. However, platform policies are in most cases not in line with all components of these definitions, as they are, among other things, unwilling to consider the component of intent, and may not be clear on their definitions of harm. This can lead to both under- and over-regulating certain aspects of the problem, and thereby potentially limiting users' freedom of expression as well as threatening information integrity.

Introduction

There are many problems that emerge in defining disinformation and eventually operationalising its moderation. First, and most obviously, publishing or sharing disinformation or other untrue content is in most cases not illegal – a person's right to free expression allows them to communicate falsehoods or present facts in a misleading way. Therefore, removing, restricting or blocking such content can violate fundamental human rights, not to mention that there is still no scientific consensus about the extent and forms of harm that the sharing of false and misleading content can cause (see Altay et al. 2023, Budak et al., 2024): from eroding trust in science and politics to mental health issues and problematic individual behaviours, scams, defamation, among others. Nevertheless, after Brexit, the Covid-19 pandemic and reports about attempted Russian interference in elections, disinformation is high on the political and policy agenda in Europe; and it is widely accepted in policy circles that the possible risks information manipulation may pose to the management of crises, as well as the functioning of democracies, require action. This, logically, leads to a trade-off: overly broad definitions of the kinds of content that require action (or disinformation per se) might risk stifling legitimate discourse, whereas overly restrictive definitions could permit potentially harmful deception with undesirable consequences such as negatively influencing individual and collective decision-making processes.

Secondly, and related to the previous considerations, a definition needs to take into account the intention behind the sharing or publishing of disinformation. Distinguishing intentional (disinformation) from inadvertent deception (misinformation) can be challenging, if not impossible, in many cases. The boundaries between disinformation, misinformation, and even satire, propaganda and hate speech, can be nebulous. The task of defining disinformation therefore involves a tension between determining what can be considered fact, identifying standards to be followed to make sure content can be considered factual, and, at the same time, allowing for subjective interpretation. This, however, might be influenced by personal biases and beliefs. Determining what constitutes disinformation is also context-dependent: a statement that is considered a truthful interpretation of facts in one context might be misleading or even harmful in another (for example, a critique of the medical profession or the public health system might have a different meaning and impact during a pandemic than in ordinary times).

In addition, it needs to be highlighted that the landscape of disinformation is constantly evolving, with new tactics emerging continuously; with this evolution, the definitions need to adapt as well. Deepfakes, AI-generated content, and other advanced methods challenge traditional concepts, requiring regular updates to definitions, in order to capture emerging forms of deception. This also means that aiming for a static and universally applicable definition may not be feasible, and perhaps not even desirable. Still, in order to enable meaningful action, there needs to be a consensus between different parties on what exactly a policy should aim for.

Definitions, as well as concrete rules and measures to mitigate the effects of what can be considered disinformation are to be found in laws and policy documents on the national and the EU-level, as well as in the terms of services and community guidelines of private online platforms – both VLOPS and other online platforms in general. Thus, on the following pages, we will focus on EU documents, such as the voluntary commitments

of the self- (or co-regulatory) Code of Practice on Disinformation (CoP or Code), the risk mitigation measures of the Digital Services Act (DSA), briefly touch on some aspects of national regulation and assess the policies of some of the largest online platforms covered either by the CoP or the DSA.

1. An EU approach to disinformation

The European Union has been at the forefront of designing policy to tackle disinformation while seeking to safeguard fundamental rights – and it has been advocating for a European approach in order to avoid a fragmented European policy landscape in light of a border-crossing problem (Nenadić, 2019, EC, 2018a). The EU's approach (in the text, we refer in most cases to the European Commission, referred to hereafter as EC) to tackling, in particular online, disinformation rests on the notion that legal content, even if it might be considered harmful 'is generally protected by freedom of expression and needs to be addressed differently than illegal content' (EC, 2018b:1) in the latter case, the removal of content is less problematic.

Within the European context, there seems to be a convergence towards three influential definitions (Ó Fathaigh et al., 2021): Wardle & Derakhshan (2017), the High Level Expert Group on Fake News and Online Disinformation (HLEG, 2018) and the European Commission's Code of Practice on Disinformation (EC, 2018b). Firstly, in 2017 Wardle & Derakhshan designed an interdisciplinary framework for research and policymaking on 'information disorder[s]' for the Council of Europe that developed one of the most well-known definitions of disinformation, and related concepts.[1] On top of the previous definition, there have been definitions both by the EC[2] and the HLEG[3] (see also Abbamonte & Gori, 2022 and Pollicino & Bietti, 2019).

[1] Wardle & Derakhshan differentiate between three categories of information disorders:

- 'Disinformation: Disinformation is false information that is deliberately created or disseminated with the express purpose to cause harm. (Producers of disinformation typically have political, financial, psychological, or social motivations.)
- Misinformation: Misinformation is information that is false, but not intended to cause harm. (For example, individuals who don't know a piece of information is false may spread it on social media in an attempt to be helpful.)
- Malinformation: Malinformation is genuine information that is shared to cause harm. (This includes private or revealing information that is spread to harm a person or reputation.)'

[2] EC's definition: '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 policymaking processes as well as public goods such as the protection of EU citizens' health, the environment or security.' originally from the Commission Communication, 'Tackling online disinformation: a European approach', COM(2018) 236 final of 26 April 2018.

[3] 'We define it as false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit. The risk of harm includes threats to democratic political processes and values, which can specifically target a variety of sectors, such as health, science, education, finance and more.'

All in all, there are four elements that, to a certain extent, are common to these different definitions (Ó Fathaigh et al., 2021): (1) the falsity or misleading nature of the information, (2) the potential for social harm, (3) the intention of the actor and (4) the possible economic gain. A report by the European Regulators Group for Audiovisual Media Services (ERGA) adds two relevant elements: (5) the information relates to a matter of public interest and (6) the information is strategically disseminated (Betzel et al., 2021, p. 18). Finally (7), we need to mention that most documents put an emphasis on disinformation content that relates to elections or other democratic processes (for example CoP, DSA, Paris Call for Trust & Security in Cyberspace), and especially since the Covid-19 pandemic, crises and national emergencies (CoP and DSA). While the first three elements can be considered necessary conditions for disinformation to fall under the category that requires action, elements 4-7 are characteristics that are often observed in the case of disinformation, but are not necessary to the definition.

1) Factuality

With regard to the factuality of the information, Wardle and Derakhshan refer to 'information that is false'. The EC, instead, specifies this by referring to 'verifiably false' information but, simultaneously, expands the definition by including 'misleading information' – the HLEG similarly widens the definition to include 'inaccurate' information.

2) Harmfulness

With regard to the harm created by disinformation, Wardle and Derakhshan develop a wider definition including harm to 'a person, social group or country' while the EC and the HLEG both only refer to 'public harm'. This public harm is, subsequently, defined by the HLEG as 'threats to democratic political processes and values, which can specifically target a variety of sectors, such as health, science, education, finance and more'. The EC aligns itself with this definition, only adding 'policymaking processes'. Notably, all three definitions share that harm does not need to have actually occurred for the information to be qualified as disinformation. Harm or potential harm is a necessary condition, as without it, action taken against a category of content that is not illegal would be hard to justify.

3) Intentionality

Intent is an important component of all definitions – distinguishing mis- from disinformation. In addition to (the potential for) harm, it is the other component that can, under certain circumstances, justify action against such content, as it signals that the actor behind the communicative act is purposefully misusing the right to freedom of expression. However, as already mentioned at the beginning of this text, proving an intent to mislead can be a complicated task. This is, for example, the reason why TrustLab, a third-party organisation tasked by the European Commission with a beta impact assessment of the Code of Practice on Disinformation through so-called structural indicators (see Nenadic et al., 2023 & 2024), opted to use the term 'mis/disinformation' to avoid making judgments about the intent of actors sharing content that can be

considered information manipulation.[4] At the same time,[5] TrustLab's report mentions that there are some signs that make it more likely that a piece of content is misleading on purpose; these can be related both to characteristics of the publisher and the published piece: 'repeat activity, size of the follower network, manipulation of images, video, or audio clips, the deliberate use of misleading headlines, or clickbait as a way to attract attention and promote false narratives' (Trustlab, 2023:12).

4) Economic motives

(4) The element of economic gain does not appear in the Wardle and Derakhshan definition, while it plays a very prominent role in the definition of the EC and the HLEG. This aspect of disinformation gained prominence after news broke of 'Macedonian teens' (Subramanian, 2017; Hughes & Waismel-Manor, 2021) who managed to capitalise financially on the increased traffic that made-up stories related to the 2016 US election campaign generated – it highlights an aspect that is widely known from older assessments of the online news environment: controversial topics and catchy titles are preferred by many online publishers, as they are believed to be more likely to attract clicks and advertising revenue than well-written and well-researched articles (as the literature on 'click-baits' shows, see Kertanegara, 2018 & Bazaco, et al. 2019). This aspect is relevant, as it shows that disinformation is about more than just actions of specific (foreign and domestic) interest groups, but has a potential for economic viability – and thus some forms of disinformation can be preemptively targeted through demonetisation. At the same time, it also has to be highlighted that not all disinformation pieces aim for monetary rewards, and even the case of those that do so, the different aspects of 'monetisation' are hard to capture, as advertising is just one of the many methods used by disinformation actors to generate revenues (others may include asking for donations or selling goods). Moreover, as it can be seen in the text of the CoP as well, monetisation is a two-way street: the platforms hosting disinformation can also become beneficiaries of this kind of content. Indeed, on the one hand, they receive a cut of the advertising revenue generated, while, on the other hand, some disinformation actors may opt to pay platforms to amplify their content (Cunningham et al., 2024; Diaz-Ruiz, 2023; Bleyer-Simon, 2024). Following the publication of the European Democracy Action Plan, the 2022 Strengthened Code of Practice on Disinformation added 'political gain'[6] as a possible complement to economic incentives – this additional aspect becomes relevant in connection with the next point: the choice of topics addressed by disinformation content.

