Who Decides What Counts as Disinformation in the EU?

Alexander Peukert 2023-10-24T07:45:20

Who decides what counts as “disinformation” in the EU? Not public authorities, because disinformation is not directly sanctioned in the Digital Service Act (DSA) or other secondary legislation. Nor Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSes), which avoid editorial decisions to maintain their legal status as intermediaries with limited liability. Instead, an analysis of reports published on the 2022 Strengthened Code of Practice on Disinformation (Disinfo Code) Transparency Centre and online self-descriptions (‘about us’) reveals that the delicate task of identifying disinformation is being undertaken by other private organisations whose place of administration and activity, purpose, funding and organizational structure appear problematic in terms of the legitimacy and even legality of the fight against disinformation. This blog post maps out the relevant (private) actors, namely the ad industry, fact checking organizations and so-called source-raters.

Brand Owners and Other Ad Industry Players

The first group of organizations that classify disinformation is primarily concerned with the demonetization of disinformation, which is a key element of the EU’s anti-disinformation policy. In the very first commitment of the Disinfo Code, “relevant signatories participating in ad placements” pledge to “defund the dissemination of Disinformation, and improve the policies and systems which determine the eligibility of content to be monetised”. Following the money in the online ad market ultimately leads to brand owners and other advertisers deciding where to place their ads. Their aggregate control of the global ad budget gives them considerable leverage, even vis-à-vis the largest ad-financed social media platforms.

Motivated by a strong commercial self-interest to preserve “brand safety”, i.e. to prevent the association of valuable brands with harmful or even only polarizing content, the global ad community indeed fully supports the fight against disinformation. The Disinfo Code has been signed by the Interactive Advertising Bureau Europe, the European Association of Communication Agencies and last but not least the World Federation of Advertisers (WFA), which claims to represent 90 % of global (!) marketing expenditure. In complying with the commitment to advance the development and use of “brand safety tools and partnerships” (Measure 1.6 of the Disinfo Code), WFA set up the Global Alliance for Responsible Media (GARM), a cross-industry initiative led by brands, which also includes agencies, ad tech companies and platforms. The GARM initiative developed a common understanding of what harmful and sensitive content is, where ads should not appear (“Brand Safety Floor”), and a common way of delineating different risk levels for sensitive content (“Brand Suitability Framework”). The Brand Safety Floor, i.e. content not

appropriate for *any* advertising support, includes i.a. the “insensitive, irresponsible and harmful treatment of debated social issues … that … incite greater conflict” and “verifiably false or willfully misleading content that is directly connected to user or societal harm”. The Brand Suitability Framework calls for enhanced advertiser controls if ads appear next to high, medium or low risk content. High risks are for example associated with “the discussion of debated social issues … in a negative or partisan context”, and medium risks with “Breaking News or Op-Ed coverage” of “partisan advocacy of a position on debated sensitive social issues” or “coverage of misinformation”. Still a low risk is triggered by an “Educational, Informative, Scientific treatment of debated sensitive social issues … including misinformation” and “news features describing various misinformation campaigns as such”. Thus, if a news publisher objectively and diligently reports about a sensitive issue like migration or a misinformation campaign, it already creates a brand safety risk and may lose ad revenues.

In order to fulfil their respective commitments under the Disinfo Code (measures 3.1 and 3.3) but first and foremost in order to meet their customers’ needs, the biggest social networks and search engines have integrated GARM classifications into their systems. In their reports under the Disinfo Code and in a separate GARM Measurement Report, Google/YouTube, META/Facebook/Instagram, Microsoft/ LinkedIn and TikTok unanimously praise their (allegedly) effective enforcement of their GARM-compliant ad policies. Even X/Twitter, which withdrew from the Disinfo Code after Elon Musk’s acquisition of the company, still features in the 2023 GARM Measurement Report and promises “enhanced brand safety controls and reporting for our advertisers”.

On the face of it, market-based cooperation between the ad industry and VLOPs/ VLOSes does not appear to be problematic. An advertiser is free to choose how and where to spend its ad budget, and preserving brand safety is a legitimate concern reflected inter alia in EU trademark law. However, the size and scope of GARM with its seamless implementation of GARM standards in the online ad environment poses a real and significant threat to media freedom and pluralism. If the partisan discussion of controversial social issues, and even the scientific treatment of a sensitive topic, raises brand safety concerns and thus may negatively affect ad revenue prospects, news providers will have good reason to refrain from publishing this kind of public interest information in the first place. The GARM system therefore tends to undermine the strict separation between editorial decisions on the one hand and the commercial interests of advertisers on the other, which has been a cornerstone of media law for decades. To prevent brand owners from exerting undue influence on editorial decisions by blacklisting websites that publish legitimate journalistic content, Member State media authorities should investigate GARM. Such oversight is also needed because several key players in the ad tech industry have their headquarters outside the EU. In addition to Google and Microsoft, this also concerns the Disinfo Code signatories MediaMath and DoubleVerify, both headquartered in New York.

