Codes of conduct: a promising tool to counter online disinformation

Authors: Nicolo Zingales*, Marina Lucena*, Leandro Rebelo*, Laise Barbosa*, Giovanna Milanese*, Henrique Bazan*

Sumário

1 The concept of disinformation

Disinformation is a complex phenomenon with important repercussions worldwide, especially in the last few years. There is a general concern with disinformation and its potential consequences, including those related to harms and risks to democracy, freedom of expression, freedom of speech, and the integrity of the electoral process, among others.

Defining what type of content is worthy of regulatory attention is a major challenge in drafting any regulation. Terms such as “fake news” and “disinformation" are employed to mean a wide variety of different conduct that vary greatly in scope, intentionality, and harm. It is difficult to envisage a simple, concise definition that provides regulators and companies with sufficient guidance as to how to act; it is worthwhile, then, to discuss some of the challenges that such a definition must face.

One option is to classify content in accordance with two axes, those of truthfulness and (potential) harm, as suggested in a 2017 report to the Council of Europe[1]. According to this categorization, false and harmful content is “disinformation”, true but otherwise harmful content (such as a leak of private information) is “mal-information”, and false but not intentionally harmful content is “misinformation”. Further, the idea of “misinformation” attempts to differentiate users who unknowingly spread false information online from those who do so to cause harm (which would then qualify as “disinformation” *per se*). Although the CoE report expands on potential motivations for spreading false information (i.e., financial, political, etc.), it seems exceedingly difficult to accurately discern a user’s motivations from the fact that they have shared certain content online - as its authors argue, it is possible to share something as a way of showing disagreement.

---

* Research Group on Platform Governance, Center for Technology and Society, Fundação Getulio Vargas in Rio de Janeiro. See https://direitorio.fgv.br/pesquisa/centro-de-tecnologia-e-sociedade.

[1] Claire Wardle and Hossein Derakhshan, ‘Information Disorder: Toward an interdisciplinary framework for research and policy making’, Council of Europe report DGI(2017)09 (the CoE Report), 5. *Apud* CAVALIERE, Paolo. The Truth in Fake News: How Disinformation Laws are Reframing the Concepts. University of Edinburgh School of Law, Research Paper Series, No. 2022/12. Available at SSRN: https://ssrn.com/abstract=4151908

Another option is to treat disinformation as "verifiably false or misleading"[2] information which "*may* [emphasis added] cause public harm", as done in 2018 by the Communication on Tackling Online Disinformation, provided by the European Commission[3]. To illustrate, the Communication explicitly excludes "reporting errors, satire and parody, or clearly identified partisan news and commentary"[4] from this concept: although such content *could* be misleading, the Communication seems to have chosen to emphasize the *intent* behind such forms of media that is implicit in certain uses of information, which could provide a justification to its otherwise potentially harmful nature.

However, these boundaries could prove difficult to navigate in concrete instances. Indeed, the idea of "verifiably false information" could be interpreted as being at odds with the CoE Report's definition of "misinformation", in the sense that under this approach users of a platform that share information could not be exempted from failing to verify that information was false (and would, therefore, be guilty of sharing disinformation). This calls into question the issue of how a person - or a platform - can judge the truthfulness of any discrete publication or idea. The European Court of Human Rights has previously advanced that "absolute truth" is not a reasonable standard to require for journalism, and, instead, it should be evaluated whether the speaker has acted in good faith and attempted to verify what they say[5].

Translating this definition to the context of online platforms is not straightforward, as users could reasonably attempt to verify disinformation, only to find other websites and sources which reproduce the same spurious claims[6] - perhaps more importantly, given that such standards were drafted for journalists, a discussion is needed regarding the necessary diligence that the average user must present. In any case, it is difficult to envisage how the "intent" of any specific user to cause harm could be measured, as there could be no real distinction in the behavior of someone deeply concerned by false information they believe in when compared to a user intentionally disseminating facts

[2] EUROPEAN COMMISSION. Tackling Online Disinformation: A European Approach. COM(2018) 236 final.
[3] EUROPEAN COMMISSION. Tackling Online Disinformation: A European Approach. COM(2018) 236 final.
[4] *Id.*
[5] Cavaliere, supra note 1.
[6] https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/

they know to be untrue. “Intent to harm”, then, could easily be construed as intent to inform other people of relevant matters of public interest.

Despite the elusiveness of defining terms such as “intent”, “harm” and “truthfulness”, it does not seem like there are easier solutions to the problem. Considering that the behaviors of users across distinct platforms can vary widely, it is not possible to determine what standards must be taken across the board. Instead, codes of conduct should indicate that platforms must discourage average users from sharing information that may be false, and otherwise prevent potentially harmful information from spreading, rather than emphasize remedial approaches. For this reason, the reports’ emphasis on “trusted” providers of information and fact-checking agencies seems well-intentioned but not entirely effective: such agents will necessarily only act after the disinformation has already widely spread and taken root. Sparse research also suggests that fact-checking may work better on users who are unaware of disinformation than on those who have been wholly convinced by it[7], which poses its own set of problems.