5) Public interest issues

[4] Later in this text, we will deal with the 2022 Strengthened Code of Practice's concept of the capitalised 'Disinformation' which includes misinformation, information influence operations and foreign interference.

[5] However, their report mentions financial incentives as a possible driver of information manipulations: 'There are four potential motivating factors: Financial: Profiting from information disorder through advertising; Political: Discrediting a political candidate in an election and other attempts to influence public opinion; Social: Connecting with a certain group online or off; and Psychological: Seeking prestige or reinforcement' (Wardle & Derakshan, 2017:26). Moreover, the text asks tech companies to eliminate the kinds of financial incentives that encourage the publishing of information manipulation.

[6] The two categories are not mutually exclusive, as the research by the Global Disinformation Index has shown that some well-known political players, like RT or Sputnik have also successfully monetised their content.

In many cases, disinformation concerns issues of public interest – as their publishers aim to capture the attention of a significant number of people, while also influencing societal processes. In most cases, these are related to democratic processes, such as elections or emergencies, as pointed out in (7). At the same time, it is also possible to intentionally spread untrue and potentially harmful content on other kinds of topics, such as celebrity gossip, in which the public interest aspect can be disputed.

6) Strategic dissemination

Similarly to the previous point, it can be assumed that in most cases actors who intentionally publish harmful untrue content do so strategically. However, not all publishers of disinformation are trained and effective communicators with clear ideas how to reach the intended audiences. While pieces of disinformation that policymakers are most concerned about are indeed the ones that reach wide audiences and have an impact on beliefs in society, one can also find purposefully published, potentially harmful pieces of content that failed to reach significant audiences.

To conclude this section, it is clear that the HLEG and the EC have narrowed the scope of their definitions of disinformation (e.g., by only considering potentially harmful content) and separated them from existing legal categories. The HLEG stated this most firmly in declaring that the concept of disinformation does not overlap with any existing legal norm. The distinction that EU policy seems to make between, on the one hand, disinformation as (potentially) harmful content and, on the other, already regulated forms of illegal content, does limit the scope of the concept. However, it also risks missing ways in which enhanced enforcement of already existing legal norms could contribute to limiting the spread of disinformation – as a specific piece of content may combine components of disinformation with illegal forms of communication, such as incitement to violence or holocaust denial.[7] In addition, we also need to mention that, despite a narrowing of the scope of pieces of false and misleading content, the three definitions can still be considered exceedingly broad for policy action, which leaves ample room for EU member states or online platforms to interpret these terms broadly and, thus, in different and arbitrary ways.

In the EU context, a range of policies are used to deal with disinformation – many of them are soft measures, such as promotion of media literacy and support of quality journalism and fact-checking. At the same time, the Digital Services Act (DSA) and the Code of Practice on Disinformation (CoP or Code) define action that should be followed by online platforms that signed it – in part to avoid having a fragmented and at times conflicting set of responses to a cross-border phenomenon –, but due to the outlined characteristics of the concept, those requirements leave a lot of leeway for platforms. In the following

[7] One can also mention the controversial EU sanctions on Russian-origin outlets, based on the Council Decision (CFSP) 2022/351 of 1 March 2022 amending Decision 2014/512/CFSP, integrated in the Council regulation (EU) 2022/350 of 1 March 2022 amending Regulation (EU) No 833/20144. This, and following sanctions on Russian news media asked for the blocking of content by specific outlets (chiefly RT and Sputnik) that are widely considered purveyors of disinformation, which meant that online platforms were legally mandated to act against content of certain publishers – but in this case, as in the cases when illegal content is removed, the justification included no reference to the falsity of content. (See Bleyer-Simon et al. 2022).

chapters, we will zoom in on national approaches, as well as the two pieces of EU (self-)regulation (DSA and the Code), followed by the platforms' own policies – in order to see how these different levels interact when it comes to defining a phenomenon that is outside the categories of clearly illegal content but can nevertheless be considered as a possible threat to society.

1.1 EU National Legislations: When Disinformation Overlaps with Illegal Content

According to a comparative assessment from Ó Fathaigh et al. (2021) and the results of a survey of legislation in EU member states conducted by the European Regulators Group for Audiovisual Media (ERGA) (Betzel et al., 2021), many EU member states have national provisions – including criminal legislation – that overlap with the notion of disinformation (even if not using the term itself). It is not difficult to realise how the disparities in national approaches can create considerable legal uncertainty and obstacles for any provider of digital content services in the European Union. National regulatory approaches differ considerably in terms of scope, addressee, and legal sanctions. For example, in the context of the Covid-19 pandemic, a 2020 Presidential Decree in Romania made it possible for the communications regulator to order the removal of content that is seen as 'promot[ing] false news', while a temporary, and since repealed, change in the Hungarian Criminal Code made the publishing of certain kinds of dis- and misinformation punishable by imprisonment (see Polyák, 2020, Ó Fathaigh et al., 2021, Bleyer-Simon, 2021). Some EU member states have passed policies aimed at regulating the conduct of online actors, the most well-known and influential being Germany's NetzDG (Mchangama & Alkiviadou, 2020). However, despite being discussed as a possible anti-disinformation measure (see Claussen, 2018), it is more appropriate in the context of illegal content.

In particular, when looking at some of the laws that were in place in the early 2020s, there are varying specific harms mentioned, including economic, public, personal harm, personal dignity, harm to election integrity, and harm to public health measures. In some cases, legal texts refer to 'false news', but there is also mention of 'slanderous noises or other fraudulent manoeuvres' (France) that are, according to most definitions, shared purposefully – while the Hungarian penal code mentions, instead of intent, the condition of 'reckless disregard for [the content's] truth or falsity' – broadening the applicability of the law to people who did not know about the shared content not being factual. The possible harm is usually seen as the effect of disturbing 'public order' or 'public peace' (Betzel et al., 2021), influencing voting behaviour (France[8]), forcing people to take unplanned measures (Greece) or decreasing one's trust in the state (Cyprus). In certain cases, the text implies that the sharing of untrue content needs to affect a group of people or a proportion of a certain population. However, precise definitions are missing. Especially in cases when disinformation or related concepts were covered in the penal code, critics pointed towards the lack of clarity of definitions that can make the use of such laws arbitrarily. This is especially worrying given that sometimes the laws prescribe sentences or fines for offenders.

[8] Ó Fathaigh et al. (2021) also mention Austria, where Art. 264 of the Criminal Code makes the dissemination of 'false news' punishable with up to 6 months imprisonment.

In the sample used for ERGA's assessment, only one EU Member State has a statutory definition for disinformation namely Lithuania, which defines disinformation as 'intentionally disseminated false information'. The definition used in Article 19 of the Law on the Provision of Information to the Public partly aligns with the EC's and the HLEG's definition, as it includes the main conditions of falsity, intention and harm. However, as Betzel et al. (2021:34) pointed out, the Lithuanian legislation 'is limited to causing harm to a specific person, does not include harm to a social group, organisation, or country; and there is no requirement of economic gain or profit-motive' (Betzel et al., 2021, p.34).

A number of Member States have legislation that aligns with the EC's, HLEG's, as well as Wardle and Derakhshan's definition of disinformation, while not specifically using the term disinformation. Instead, the most common legislative terms for 'information that is false and deliberately created to harm a person, social group, organisation or country' are laws on 'false news' and 'false information'. This observation echoes the mentioned ERGA report, which noted that false information and disinformation are 'different ways to indicate the same concept,' and the European Commission's finding in June 2020 that 'Several Member States already had provisions, including of criminal nature, related to disinformation' (EC, 2020b). When it comes to defining possible action against disinformation on online platforms, this situation might contribute to the complexity of the problem: even if the platforms have their own definitions and concepts used for certain pieces of disinformation content (which, as we will see later, not just national laws, but also online platform policies might refer to using different terms), their content moderators will need to apply different forms of treatment for a subset of them, in line with the requirements set for illegal content.

1.2 The Digital Services Act

Regulatory differences at the national level and the push to limit the harm caused by specific forms of disinformation have potentially far-reaching implications for fundamental rights. In response to the emergence of divergent approaches, the European Commission adopted a landmark piece of EU legislation: the Digital Services Act, which imposes a whole set of new 'uniform rules' for digital platforms to harmonise different national approaches thus making it applicable in all 27 EU member states. The Regulation also introduces a new classification, associated with extra duties, for platforms and search engines that have more than 45 million users per month in the EU: they are labelled as very large online platforms (VLOPs) or very large online search engines (VLOSEs).[9]

In the DSA, the concept of 'illegal content' is central. Article 3 of the DSA defines 'illegal content' as 'any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law'. Due to its reference to compliance with Member State law, this definition includes all those instances of disinformation that are considered illegal by national laws – including some examples from the previous chapter, such as false or misleading communications that might influence voter behaviour in the French or Austrian context. While the DSA's provisions do not specifically mention, nor define, disinformation (or related concepts), the new rules, among other things, on systemic risks

[9] For an updated list see: https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses

(Arts. 34 and 35), ad libraries (Art. 39), and codes of conduct (Art. 45), clearly envision being applicable to the kind of disinformation that is not considered illegal in Member States, in cases in which they can be considered sources of harm. While disinformation is not defined in the text of the DSA, there are preceding EU policies, such as the European Democracy Action Plan, in which the term is defined, and can therefore apply in this case. Whether the Code's definition (once it turns into a Code of Conduct) will become the definition reference under the DSA remains to be seen.