Fact-Checking Organisations

While brand owners and other advertisers define brand safety and thereby general disinformation standards, they do not decide whether a particular piece of content qualifies as dis- or misinformation. This role has instead been assumed by fact-checking organisations, whose articles are used by VLOPs and VLOSes in various ways to identify and down-rank disinformation. Meta, for example, reports that it used over 190,000 distinct fact-checking articles to both label and reduce the amplification of over 40 million pieces of content on Facebook in the first half of 2023.

On the one hand, fact-checking organisations operate under strict rules of independence and transparency and in a relatively well-established institutional framework. The European Fact-Checking Standards Network (EFCN) is a signatory of the Disinfo Code and represents dozens of fact-checking organisation. Its members agree to be evaluated by two independent academic experts every two years on their compliance with the European Code of Standards for Independent Fact-Checking Organisations, and they are subject to an EFCN complaint procedure.

On the other hand, fact-checks still raise problems regarding freedom of expression and the media. In two decisions (here and here), the Karlsruhe Court of Appeal held that a fact-checking organisation (CORRECTIV) and Facebook engaged in unfair competition against an online news publisher by placing negative fact-checking labels on two articles, thereby reducing their reach. The court reasoned that the fact-check denigrated the services of the publisher in an unjustified way because it falsely represented the content of the original article: misinformation on alleged disinformation.

This case-law demonstrates that fact-checking needs to be independently monitored and, if necessary, corrected. Considering the scale, effect and speed of fact-checks, a punctual and expensive ex-post regulation by ordinary courts appears inadequate. As of today, content providers who are confronted with a negative fact-check are also unable to rely on art. 20 paras 4-6 DSA, which will only oblige online platforms from 17 February 2024 to reverse unjustified content moderation decisions without undue delay. However, a fact-check that violates unfair competition law is at the same time an “illegal content” under the DSA. VLOPs and VLOSes are therefore obliged to assess and if necessary mitigate the risk of the dissemination of such illegal fact-checks through their services (art. 34(1)(a) DSA). Repeat infringers (i.e. bad quality fact-checkers) should be excluded from the scheme, and ratings have to be precise and sufficiently supported by the check. It is questionable whether the Facebook categories “Partly false” and “Missing context” satisfy this threshold.

Source-Raters

The third group of actors engaging in the classification and identification of disinformation are source-raters. Unlike fact-checkers, these organisations do not evaluate individual pieces of content, but entire websites, particularly those of

news providers. Source-raters thus operate at a more fundamental level of content curation than fact-checkers. Despite this, source-raters have received less attention from the public, academia and regulators than fact-checkers. Three of these actors will therefore be briefly presented.

Reporters without Borders (RSF) is an international non-profit organisation working to defend and promote the freedom, pluralism and independence of journalism. RSF, a signatory of the Disinfo Code, is engaged with the issue of disinformation primarily via the Journalism Trust Initiative (JTI), a self-regulatory standard that RSF initiated and is currently in the process of rolling out. JTI translates the self-assessments and independent third-party audits of media outlets regarding transparency and guarantees of editorial independence and processes into a machine-readable trustworthiness indicator. The JTI certi#cation mark will be awarded for two years to media outlets that meet the requirements. When integrated into Big Tech's algorithms, the compliance certificate should improve the ranking of the certified media outlets.

The second source-rater, which signed the Disinfo Code is NewsGuard, a company established 2018 and headquartered in New York. NewsGuard provides credibility ratings and detailed "Nutrition Labels" for thousands of news and information websites across the U.S., U.K., Canada, Germany, France, Italy, and Austria. The rating, done by local journalists, depends inter alia on whether a website repeatedly publishes false content, whether it gathers and presents information responsibly and avoids deceptive headlines. Microsoft reports to employ NewsGuard trustworthiness indicators for numerous products, including the Edge browser, the MSN news aggregator, Bing Search and the AI-powered Bing Chat. NewsGuard furthermore offers Misinformation Fingerprints™, a catalog of "top hoaxes", which can be used as a unique identifier to seed AI tools searching for disinformation. This product was positively evaluated in a test run with AI "social listening tools" by U.S. Cyber Command to monitor state-sponsored disinformation in near real time. A similar service to analyse "digital chatter" is provided by Crisp, a Code signatory owned by a large U.S. risk and financial advisory company and headquartered in the UK.