While intention is difficult to ascertain, technical measures that curb massive, rapid dissemination of content[8] could hinder large-scale automatic messaging (such as by bots) without significant harms to freedom of speech. Incentivizing users to read links before forwarding them[9], for example, can be an unobtrusive way to ask users to verify the information they are sharing - although this, once again, leaves open the question of intent.

In conclusion, while it is possible to reach reasonable working definitions of “disinformation”, operationalizing the ideas of “harm”, “intent” and “truthfulness” appears perplexing, but this should not discourage the work of platforms. Contextual clues are important, and each platform is more acutely aware of its users’ specific habits. As such, codes of conduct should attempt to reach a definition of disinformation that articulates the ideas of truthfulness, harm and intent, and then provide general guidance on different types of behavior (such as, for example, indicating that sharing content by mistake is less grave than doing it intentionally), categories which must be further defined by each platform. As the cited documents agree, inaction is not a welcome possibility;

[7] https://www.brookings.edu/articles/when-are-readers-likely-to-believe-a-fact-check/
[8] https://faq.whatsapp.com/1053543185312573
[9] https://twitter.com/Support/status/1270783537667551233

while edge cases will always permeate this system, focusing on the concrete effects of acts of sharing content appears to be a worthwhile start, leaving room for users to demonstrate "intent" as an exculpating factor.

2 Codes of conduct and their role in regulatory governance

Codes of conduct can be part of different regulatory strategies, ranging from self-regulation to co-regulation, which in turn can be more or less constrained by public law. Thus, the term can be used to describe different industry roles[10]. The terminology used also can be different from one instrument to another. The concept of Code of Practice is typically seen as a manifestation of a self-regulatory regime. However, a Code of Practice would be considered a Code of Conduct if it "becomes an instrument for fulfilling legally prescribed obligations, namely, to assess and mitigate the risk of disinformation"[11]. The European Commission's Code of Practice on Disinformation (2022), for instance, aims to become a Code of Conduct, according to its preamble, (i) and under article 35, DSA.

Codes of conduct can be voluntary industry codes, and in this case, it is a type of self-regulation, through which industry participants choose their standards and commit to comply with them. Self-regulation is characterized by rules formulated by the industry. In this case, there are codes of conduct and the industry itself is responsible for their enforcement. In the case of misinformation online, a self-regulatory instrument seeks to encourage the online platforms to act on their own – for instance, by tackling the dissemination of false and harmful content, ensuring transparency, and promoting trustworthy sources and fact-checking organizations.

Co-regulation occurs when an industry develops its own arrangements against the backdrop of government legislation and enforcement. In the co-regulation regime, the Codes seek to implement some provisions already established in a government law.

[10] TERRY, Andrew. The Unusual Place of Industry Codes of Conduct in the Regulatory Framework. UNSW Law Journal, volume 45(2), 2022, p. 649-687. Available at: www.unswlawjournal.unsw.edu.au/wp-content/uploads/2022/07/Issue-452-06-Terry.pdf.
[11] NENADIC, Iva; BROGI, Elda; BLEYER-SIMON, Konrad. Structural indicators to assess effectiveness of the EU's Code of Practice on Disinformation. Robert Schuman Centre for Advanced Studies. 2023. Available at SSRN: https://www.ssrn.com/abstract=4530344, p. 8.

The creation of Codes of Conduct typically begins with proposals from organization leaders or members, followed by extensive deliberations and debates involving experts and stakeholders. This consensus-driven approach[12] relies on input from committee members and participants to ensure broad agreement. However, it's important to mention the risks and difficulties associated with seeking this consensus, such as the need to make sacrifices to reach compromise over certain aspects in the pursuit of common ground among stakeholders. Furthermore, companies with more resources tend to have an advantage in this context, potentially dominating the consensus-building process towards their own interests.