The EU's regulatory approach is based on operators' duties of care. As such, the DSA doesn't provide clear and specific rules for platforms on how they need to deal with disinformation, but rather sets out a general requirement to mitigate the risks posed by possibly harmful content – including disinformation. This lack of clear rules, of course, leaves some leeway for online platforms, when it comes to deciding what needs to be acted on and how. More detailed requirements can be found in the Code of Practice on Disinformation which outlines a set of commitments for platforms to make their services safe for their users – some of which all large platform signatories are expected to follow. While the Code is currently only a set of voluntary commitments, it is set to evolve into a Code of Conduct under the DSA. Once this evolution takes place, the Code can be considered a description of clear action for platforms, by which their risk-mitigation efforts can be assessed. Griffin and Vander Maelen (2023) point out that the codes of conduct under the DSA facilitate the implementation and enforcement of the DSA by concretising its provisions and thereby creating de facto obligations for those platforms and search engines that fall under the category of VLOPs or VLOSE.

1.3 The Code of Practice on Disinformation

The Code of Practice on Disinformation is a self-regulatory code (evolving towards so-called 'co-regulation') with the participation of the leading online platforms, representatives of the advertising industry, fact-checkers as well as civil society organisations with an interest in action against disinformation (who are signatories of the Code).

Contrary to the DSA, the CoP already had a definition of disinformation in its early formulation in 2018, which was seen as the basis of action for signatories. The primary focus of the first CoP of 2018 was based on the definition of the Commission communication 'Tackling online disinformation: a European approach', and in line with the HLEG definition of disinformation – but, in addition to the forms of disinformation defined by the earlier outlined components of falsity, harm, intention and possible economic gain,[10] the text was also considering certain strategies of publishing and spreading content, such as coordinated efforts and the use of bots (under the pillar 'integrity of services'), as well as political and issue based advertising.

With the European Commission Guidance on Strengthening the Code of Practice on Disinformation in 2021 and the Strengthened CoP of 2022, the scope of the document was extended, and the term Disinformation (with capital D) included, besides

[10] 'Political gain' was only spelled out in later versions of the text.

disinformation, three additional forms of information manipulation: (certain forms of) misinformation, information influence operations and foreign interference.[11]

The Code assigns a number of commitments to its signatories – including measures that increase the transparency of their services, introduce safeguards against the misuse of monetisation tools, and contribute to online audiences' resilience towards information manipulation, among other things through increasing media literacy and supporting fact-checking. Moreover, it introduces key performance indicators to assess the Code's effectiveness – both on the service and at the structural level. Service-level indicators require platforms to report on their actions taken in connection to disinformation on their services – and requires them to have a working definition in their reports to identify what content falls into this category; this will be assessed, together with the relevant platform policies, in the next section. Under Commitment 41 of the Code, signatories commit to cooperate with experts and stakeholders on developing structural indicators, designed to assess the effectiveness of the Code in reducing the spread of online disinformation for each of the relevant Signatories and for the entire online ecosystem in the EU and at Member State level.

Importantly, the first proposal of Structural Indicators (Nenadic et al., 2023) compels platforms to explain, in an accessible way, how they define disinformation and how they identify sources and content of disinformation. The current proposal foresees to leave this definition up to the Signatories (Nenadic et al., 2024). Some experts consulted in the context of the Expert Group on Structural Indicators believe that leaving this up to the platforms is not ideal, both due to issues of comparability and general distrust in the platforms and calls for a common definition. However, there was a dissenting opinion that imposing any sort of definition on the platforms would be a 'regulatory overreach'. At the same time, not even the authors of the first beta assessment of the structural indicators were able to or willing to use a definition that would have allowed them to clearly differentiate between disinformation and misinformation.

2. (Very Large) Online Platforms

As it was highlighted earlier in this text, one of the key characteristics of disinformation is that in its pure form (if not overlapping, for example, with illegal content such as hate speech, see Wardle, 2024) it is not illegal in most jurisdictions. As such, platforms do not have the same legal obligations and (relatively) clear rules to act on disinformation content, as in the case of illegal content. Terrorist content, child sexual abuse material, hate speech, or even copyrighted content is usually removed when detected. At the same time, we can see in practice that platforms' community guidelines can also prohibit the publication of content that is not illegal, but may be considered harmful to the users, such

[11] Misinformation is false or misleading content shared without an intent to mislead, but its effects can still be harmful. Information influence operations are 'coordinated efforts by either domestic or foreign actors to influence a target audience using a range of deceptive means, including suppressing independent information sources in combination with disinformation'. Foreign interference refers to 'coercive and deceptive efforts to disrupt the free formation and expression of individuals' political will by a foreign state actor or its agents'.

as nudity and porn, spam, or false and misleading information, but also certain kinds of activities, such as the use of automated software applications.[12]

This means that in online platform content moderation, online intermediaries (or platforms) may prohibit otherwise legal forms of disinformation in their terms and conditions and enforce this contractual prohibition through (semi)automated or manually administered content moderation measures – even if there are some limits to this, as platforms are expected to respect human rights standards under their guiding principles, and in some EU country cases, such as Germany and Italy, courts have already argued that all of platform's content removal decisions cannot be justified simply by a contractual relationship (Pollicino, 2023:32). At the same time, it appears that the trend is going in the opposite direction: many of these platforms are reluctant to impose, what they would consider, unnecessary limitations on content. This is, among other things, because many of the currently popular online platforms emerged from a start-up culture that assigns to freedom of expression a particularly high value (e.g. Rozenshtein, 2021), but also because bans on specific kinds of content may alienate users, content moderation can be expensive (Keller, 2022), and some forms of problematic content, such as divisive disinformation, can generate more user engagement (Corzi, 2024). That said, it also needs to be mentioned that a preference towards an unregulated environment can by no means be understood as free speech absolutism: when platforms decide to take action, they often tend to over-moderate content[13], as a thorough assessment of content would require extensive investment in trained human content moderators (Keller & Leerssen, 2020).

On the following pages, we assess the internal policies of some online platform services that can be susceptible to information manipulations. These platforms are five of the Very Large Online Platform (VLOP) signatories of the Code of Practice, Instagram and Facebook (Meta) (which we assess together, due to the overlapping policies), LinkedIn (Microsoft), YouTube (Alphabet), and TikTok, the non-signatory VLOPs X (formerly Twitter), Wikipedia and Snapchat, as well as the non-VLOP signatories of the Code, Avaaz,[14] Twitch, Clubhouse and Vimeo. We do not consider search engines (the two Very Large Search Engines - VLOSEs, Google and Bing, and the smaller signatory search engine Seznam) in this assessment, as the content available through these services is not created by users who agree to terms of services, instead they index content available across the web. We included Snapchat as the only messaging platform, as it is designated as a VLOP, and has semi-public and public sharing options that make it similar to social networking sites (the messaging services of signatories are in theory covered by the Code of Practice, but none of them are considered VLOPs, and the challenges of defining and enforcing policies on end-to-end encrypted services would require a dedicated research).

As of April 2024, and based on the overview of online platform policies, we can see that mainstream social media platforms prefer to use the term ‘misinformation’ rather than ‘disinformation’ in their policies and reports to the Commission – this can be explained

[12] A list of policies can be found in Annex 2.

[13] With over-moderating we refer more broadly to two specific phenomena: ‘over-removing’ and ‘over-blocking’, where content is removed, on the one hand, and accounts are suspended, on the other hand, for pure opportunity (and fear of sanctions) even when it is legally permitted (Heldt, 2019).

[14] The Commission refers to Avaaz as a civil society/research organisation, rather than a platform.

by the fact that the distinction between the two would require an assessment of the communicator's intention, which is complicated, and in many cases nigh on impossible. However, such a decision also can make it more likely that platforms leave certain manifestations of the disinformation problem unaddressed on their services and justify possible inaction with the protection of users' freedom of expression or opt to take forms of action that may spare content from removal, such as adding fact-checking labels or other notes to the content.

Signatory	Term used	Definition
Meta (Facebook and Instagram)	Misinformation	No clear definition. Instead, Meta lists categories of misinformation.
Alphabet (YouTube)	Misinformation (and deceptive practices)	'Certain types of misleading or deceptive content with serious risk of egregious harm are not allowed on YouTube. This includes certain types of misinformation that can cause real-world harm, certain types of technically manipulated content, or content interfering with democratic processes.'
TikTok	Misinformation	'We do not allow inaccurate, misleading, or false content that may cause significant harm to individuals or society, regardless of intent.'
X	Synthetic and manipulated media/ misinformation/	'[S]ynthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm'
Microsoft (LinkedIn)	False or misleading content	'Specific claims, presented as fact, that are demonstrably false or substantially misleading.'
Avaaz	False material/ False or misleading information	'User Contributions must not: Contain any material which is false, defamatory, obscene, indecent, abusive, offensive, harassing, violent, hateful, inflammatory, endangers Avaaz's broader mission, or is otherwise objectionable.'
Twitch	Harmful Misinformation	Policy applies to users whose activity is 'dedicated to (1) persistently sharing (2) widely disproven and broadly shared (3) harmful misinformation topics, such as conspiracies that promote violence.'
Clubhouse	Harmful Misinformation/ Disinformation	Not defined

Vimeo	False or misleading claims/False or misleading information	Content that 'promotes fraudulent or dubious money-making schemes, proposes an unlawful transaction, or uses deceptive marketing practices; - Contains false or misleading claims about (1) vaccination safety, or (2) health-related information that has a serious potential to cause individual or public harm; Contains false or misleading information about voting or seeks to obstruct voting; Contains (1) claims that a real-world tragedy did not occur; (2) false claims that a violent crime or catastrophe has occurred; or (3) false or misleading information (including fake news, deepfakes, propaganda, or unproven or debunked conspiracy theories) that creates a serious risk of material harm to a person, group, or the general public; or - Violates any applicable law.'