The third source-rater to be briefly introduced is the Global Disinformation Index (GDI). GDI, also established in 2018 and based in the UK, is a not-for-profit organisation, which means that all income is reinvested in support of its mission to defund disinformation. GDI is not a signatory to the Disinfo Code, but – like the U.S.-based Alliance for Securing Democracy – expressly referenced in Microsoft's report to the Code. GDI uses both expert human review and AI to assess disinformation risks across the open web. It claims to have assessed more than 700 million websites in over 40 languages. GDI's operations are funded by ad tech licenses of GDI's dynamic exclusion list, philanthropic organisations such as the Knight Foundation, and governments (U.S., UK, Germany, EU) to conduct studies of the news ecosystems in countries under the influence of authoritarian regimes, such as China and Russia. GDI stresses, however, that it "never worked with – and will never work with – any government to conduct research into that government's own country". GDI's close ties to public authorities on both sides of the Atlantic are also reflected in the biographies of its two founders. One (Clare Melford) serves on the

Advisory Council for the EU-funded European Digital Media Observatory (EDMO), which in turn is represented in the permanent task force of the Disinfo Code, the other one (Daniel Rogers) has a background in the U.S. intelligence community. With this in mind, it is perhaps not surprising that GDI's foundational study on the concept of disinformation begins with the McLuhan quote "World War III is a guerrilla information war with no division between military and civilian participation."

Such a credo should raise concerns when it comes to regulating the digital public sphere in the EU. Under German constitutional law, the press must remain free from state influence, let alone military influence (which, by the way, is also very present in the DISARM Framework supported by the Code signatory Alliance4Europe). In its proposal for a European Media Freedom Act, the Commission stresses that opaque and biased allocation of state advertising can be a powerful tool to 'capture' media service providers to the detriment of fundamental rights. Yet in the context of the particularly delicate war against disinformation, we observe an indirect but still intense involvement of public authorities, including intelligence agencies, in the classification and rating of alleged disinformation sources. This is most obvious in the case of the GDI, but three of the four co-funders of RSF's Journalism Trust Initiative are also either public authorities/agencies (European Commission, Agence Française de Développement) or fully publicly funded (U.S. National Endowment for Democracy). If there is any truth to the adage that he who pays the piper calls the tune, such a funding structure is another cause for concern. In addition, the JTI trust certificate suffers from potentially biased self-assessments by participating media organisations seeking privileged visibility in the hotly contested news market.

Considering this, NewsGuard appears to be the least objectionable source-rater. If their ratings are inadequate, no-one will be willing to pay for them. Yet it remains a remarkable fact that NewsGuard is headquartered in a third country, namely the U.S., where EU laws, for example regarding data protection, do not apply. It is indeed strange that the European Commission, on the one hand, proclaims that it wants to strengthen Europe's "digital sovereignty" and become a global standard-setter rather than follower, while, on the other hand, it accepts and even encourages the outsourcing of the highly sensitive question of which website is a purveyor of disinformation to a U.S. company and a UK not-for-profit (GDI). It is true that NewsGuard, under Measure 22.4 of the Disinfo Code, committed to "ensure that information sources are being reviewed in a transparent, apolitical, unbiased, and independent manner, applying fully disclosed criteria equally to all sources and allowing independent audits by independent regulatory authorities or other competent bodies". Whether the company meets these standards is another matter. In a case currently pending before the Frankfurt Regional Court, a German news provider claims that NewsGuard's negative rating of its website is unjustified and thus unfair and illegal. If the court agrees, Microsoft and other VLOPs/VLOSes should reconsider integrating NewsGuard's source ratings into their algorithms.

Implications

Until now, the activities of disinformation classifiers have largely flown under the radar of public and academic attention. Considering that the European Approach

to tackling misinformation “should strictly respect freedom of expression” and include safeguards that prevent the misuse of respective measures, “for example, the censoring of critical, satirical, dissenting, or shocking speech”, this is an unsustainable state of affairs. There is an urgent need for a critical assessment of the operations of disinformation classifiers, which are not subject to the relatively strict rules of art. 22 DSA on trusted flaggers regarding illegal content. In view of this regulatory lacuna, the Commission is called upon to implement the DSA and in particular the rule on systemic risks for fundamental rights (art. 34(1)(b) DSA) in a balanced way that transforms the current, one-sided mode of anti-disinformation self-regulation into a co-regulatory scheme that pays due respect to the freedom of expression, information and the media. The possibly imminent DSA investigation into several online platforms provides a perfect opportunity to achieve this aim.