Table 1. A summary of the definition of dis/misinformation of relevant CoP signatories.

Overall, defining the phenomenon of false or misleading content as 'misinformation' is the most common approach in platform policies. It is done by the platforms of Alphabet, Microsoft, TikTok, Twitter/X, and also Meta – often, it is emphasised that false or misleading information has to be harmful or potentially harmful to be considered under this category. But intent to mislead is not covered in policies. In some cases, definitions are also insufficient to clearly determine what kind of content would be considered false or misleading and harmful: Meta, for example, mentions in its policy that 'there is no way to articulate a comprehensive list of what is prohibited'.

In many cases, platforms are outsourcing decisions on falsity to external subject matter experts, such as civil society, public health authorities, fact-checking organisations (e.g. Meta[15], TikTok), refer to lists of previously identified deceptive media (Meta) or mention that content is considered false if it contradicts guidance from authoritative sources, such as the World Health Organization's communications during the Covid-19 pandemic (e.g. Microsoft).

There were only two cases in which we found 'disinformation' as a term used in platform policies – in the case of Clubhouse and Microsoft. For Clubhouse, the term was mentioned, but not defined – as a reason for action it referred to the potential or intention of causing harm and the intention to make money through deception. In practice, this provides a justification for the platform to act against content that falls into certain risky

[15] Meta, for example, works with fact-checkers certified by the International Fact-Checking Network, who categorise content as 'False, Altered, or Partly False' or rate it as 'Missing Context'. See: Meta, Transparency Policy: Fact-Checked Misinformation https://transparency.meta.com/features/approach-to-ranking/content-distribution-guidelines/misinformation; TikTok has a dedicated page for 'Safety partners' where it mentions the partners 'Australian Associated Press (AAP), Agence France-Presse (AFP), Animal Político, Code for Africa, dpa Deutsche Presse-Agentur, Estadão Verifica, Facta, Lead Stories, Logically, Newschecker, Newtral, PolitiFact, Reuters, Science Feedback, and Teyit'. https://www.tiktok.com/safety/en/safety-partners/

categories or can be considered a scam, but as 'potential' is added as an additional criterion to intention, in practice the latter doesn't need to be assessed by the platform. One can also argue that Clubhouse is a rather small platform that does not garner much public attention, thus, it devotes less attention to the framing of its policies.

In the case of Microsoft, the term 'disinformation' can be found in a policy related to the company's advertising offer – according to which 'Microsoft will not willfully profit from disinformation nor fund disinformation actors', thus advertisements or sites containing or leading to disinformation are excluded from the program. However, an exclusion of disinformation in the context of advertising services might have similar limitations as the use of 'misinformation' in the user-facing policies: if the platform is unable to assess the intent to mislead through publishing content, the use of the term 'disinformation' will become an empty signifier – as per default untrue content will be considered misinformation, until intent is proven.

The actions against misinformation-related content on VLOPs are topic- or context-driven: chiefly elections and emergencies. In certain cases, we find dedicated public health and election integrity policies, going beyond mis/disinformation, covering a number of activities that can contribute to election fraud or manipulation. The most common situation in which action against false or misleading content is emphasised is related to the integrity of elections, as it was already highlighted in the *Paris Call for Trust & Security in Cyberspace*, a multi-stakeholder cybersecurity initiative with more than thousand supporters, including national governments, companies, and the European Union[16]. The focus on elections is highlighted in the Call's Principle 3 'Defend electoral processes' with the aim of protecting the integrity of elections from interferences stemming from local and cross-border actors that aim to influence outcomes and electoral processes. As highlighted earlier in this text, elections were also a key area in EU documents, starting from the High Level Expert Group on fake news and online disinformation. Another important context in which false and misleading content is mentioned as requiring action is that of public health emergencies – one of the reasons why platforms treat this context as a priority is the global Covid-19 pandemic, which was accompanied by an increase of health-related mis- and disinformation. The phenomenon was also referred to as 'infodemic' (Zarocostas, 2020, Simon & Camargo, 2023) and triggered theoretical and policy discussions across the EU and beyond[17].

All VLOP signatories of the Code (including former signatory X) tie the policy to the potential of causing harm, while intention behind sharing certain content (except scams and possible undeclared benefits) is not considered. All signatory platforms mention misleading posts that can threaten the integrity of elections (depending on the exact definition, this may include false information about eligibility to participate, incitement to interfere with elections, or false claims about fraud) as possibly subject to removal or other action (such as reduced visibility or after multiple violations, a temporary locking of the account), while all but X include health/medical emergencies when they list content

[16] Paris Call for Trust & Security in Cyberspace and its 9 principles are available under https://pariscall.international/en/principles

[17] See, for example, the World Health Organisation's dedicated page on 'infodemic' and the Infodemic Management News Flash. https://www.who.int/health-topics/infodemic#tab=tab_1

that is not allowed on their platforms[18]. In certain cases, climate change disinformation is also dealt with by platforms, but in many cases, it is covered in a separate policy domain, not under disinformation (see also Romero-Vicente, 2023). However, in most cases there is no clear definition of what exactly is considered harm, and in certain cases, the platforms even distinguish between different forms of potential harmful impact, which determine the severity of platform response – for example, YouTube's policy mentions 'egregious harm' as a reason for removal (leaving space for possible other forms), Meta's policy mentions that content that 'directly' contributes to the risk of harm will be removed (and other 'hoaxes' and 'viral misinformation' will be demoted), while X mentions that there are certain kinds of content that are 'not likely to result in immediate harm but still have a potential to impact public safety, result in harm, or cause widespread confusion towards a public issue'.

Often, overlaps with categories of illegal content are mentioned, as well as with not illegal but prohibited categories. Manipulated media (including deepfakes) are covered by all platforms, either as part of the misinformation policies or in a separate policy – as well as scams, coordinated inauthentic activities and certain uses of a fake identity, such as impersonation. The platforms act on content that is illegal, and hate speech is mentioned in all cases as content that is not tolerated (even if in some cases different terms are used, they significantly overlap with hate speech and incitement to violence).

In the case of non-VLOP/VLOSE signatories of the CoP, Clubhouse and Vimeo have policies that are similar to the previously described signatories, albeit with less detailed descriptions in their reports and policies. Twitch differs by extending its criteria for action against certain users by mentioning their behaviour outside the platform as well (which would require the platform to have a clear identity profile of its users – spanning well beyond registering its legal identity), and the policies of Avaaz, as a platform that shares petitions, define stricter limits to what is acceptable on the service, which could almost be understood as a self-imposed editorial responsibility to protect the integrity of a mission-driven service.

There are two VLOPs that are not signatories of the Code: Wikipedia and Snapchat. Snapchat's rules are similar to those of VLOP signatories, as it refers to 'false or deceptive information' or in some cases misinformation and uses the potential harm as the criteria for action but doesn't take the intent of the actor into consideration. Similarly to Microsoft, it takes the guidance of health authorities as a benchmark for determining what can be considered health misinformation. As the CoP-signatories, Snapchat also looks at the integrity of elections and manipulated content. Wikipedia, on the other hand, takes a different approach. It doesn't refer to disinformation and misinformation, instead it has a policy that requires contributors not to 'lie'. As Wikipedia acts as a constantly updated online encyclopaedia, it aims for verifiability, and all statements on its page need to be pre-published and backed up by 'reliable' sources.

[18] X mentions 'public health emergencies' under its Crisis misinformation policy, but the policy itself is aimed at armed conflicts.

3. Discussion

On the following pages, we highlight certain considerations that impact the discussion on the usefulness of definitions – looking at the appropriateness of existing measures, the opportunities provided by re-interpretations of the definitions and policies, as well as the cases for changing the approach.

Disregarding intention is first and foremost a problem for overregulation. In our assessment we found that online platforms' policies related to misleading and harmful content only rarely mention the word disinformation and opt for the term 'misinformation' or other related concepts instead, in order to avoid assessing the intentions behind the publishing of harmful content. Possible justifications for this decision might be (i) the difficulty and cost of assessing intent, especially if such assessment has to be done in the case of a large number of content pieces, or (ii), as TikTok writes in its 'Combating Harmful Misinformation' policy, the consideration of intent as an unnecessary condition 'as the content's harm is the same either way'. While it is true that intention is not relevant when determining the societal harm that a piece of content may cause, when leaving intention out of the consideration, it can be argued that platform action will inevitably fall short when it comes to protecting freedom of expression, as the Code clearly asks signatories to consider 'the delicate balance that must be struck between protecting fundamental rights and taking effective action to limit the spread and impact of otherwise lawful content' (EC, 2022a:1). Not to mention that in the case in which the Code of Practice on Disinformation and its signatories use different wording and criteria to define the issues that they are acting upon, the effectiveness of the Code, as well as platforms' compliance, will become hard, if not impossible, to measure.

It is hard to expect more than what is in the policy. Falling short on certain aspects is especially concerning if we consider that the Code is a set of minimum actions and standards to uphold the integrity of the online environment, and it is unlikely that platforms would go further than the commitments of the Code, when it comes to taking actions that may affect their profitability. For example, in the case of topics covered by the mis/disinformation policies, we rarely see platforms extending their topical coverage beyond electoral and public health issues. Environmental content (which was mentioned in the 2018 CoP, but not in its 2022 iteration) is only explicitly covered in the mis/disinformation policies of TikTok. Meta's platforms cover the topic in their climate policy, and there are also indications that YouTube takes action against climate disinformation, despite not having a definition or a policy (see also Romero-Vicente, 2023).

Intention is not the only problematic concept; the operationalisation of harm is also arbitrary. When it comes to platform action, in most cases, the response to harmful misinformation and related content is removal, with possible account termination for repeat offenders. However, the unclear wording of policies and the vague criteria regarding what could in fact be considered harm, leaves a lot of discretion to the platforms. The platform X (not a signatory of the Code anymore), for example, differentiates between two types of harm, and it foresees lighter responses when a piece of content is 'not likely to result in immediate harm but still have a potential to impact public safety, result in harm, or cause widespread confusion'. If we assess platforms simply based on their definitions and clarity of policies (a proper assessment of

enforcement requires dedicated research), we can see that TikTok is closest to what the Code envisions. It is transparent in its criteria, provides a justification why it doesn't differentiate between disinformation and misinformation (it sees harm as a much more important determinator), introduces special actions and labels for borderline, 'unverified' content and provides a list of partner organisations it cooperates with when determining the veracity and harmfulness of content.

A constantly changing problem cannot justify the lack of definitions. Meta justifies its lack of a definition with the fact that the forms of information manipulations are constantly evolving. While it is true that the concept of dis- and misinformation is a moving target and the form it takes is constantly changing – due to changes in technology and tactics – this justification belies the fact that the components at its core of the definition are constant, such as the falsity of content and the potential of harm. A dynamic definition is, therefore, needed. A lack of definition makes it harder to identify the scope of any kind of action, not simpler.

There is no one-size-fits-all, but on most platforms, a blanket ban of fake and misleading content should be a no-go. Uniform policies – or even definitions – cannot be envisioned because all platforms are used differently and may attract divergent audiences. Thus, the different types of information manipulations, and specifically disinformation and misinformation, might take different forms across platforms. Current policies highlight that the platform's profile can also shape action: Avaaz and Wikipedia, for instance, argue that all kinds of false and misleading content would undermine the service they provide. Avaaz is a platform that publishes petitions, and therefore content is assessed in relation to Avaaz's mission – meaning that factual statements are required from authors/campaigners if they want to collect support. Wikipedia defines itself as an online encyclopaedia which needs to be factual, even if content can be produced by any user. To make sure that its 'don't lie' policy is enforced, the authors need to provide proof or reference to the statements included in published content, and all information has to be previously published in 'veritable sources'.[19] The CoP-signatory LinkedIn, at the same time, considers itself a 'real identity online social networking service for professionals', which, on the one hand, means that the platform claims to be less prone to mis/disinformation than others, and, on the other hand, extends the scope of content to be subject to platform action, as it also acts on non-harmful but false or misleading content, by limiting its reach. At the same time, most VLOP social media or messaging platforms (possibly including LinkedIn) are so widely used for all kinds of personal communications that limitations to lawful content need to be based on careful considerations.

Proxies can be utilised to differentiate between some forms of mis- and disinformation, but there are limits. While online platforms seem to be keen not to consider intentions behind publishing, we must ask ourselves whether it is possible to find criteria that allow for an effective differentiation between disinformation and misinformation.[20] And indeed, there are indications that proxies can be found that help

[19] This does not mean that Wikipedia cannot be used to spread disinformation, but the tactics might be different, and, according to news reports, Wikipedia has its own 'custodians' working on protecting its content. (See Borak, 2022)

[20] Many authors referred to the difficulty of distinguishing harmful disinformation from legitimate expressions of (a misguided) opinion – Bayer et al. (2021), for example point out that intent is 'an elusive

us identify, at least in certain cases, whether a piece of content is dis- or misinformation. TrustLab's assessment of the Code of Practice's impact (TrustLab, 2023), for example, considers scale as an indication of intention: focusing on 'visible signs from the user who posted the content such as (but not limited to) repeat activity, size of the follower network, manipulation of images, video, or audio clips, the deliberate use of misleading headlines, or clickbait as a way to attract attention and promote false narratives' the research referred to certain content publishers as 'disinformation actors' (while on the content-level, it still relied on the term 'mis/disinformation').[21] While this approach can be helpful as a first step towards identifying intentionally misleading content, it is more suitable for ex-post assessment, and has its limitations for timely platform action, as it can be complicated to assess every mis/disinformation content in the context of its publishers' other activities – not to mention that some characteristics, such as size of followership, only increase the likelihood of someone being a purveyor of disinformation, but does not provide proof.

Looking at additional categories of content addressed in the Code, might also help find a solution. When dealing with the problems of identifying intentions and preventing possible limitations to the freedom of expression, we also need to take into consideration that the Commission's Guidance on Strengthening the Code of Practice on Disinformation introduced an additional level of complexity to the definition of disinformation, as it stated that disinformation can be interpreted in 'the narrow sense', according to the definition discussed, but also as an 'overarching term' that may refer to a range of phenomena, such as 'misinformation, as well as information influence operations and foreign interference in the information space' (EC, 2021:5). This was reiterated in the 2022 Strengthened CoP, which also included actions against these broader categories in its commitments. While the use of the extended Disinformation category can be misleading to many observers wanting to use a clear definition, an assessment of the terms can in fact help us operationalise the definition. Peukert (2023) develops a multi-layered approach (see Table in Annex 3), which identifies key elements of these terms, which allow us to observe differences between mis- and disinformation (i.e., behaviour, content, and degree), as well as similarities between disinformation and influence operations or foreign interference (i.e., behaviour, degree and actor). Especially some characteristics of the publishing actor (large organisations or states that are expected to be informed communicators) and the degree of operation (that implies coordination between different actors) can be considered signs of a piece of content being intentionally spread.

Certain actors and techniques can justify anti-disinformation action. Given that 'foreign interference' and 'influence operation' do not have a specification on the factuality of the published content, one can also make the argument that these two are not distinct categories of information manipulation, but characteristics of actors and techniques that can often be observed in the context of disinformation as well. Thus, evidence of

---

criterion' as it 'requires one to inquire into the state of mind of the perpetrator, which may entail a complex evidentiary procedure.'

[21] TrustLab labels a user a 'disinformation actor' based on an assessment of up to 15 of posts (all mis/disinformation posts collected and some of the most recent posts). If at least 3 posts are mis/disinformation, the user will be considered a 'Level 1 actor', while large following (5000+), frequent posting (3+ times a month on a specific topic) can contribute to a denotation as 'Level 2 actor', if a secondary reviewer agrees.

influence operation and foreign interference can also be considered strong proxies of intent behind a piece of content.[22] This means that, to an extent, relevant insights can be gained from the publisher of a piece of content as well – for example, if one can be considered a proven state actor (or its proxy), a large organisation or a media outlet that needs to follow clear rules for content verification (and the publishing of non-factual content appears repeatedly), its non-factual publication may be considered disinformation, due to the higher standards such publishers would need to follow – or due to their conscious disregard for the truth. While these components will (in most cases) not make these actions illegal, they can provide guidelines for moderators taking action, as well as for researchers who want to assess the extent and the nature of disinformation (while also making a meaningful differentiation between mis- and disinformation), or the possible impact of anti-disinformation policies.

There is already a spillover effect from acting against overlapping categories. It has to be highlighted that most VLOP signatories already have policies against certain overlapping categories, such as fake accounts, fraud, coordinated inauthentic behaviour, adversarial threats or foreign interference, and also take action against content identified as disinformation by third parties, which will contribute to a decreased prevalence of disinformation on their services. This can likely be observed in future assessments done on the basis of an operationalised CoP definition. However, in this case, we can only speak of the spillover effects of policies that are officially not aimed at mis/disinformation.

Should the current policy approach potentially be reassessed? Coming back to the changes introduced by the Guidance and the Strengthened Code of Practice, one can also try to turn around the argument, pointing out that in certain cases the differentiation between misinformation and disinformation (or the focus on intentions) is not always justified: when it comes to issues like 'impermissible manipulative behaviour' (covered under the commitments of Integrity of Services, where the text highlights both mis- and disinformation) or misrepresented identities, the characteristics of the publisher already imply a conscious decision to mislead. At the same time, the use of monetisation services (either making money through publishing or using advertising services) should require higher standards from users and a commitment to truthfulness, as it would be the case

[22] In addition, the 2022 CoP lists a number of behaviours and tactics that can further add to our understanding of intent. Such behaviours and practices, which should periodically be reviewed in light with the latest evidence on the conducts and TTPs employed by malicious actors, such as the AMITT Disinformation Tactics, Techniques and Procedures Framework, include:

- The creation and use of fake accounts, account takeovers and bot-driven amplification,
- Hack-and-leak operations,
- Impersonation,
- Malicious deep fakes,
- The purchase of fake engagements,
- Non-transparent paid messages or promotion by influencers,
- The creation and use of accounts that participate in coordinated inauthentic behaviour,
- User conduct aimed at artificially amplifying the reach or perceived public support for disinformation.'

While some of these techniques fall under additional platform policies, they can help with the operationalisation of the difference between dis- and misinformation, given that they clearly imply action that is more likely to be taken by actors that purposefully spread manipulative messages.

in advertising and publishing, where actors are held to account for being untruthful or misleading possible consumers/audiences. This complexity, however, is not sufficiently highlighted in the wording of the Code, therefore, at some points, the ambiguity associated with the inclusion of misinformation in the wider definition of disinformation might even give the impression that the text doesn't sufficiently highlight in what cases unintentional manipulations of users need to be protected, and thus contradicts its stated goal of acting against information manipulations without unnecessarily infringing on users' freedom of expression. Whether this potential problem can be solved by introducing more clarity or whether the solution would be the introduction of a new definition, or even new terms to be used for the sake of this policy is out of the scope of this paper but needs to be discussed further.

4. Conclusion

This investigation into how 'disinformation' is defined in EU documents, national jurisdictions and online platforms not only confirms previous observed and well-known dynamics but also offers us insights into the ways in which definitions of different actors can raise concerns.

Firstly, it clearly highlights the inherent intricacies to set boundaries for disinformation. Identifying content that straddles the line between truth and falsehood, between malicious intentionality and naive unintentionality, between potential and actual harm, is a daunting task fraught with philosophical and practical challenges. Disinformation indeed overlaps with many other categories of harmful, misleading or manipulated content – some of which are illegal in some or many countries or may not be allowed by some platforms' terms and services. The trade-off faced is a democratically relevant one; on the one hand, broader definitions of disinformation might result in an excess of content removals, and therefore in censorship and a reduction of a diversity of voices which are essential for fostering critical thinking and informed decision-making whereas, on the other hand, tighter definitions of disinformation might concur to a 'post-truth' information landscape where misleading, divisive and polarising content would reduce societal resilience and decision-making abilities. Currently both overregulation and underregulation happen on platforms, sometimes as a misguided or badly calibrated response to requests by regulators or policy makers.

Secondly, the analysis shows how multiple governance layers may compete in the definition of disinformation – including the EU, the member states and the platforms' own community rules – latter enabling platform operators to prohibit certain, otherwise legal, forms of communication or user action.

While in EU documents disinformation is defined in a way that can provide a reference for action, platforms opt for defining only misinformation because they either see intention as an unnecessary condition if harm needs to be prevented (e.g. TikTok), or because it would be burdensome (and in certain cases impossible) for them to assess the intentionality of all the dis/misinformation content which is shared on their services.

It may be prudent for platforms to refrain from establishing clear and rigid definitions of disinformation. This approach allows for a broader range of interpretations of disinformation across diverse cultural contexts as well as in critical times. Indeed, the harmful impact of misleading content can vary significantly depending on geographical locations or specific circumstances, such as during elections or outbreaks of public health emergencies. Consequently, contextualising disinformation may help to mitigate the risk of inadvertently restricting free speech to a greater extent than the harm caused by disinformation itself. As such, minimising disinformation content moderation can prove to be more in line with liberal principles, which prioritise the free exchange of ideas and aligns with the predominant cultural and legal traditions in the United States where most of the platforms analysed ultimately originate from, and where its management decisions are made.

By concentrating on misinformation, these platforms avoid addressing the more insidious problem of disinformation, which often involves malicious intent and organised efforts to deceive. Addressing disinformation more directly would involve stricter policies, increased transparency, and a commitment to combating coordinated efforts to deceive users. Ultimately, the prioritisation of misinformation over disinformation may be a strategic choice for social media platforms to avoid taking an editorial responsibility and the burden of balancing human rights and justify their decisions, but it likely does not fully address the complex challenges posed by deceptive content on their platforms. The use of the terms implies that platforms choose options that limit their responsibility to take action, as misleading content may be considered unintentional by default. This is also the case when the term ‘disinformation’ is used in a platform policy: in Microsoft’s advertising offer a ban on placing advertisement next to disinformation content in the context of advertising services might have similar limitations as the use of ‘misinformation’ in the user-facing policies: if the platform cannot assess intent, the use of the term ‘disinformation’ is only an excuse for inaction – as per default untrue content will be considered misinformation, until intent is proven.

Finally, we observe how defining disinformation and operationalising actions against it represent two sides of the same coin. To fully grasp the definitions of disinformation it is indeed required to conduct further investigation on how these definitions are used as basis of action, in practice. The definitions provided by platforms ultimately offer limited insight into how they enact disinformation moderation, as they tend to be overly broad and thus lack the specificity needed to anticipate the course of content moderation practices until a decision is made. Indeed, the operationalisation of these definitions involves multiple stages, primarily centred around determining the categories that constitute disinformation and subsequently outlining the appropriate measures to address it. As the remedies taken by platforms to moderate online disinformation are various and often opaque, the next analysis will focus exactly on this with the goal of providing a comprehensive view of how disinformation is legally defined and practically identified and, eventually, moderated.

References

Law, regulation and policy documents

EC - European Commission (2018a). *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. Https://Eur-Lex.Europa.Eu/Legal-Content/EN/TXT/?Uri=CELEX:52018DC0236

EC - European Commission (2018b). *Code of Practice on Disinformation*. https://ec.europa.eu/newsroom/dae/redirection/document/87534

EC European Commission (2020a). *Communication from The Commission to The European Parliament, The Council, The European Economic and Social Committee and The Committee of The Regions on The European Democracy Action Plan* COM/2020/790 Final

EC – European Commission. (2020b). *Communication on Tackling COVID-19 disinformation—Getting the facts right.* JOIN/2020/8 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020JC0008

EC - European Commission (2021). *Guidance on Strengthening the Code of Practice on Disinformation* (COM(2021) 262 final). https://ec.europa.eu/newsroom/dae/redirection/document/76495

EC – European Commission (2022a). *Strengthened Code of Practice on Disinformation.* https://disinfocode.eu/wp-content/uploads/2023/01/The-Strengthened-Code-of-Practice-on-Disinformation-2022.pdf

EC - European Commission (2022b). Digital Services Act. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065

HLEG – High Level Expert Group on Fake News and Online Disinformation (2018). A multi-dimensional approach to disinformation - Report of the independent High level Group on fake news and online disinformation. *European Commission.* https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271

Bibliography

Abbamonte, G. B. & Gori, P. (2023). Freedom of speech and the regulation of fake news in the European Union: the EU policy to tackle disinformation. In: Pollicino, O. (ed.), Freedom of speech and the regulation of fake news, Cambridge : Larcier-Intersentia, 2023, pp. 129-166

Altay, S., Berriche, M., Heuer, H., Farkas, J., & Rathje, S. (2023). A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*, 4(4), 1-34.

Bayer, J., Holznagel, B., Lubianiec, K., Pintea, A., Schmitt, J. B., Szakács, J. & Uszkiewicz, E. (2021). Disinformation and propaganda: impact on the functioning of the rule of law and democratic processes in the EU and its Member States - 2021 update. *European Parliament Think Tank*. https://www.europarl.europa.eu/thinktank/en/document.html?reference=EXPO_STU%282021%29653633

Bazaco, Á., Redondo, M., & Sánchez-García, P. (2019). Clickbait as a strategy of viral journalism: conceptualisation and methods. *Revista Latina de Comunicación Social*, (74), 94.

Betzel, M., Nyakas, L., Papp, T., Kelemen, L., Monori, Z., Varga, Á., Marrazzo, F., Matějka, S., Ó Fathaigh, R., & Helberger, N. (2020). Notions of Disinformation and Related Concepts (ERGA Report) [Report]. European Regulators Group for Audiovisual Media Services. https://erga-online.eu/wp-con tent/uploads/2021/03/ERGA-SG2-Report-2020-Notions-of-disinformation-and-related-concepts-fin al.pdf

Bleyer-Simon, K. (2021). Government repression disguised as anti-disinformation action: Digital journalists' perception of Covid-19 policies in Hungary. *Journal of Digital Media & Policy*, *12*(1), 159-176.

Bleyer-Simon, K. (2024). (De)monetisation of Disinformation: Can the actions of large online platforms be measured? *Centre for Media Pluralism and Media Freedom*. https://cmpf.eui.eu/demonetisation-of-disinformation/

Borak, M. (2022, 17 Oct.). The Hunt for Wikipedia's Disinformation Moles. *Wired.* https://www.wired.com/story/wikipedia-state-sponsored-disinformation/

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

Corsi, G. (2024). Evaluating Twitter's algorithmic amplification of low-credibility content: an observational study. *EPJ Data Science*, *13*(1), 18.

Cunningham, T., Pandey, S., Sigerson, L., Stray, J., Allen, J., Barrilleaux, B., ... & Rezaei, B. (2024). What We Know About Using Non-Engagement Signals in Content Ranking. *arXiv preprint arXiv:2402.06831*.

Diaz Ruiz, C. (2023). Disinformation on digital media platforms: A market-shaping approach. *New Media & Society*, 0(0).

Etienne, H. (2021). The future of online trust (and why Deepfake is advancing it). *AI and Ethics*, 1(4), 553-562.

Fallis, D. (2015). What is disinformation?. *Library trends*, *63*(3), 401-426.

Frau-Meigs, D. (2021). Addressing the Risks of Harms Caused by Disinformation: European vs. US Approaches to Testing the Limits of Dignity and Freedom of Expression Online. *The Handbook of Communication Rights, Law, and Ethics: Seeking Universality, Equality, Freedom and Dignity*, 135-146.

Griffin, R., & Vander Maelen, C. (2023). Codes of conduct in the digital services act: exploring the opportunities and challenges. *SSRN*. https://dx.doi.org/10.2139/ssrn.4463874

Hegelich, S., Dhawan, S., & Sarhan, H. (2023). Twitter as political acclamation. *Frontiers in Political Science*, *5*.

Heldt, A. P. (2019). Merging the social and the public: How social media platforms could be a new public forum. *Mitchell Hamline Law Review*, 46, 997-1042.

Hughes, H. C., & Waismel-Manor, I. (2021). The Macedonian fake news industry and the 2016 US election. PS: Political Science & Politics, 54(1), 19-23.

Keller, D. (2022, Feb. 24). *The DSA's Industrial Model for Content Moderation, VerfBlog.* https://verfassungsblog.de/dsa-industrial-model/.

Keller, D., & Leerssen, P. (2020). Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation. In N. Persily & J. A. Tucker (Eds.), *Social Media and Democracy* (pp. 220–251). Cambridge: Cambridge University Press.

Kertanegara, M. R. (2018). Clickbait Headline and Its Threat in The National Resilience. *CoverAge: Journal of Strategic Communication*, *8*(2), 57-62.

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. Proceedings of the National Academy of Sciences, 120(7), e2210666120.

Macdonald, S., & Vaughan, K. (2023). Moderating borderline content while respecting fundamental values. Policy & Internet.

Mchangama, J. & Alkiviadou, N. (2020). *The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship - Act two.* Justitia. https://justitia-int.org/wp-content/uploads/2020/09/Analyse_Cross-fertilizing-Online-Censorship-The-Global-Impact-of-Germanys-Network-Enforcement-Act-Part-two_Final-1.pdf

Nenadić, I. (2019). Unpacking the" European approach" to tackling challenges of disinformation and political manipulation. *Internet Policy Review*, *8*(4), 1-22.

Nenadić, I., Brogi, E., & Bleyer-Simon, K. (2023). Structural indicators to assess effectiveness of the EU's Code of Practice on Disinformation.

Nenadic, I., Brogi, E., Bleyer-Simon, K., & Reviglio, U. (2023). Structural Indicators of the Code of Practice on Disinformation: The 2nd EDMO report. *European Digital Media Observatory.* https://edmo.eu/wp-content/uploads/2024/03/SIs_-2nd-EDMO-report.pdf

Ó Fathaigh, R., Helberger, N., & Appelman, N. (2021). The perils of legally defining disinformation. *Internet policy review*, *10*(4), 2022-40.

Peukert, A. (2023). The Regulation of Disinformation in the EU–Overview and Open Questions. Research Paper of the Faculty of Law of Goethe University Frankfurt/M, (2).

Pollicino, O., & Bietti, E. (2019). Truth and deception across the Atlantic: a roadmap of disinformation in the US and Europe. *Italian Journal of Public Law*, *11*, 43.

Polyák, G. (2020). Hungary's two pandemics: COVID-19 and attacks on media freedom. *European Centre for Press and Media Freedom Legal Opinion*, Update 28 June, https://www.ecpmf.eu/hungarys-two-pandemics-covid-19-and-attacks-on-media-freedom/

Romero-Vicente, A. (2023). Platforms' policies on climate change misinformation. *EU Disinfo Lab.* https://eu.boell.org/sites/default/files/2023-09/factsheet_platforms_climate_misinformation_final.pdf

Rozenshtein, A. Z. (2021). Silicon Valley's Speech: Technology Giants and the Deregulatory First Amendment. *Journal of Free Speech Law 1*:337-376. https://ssrn.com/abstract=3911460

Simon, F. M., & Camargo, C. Q. (2023). Autopsy of a metaphor: The origins, use and blind spots of the 'infodemic'. *New Media & Society*, 25(8), 2219-2240.

Singhal, M., Ling, C., Paudel, P., Thota, P., Kumarswamy, N., Stringhini, G., & Nilizadeh, S. (2023, July). SoK: Content moderation in social media, from guidelines to enforcement, and research to practice. In 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P) (pp. 868-895). IEEE.

Subramanian, Samanth. 2017. "Meet the Macedonian Teens Who Mastered Fake News and Corrupted the US Election." Wired, February 15. Available at www.wired.com/2017/02/veles-macedonia-fake-news.

Trustlab (2023). Code of Practice on Disinformation. A Comparative Analysis of the Prevalence and Sources of Disinformation across Major Social Media Platforms in Poland, Slovakia, and Spain. *Code of Practice on Disinformation Transparency Centre*. https://disinfocode.eu/wp-content/uploads/2023/09/code-of-practice-on-disinformation-september-22-2023.pdf

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media+ Society, 6(1), 2056305120903408.

Wardle, C. (2018). Information disorder: The essential glossary. Harvard, MA: Shorenstein Center on Media, Politics, and Public Policy, Harvard Kennedy School.

Wardle, C. (2018). The need for smarter definitions and practical, timely empirical research on information disorder. Digital journalism, 6(8), 951-963.

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Vol. 27, pp. 1-107). Strasbourg: Council of Europe.

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, 395:676.

Zeng, J., & Brennen, S. B. (2023). Misinformation. *Internet Policy Review*, *12*(4).

Annex 1: Platforms’ terms, definitions, criteria, categories and related concepts (as of 31 May 2024).

VLOP signatories of the CoP

VLOP	Term used	Definition	Criteria & Intent	Categories	Related concepts[23]
Meta (Facebook and Instagram)	Misinformation (at times also 'false news' and 'false information')	No clear definition	Identified as false or altered content and the potential of causing harm. Intent is not covered. (fact-checking policy mentions opinion and speeches by political as not falling under content to be checked, but action can be taken on 'content presented as opinion but based on underlying false information')	Categories requiring removal: • content contributing to the risk of imminent physical harm or violence • harmful health misinformation (misinformation about vaccines, 'about health during public health emergencies' and 'promoting or advocating for harmful miracle cures') • contributing to interference with the functioning of political processes • content from certain highly deceptive manipulated media Categories that require reducing prevalence: • 'hoaxes' and 'viral disinformation' (not defined) Categories that require informative label:	Hate speech, fake accounts, fraud, coordinated inauthentic behaviour

[23] Not an exhaustive list

				• 'Content Digitally Created or Altered that May Mislead' (but otherwise not in violation of Community Standards) • content flagged by fact-checkers for 'missing context' Categories that require no action: • harmless misinformation 'to exaggerate a point' ('This team has the worst record in the history of the sport!') • humour & satire ('My husband just won Husband of the Year.')	
Alphabet (YouTube)	Misinformation (and deceptive practices)	Misinformation: 'Certain types of misleading or deceptive content with serious risk of egregious harm are not allowed on YouTube. This includes certain types of misinformation that can cause real-world harm, certain types of technically	Content is considered if misleading or deceptive and high risk of 'egregious' harm; intent is not mentioned Misleading/deceptive means 'obviously doctored or manipulated, or […] taken out of context'	- Suppression of census participation: a set of criteria that can prevent a person from participating in a census - Manipulated content: when manipulation or doctoring can mislead users and 'may pose a serious risk of egregious harm'. - Misattributed content: old footage presented as portraying a current event. Additional policy dedicated to election misinformation • Voter suppression • Candidate eligibility	scam, spam, impersonation, fake engagement, external links, violent or graphic content, hate speech

		manipulated content, or content interfering with democratic processes.'		• Incitement to interfere with democratic processes • Distribution of hacked materials • Election integrity and medical misinformation, related to: • Prevention • Treatment • Denial	
TikTok	Misinformation	Misinformation refers to 'inaccurate, misleading, or false content that may cause significant harm to individuals or society, regardless of intent.' (Conspiracy theories are added as 'beliefs about unexplained events or involve rejecting	It is highlighted that intent is not considered. Significant harm is defined: it 'includes physical, psychological, or societal harm, and property damage. It does not extend to commercial and reputational harm, nor does it cover simply inaccurate information and myths.' (societal harm is further explained: 'including undermining	The following forms of 'misinformation' are not allowed under the platforms' policy: '• Misinformation that poses a risk to public safety or may induce panic about a crisis event or emergency, including using historical footage of a previous attack as if it were current, or incorrectly claiming a basic necessity (such as food or water) is no longer available in a particular location • Medical misinformation, such as misleading statements about vaccines, inaccurate medical advice that discourages people from getting appropriate medical care for a life-threatening	Influence Operations, Civic and Election Integrity, Synthetic and Manipulated Media, Fake Engagement, Unoriginal Content and QR Codes, Spam and Deceptive Account Behaviors, Bullying, Harassment

		generally accepted explanations for events and suggesting they were carried out by covert or powerful groups.')	fundamental social processes or institutions, such as democratic elections, and processes that maintain public health and public safety')	disease, and other misinformation that poses a risk to public health • Climate change misinformation that undermines well-established scientific consensus, such as denying the existence of climate change or the factors that contribute to it • Dangerous conspiracy theories that are violent or hateful, such as making a violent call to action, having links to previous violence, denying well-documented violent events, and causing prejudice towards a group with a protected attribute • Specific conspiracy theories that name and attack individual people • Material that has been edited, spliced, or combined (such as video and audio) in a way that may mislead a person about real-world events' Under the Civic and Election Integrity section: 'Election misinformation, including the following:	

				- How, when, and where to vote or register to vote - Eligibility requirements of voters to participate in an election, and the qualifications for candidates to run for office - Laws, processes, and procedures that govern the organisation and implementation of elections and other civic processes, such as referendums, ballot propositions, and censuses - Final results or outcome of an election'	
X/Twitter	Synthetic and manipulated media / misinformation / misleading content / misleading information	'You may not share synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm' (Synthetic and manipulated media policy) 'We define misleading content ('misinformation')	'In order for content with misleading media (including images, videos, audios, gifs, and URLs hosting relevant content) to be labeled or removed under this policy, it must: • Include media that is significantly and deceptively altered, manipulated, or fabricated, or • Include media that is shared in a deceptive	● Synthetic and manipulated media (image & video) ● Crisis misinformation Civic integrity	Misleading and deceptive identities, distribution of hacked materials, financial scam, platform manipulation and spam policy, 'copypasta' and duplicate content, ban evasion

		as claims that have been confirmed to be false by external, subject-matter experts or include information that is shared in a deceptive or confusing manner.'[24]	manner or with false context, and • Include media likely to result in widespread confusion on public issues, impact public safety, or cause serious harm'		
Microsoft (LinkedIn)	'False or misleading content' (Only advertising policies refer to 'disinformation')	LinkedIn defines misinformation as 'specific claims, presented as fact, that are demonstrably false or substantially misleading.' [25]	Content that falls under this policy is demonstrably false or substantially misleading. In case of public health it is also mentioned that it 'directly contradicts guidance from leading global health organizations and	Examples of content considered: ● related to upcoming or recent elections (time, location, means, or eligibility requirements for voting) ● Claims that may induce panic or discourage others from taking safety precautions during an emergency	scam/fraud, misrepresentation, hate speech

[24] This definition was not available anymore on X's website in May 20224, but it is also quoted in Hegelich et al. (2023).

[25] Microsoft advertising services use a broader definition, which also includes 'disinformation': 'Microsoft prohibits misleading deceptive content, or harmful content, or content that otherwise threatens public or personal safety, physical, mental, or financial health, or content whose primary purpose is to create controversy. Examples include, without limitations: Unsubstantiated claims; fraudulent free offers or pricing claims; sensationalized text or images; content that isn't related to the product/service being promoted; misrepresentations; unauthorized promotion of third-party products and services; information influence operations, foreign interference, false or misleading content that may cause public harm, or other similar behaviors ('disinformation').'

			public health authorities' Intent behind publishing is not considered (except undeclared personal benefits that are related to the content) If a post can cause harm: the action is removal If it is not considered as possibly causing harm: it will be restricted to user's network.	• content taken out of context, related to human rights abuses or military conflict • Synthetic or manipulated media • false or misleading content related to public health, pandemics, miracle cures	

Non-VLOP/VLOSE signatories of the CoP

Signatory	Term used	Definition	Criteria & Intent	Categories	Related concepts
Avaaz	'false material', 'false or misleading information'	'User Contributions must not: Contain any material which is false, defamatory, obscene, indecent, abusive, offensive, harassing, violent, hateful,	Intent is not specified, content can be removed based on its factuality	no	no

		inflammatory, endangers Avaaz's broader mission, or is otherwise objectionable.'			
Twitch	'harmful Misinformation'	Policy applies to users whose activity is 'dedicated to (1) persistently sharing (2) widely disproven and broadly shared (3) harmful misinformation topics, such as conspiracies that promote violence.'	multiple offences lead to action (either on the platform or elsewhere)	- Misinformation against protected groups (covered in the Hateful Conduct & Harassment Policy) - Public health: 'Harmful health misinformation and wide-spread conspiracy theories related to dangerous treatments, COVID-19, and COVID-19 vaccine misinformation' - Conspiracy theories 'Misinformation promoted by conspiracy networks tied to violence and/or promoting violence' - Election 'Civic misinformation that undermines the integrity of a civic or political process	Impersonation, Hateful Conduct & Harassment

				- Promotion of verifiably false claims related to the outcome of a fully vetted political process, including election rigging, ballot tampering, vote tallying, or election fraud' - Public safety in cases of public emergencies (identified case-by-case, based on possible impact)	
Clubhouse	'Harmful Misinformation' and 'Disinformation'	not defined	- potential of causing harm - intention (of causing harm or making money through deceiving)	- health misinformation - misinformation for money - elections and other civic processes - "synthetic" or manipulated media	- impersonation or misrepresentation (including the use of pseudonyms, if not justified by artistic or human rights reasons) - spam, bots, artificial behaviour (including 'artificial amplification or suppression of information')

Vimeo	'false or misleading claims' / 'false or misleading information'(misinformation mentioned in statement about 'updates to our content policies', but not in the acceptable use policy)	Content that - 'Promotes fraudulent or dubious money-making schemes, proposes an unlawful transaction, or uses deceptive marketing practices; - Contains false or misleading claims about (1) vaccination safety, or (2) health-related information that has a serious potential to cause individual or public harm; - Contains false or misleading information about voting or seeks to obstruct voting; - Contains (1) claims that a real-world tragedy did not occur; (2) false claims that a violent crime or catastrophe has occurred; or (3) false or misleading information (including fake news, deepfakes,	intent not mentioned, only potential of harm (and possible topics)	- fraudulent commercial activity - risk to public health - obstruction of voting - emergency or making up an emergency - violating applicable laws	- impersonation (or acting in 'deceptive manner') - using misleading metadata - inauthentic use and spamming (mentioned in CoP report, cannot be found on Vimeo website. Refers to 'use of bots, scripts, or other automated tools for any purpose' and 'creating fake accounts, liking and commenting on your own content using another account, and purchasing likes or comments from third-parties'.)

		propaganda, or unproven or debunked conspiracy theories) that creates a serious risk of material harm to a person, group, or the general public; or - Violates any applicable law.'			

VLOP, not signatory of CoP

VLOP (non-signatory)	Term used	Definition	Criteria & Intent	Categories	Related concepts
Wikipedia	'lie'	'There are many ways that editors can lie on Wikipedia, such as deliberately using a quote out of context to mislead readers, fabricating a reference, stating content is not included in an article when it	Factuality of content, based on: - verifiability: 'other people using the encyclopedia can check that the information comes from a reliable source' - sourcing (footnotes) neutral point of view: multiple sources to be mentioned, in case of disagreement	N/A	N/A

		actually is, or making untrue accusations about the conduct of another editor.'	- previously published information (no original research is allowed) - published in a reliable source		
Snapchat	Harmful False or Deceptive Information (misinformation comes up in the description of actions that undermine the integrity of the civic process)	'false information that causes harm or is malicious'	- 'causes harm or is malicious' - intent is not considered - 'our teams take action against content that is misleading or inaccurate, irrespective of whether the misrepresentations are intentional' (but: intent is mentioned in the context of manipulating content) - accuracy vis-a-vis authoritative sources - In case of health misinformation, accuracy is determined based on correspondence with health agencies' guidance	- denying the existence of tragic events, - unsubstantiated medical claims, - undermining the integrity of civic processes (broader category than mis/disinformation, ie. 'intimidation to personal safety' and 'content that encourages people to misrepresent themselves to participate in the civic process or to illegally cast or destroy ballots') - manipulating content for false or misleading purposes (whether through generative AI or through deceptive editing).	- Misrepresentation ('pretending to be someone (or something) that you're not, or attempting to deceive people about who you are. This includes impersonating your friends, celebrities, public figures, brands, or other people or organizations for harmful, non-satirical purposes.') - spam (such as, 'pay-for-follower promotions or other follower-growth schemes, the promotion of spam applications, or the

					promotion of multilevel marketing or pyramid schemes') - fraud and deceptive practices ('including the promotion of fraudulent goods or services or get-rich-quick schemes, or imitating Snapchat or Snap Inc.')

ANNEX 2: Platform policies that refer to disinformation and related phenomena.

Company	Main Source(s)
Meta	- Meta Community Standards (Misinformation) and Content Distribution Guidelines - Manipulated Media policy - Guidelines: Fact-Checked Disinformation - January 2023 Report in the Code of Practice Transparency Centre
Alphabet	- YouTube misinformation policies (including Elections misinformation policies and Medical misinformation policy) - January 2023 Report in the Code of Practice Transparency Centre
Microsoft	- False or misleading content policy - Misinformation and inauthentic behavior - Microsoft Advertising. Disallowed content - January 2023 Report in the Code of Practice Transparency Centre
X / Twitter	- Synthetic and manipulated media policy - January 2023 Report in the Code of Practice Transparency Centre
TikTok	- Transparency Center: Combating Harmful Misinformation - Community Guidelines: Integrity and Authenticity - January 2023 Report in the Code of Practice Transparency Centre
Snapchat	- Snapchat Community Guidelines - Harmful False or Deceptive Information. Community Guidelines Explainer Series - Our Approach to Preventing the Spread of False Information
Wikipedia	- Wikipedia: Don't lie - Wikipedia Conduct policy - Wikipedia: Verifiability
Avaaz	- Privacy Policy & Terms of Use - January 2023 Report in the Code of Practice Transparency Centre
Twitch	- Community Guidelines - January 2023 Report of Twitch, Transparency Centre - Post: Preventing Harmful Misinformation Actors on Twitch - January 2023 Report in the Code of Practice Transparency Centre
Clubhouse	- Clubhouse Community Guidelines - January 2023 Report in the Code of Practice Transparency Centre
Vimeo	- Vimeo, acceptable use policy - Updates to content policy - January 2023 Report in the Code of Practice Transparency Centre

ANNEX 3: Peukert' categorisation of disinformation and related concepts in the 2022 Strengthened Code of Practice on Disinformation

	Misinformation	Disinformation	Influence Operation	Foreign Interference
Actor	Any, less likely a large organization or state actor	Any	Any, but likely a large organization or state actor	State actor and/or its proxies
Behaviour	No evidence of an intent to deceive	Evidence of deliberately deceptive behaviour	Coordination of various techniques aimed at a common goal	Coordination of various techniques aimed at a common goal
Content	Often legitimate expression of an opinion	verifiably deceptive or untrue elements[37]	Any, often multiple types of measures	Any, often multiple types of measures
Degree	Limited evidence of coordination	Any	Scale of the operation indicates coordination	Any
Effect	Any	Any	Any, but should further the objective(s) of the actor	Any, but should further the objective(s) of the actor

Peukert, A. (2023). The Regulation of Disinformation in the EU–Overview and Open Questions. Research Paper of the Faculty of Law of Goethe University Frankfurt/M, (2).

EDMO

EUI SCHOOL OF TRANSNATIONAL GOVERNANCE

www.edmo.eu

The European Digital Media Observatory has received funding from the European Union under contract number LC-01935415