Regulation is intended to address important economic and social problems. However, the most suitable governance model may depend on the country, social context, and culture. Nowadays, the use of cooperative instruments of shared responsibility is increasingly proposed to approach complex problems in the governance of online platforms[13]. This is notably the case when it comes to social media platforms and the spread of disinformation. Social media platforms are dealing with the challenges related to the definition of "disinformation" and how to address the problem – what is the best path to moderate content without compromising freedom of speech, for instance. Even if disinformation does not occur only on social media platforms, these constitute an important channel to share all types of false and harmful content. Similar considerations apply to instant messaging applications and ad networks, which, however, have different characteristics and therefore may require different containment measures when it comes to combating disinformation. As a general rule, the overarching principles guiding the elaboration of codes must maintain a certain level of openness in order to accommodate for the different roles that these companies have. It is also important, in this context of quick and constant technological and social changes, for signatories to have the opportunity to define their own criteria, following the rules and principles established both in the law and in those very codes where the adopted measures originate. Nevertheless, the definition of the boundaries and metrics of application of the code must be transparent, in order to ensure meaningful accountability. One good example is

[12] COGLIANESE, C (2023). Private Codes and Standards. U of Penn Law School, Public Law Research Paper, No. 23-34. P. 1. Available at SSRN: https://ssrn.com/abstract=4592696.
[13] HELBERGER, N., PIERSON, J., & POELLl, T. (2018). Governing online platforms: From contested to cooperative responsibility. The Information Society, 34(1), 1-14. https://doi.org/10.1080/01972243.2017.1391913.

provided by the EU Code of Practice on Disinformation, which can be seen as a key instrument in the European Union to address the disinformation problem.

3 Case study: EU Code of Practice on Disinformation

The Code of Practice on Disinformation was launched in 2018, and revised for the first time in in 2019 with new commitments. Another update came in 2022, defining 44 commitments and 128 proposed measures. Most importantly, the 2022 improved version includes structural indicators to measure disinformation. Although it can be challenging to measure disinformation to understand this complex phenomenon, it is important to discuss the indicators[14].

The 2022 version of the Code provides two different key performance indicators: Service-Level Indicators and Structural Indicators. Service-Level Indicators are related to specific measures adopted according to the Code by each signatory in order to meet some quantitative reporting elements with regard to the six areas of concerns outlined in the Code: scrutiny of ad placements, political ads, integrity of services, empowering users and empowering the research community. Structural Indicators, on the other hand, aim to analyze the disinformation phenomenon and the Code's effectiveness in Europe. There are some Structural Indicators to measure the effectiveness of the Code of Practice on Disinformation in Europe. They were the result of a complex process involving consultations with experts, academic researchers, civil society organizations, and members of EDMO (European Digital Media Observatory) and ERGA (European Regulators Group for Audiovisual Media Services). Additionally, a literature review was conducted to approach methods and metrics used in empirical studies to measure disinformation and misinformation.

The Proposal contains six structural indicators. The first one is the prevalence of disinformation, that is, the total number of contents identified as disinformation and harmful misinformation. This can be somewhat problematic to the extent that it is based

[14] NENADIC, Iva; BROGI, Elda; BLEYER-SIMON, Konrad. Structural indicators to assess effectiveness of the EU's Code of Practice on Disinformation. Robert Schuman Centre for Advanced Studies. 2023. Available at SSRN: https://www.ssrn.com/abstract=4530344.

on self-reporting by a particular platform, which does not guarantee the accuracy of this self-assessment. The second indicator mentions the sources of disinformation, identifying what is the total number of such sources, including the originators and the superspreaders. The third is the audience of disinformation, which aims to monitor the users exposed to disinformation, in an aggregated and anonymized way. Fourth, the demonetization of disinformation approaches the monetization strategies used by disinformation providers and the revenue obtained. The fifth structural indicator is the collaboration and investments in fact-checking, including its availability, the extent of collaboration within the platforms, and funding by platforms. Finally, the sixth indicator approaches the investments in the overall implementation of the Code, including the financial and human resources invested by signatories to meet the commitments and objectives established in the Code.

The question about the possibility of applying the same indicators settled in Europe in Brazil and other countries from the named "Global South" remains. However, in general, the Structural Indicators mentioned above do not seem related to a specific social-demographic context, which means they can be useful to assess online disinformation in other countries. On the other hand, Service-Level indicators may vary greatly to address very different scenarios, even holding the areas of concern constant. Furthermore, those areas might necessitate different types of actions depending on the environment in which they are brought to bear: for instance, empowering the community of researchers in a Global South context might require greater investment in digital literacy and affordances to be provided by signatories.

It is also important that signatories of the Code may commit to specific points, without being obliged to comply with the entire CPD. From this context, two central characteristics of codes of practice (also referred to as codes of conduct) can be inferred: there is greater flexibility to update them compared to hard laws; and as they are optional, companies have the choice to opt in and select which provisions to comply with.

The Digital Service Act (DSA) was launched in 2022, although some provisions regulating very large online platforms only came into force in 2023. The Act establishes rules for transparency, governance, and accountability for content moderation, as well as addressing issues related to the liability regime for digital platforms.

The Code of Practice on Disinformation and the Digital Service Act have similar purposes but differ primarily due to the former's voluntary nature and the latter's binding nature. This difference is a primary criticism of codes of practice in general, directed at the CPD due to its lack of genuine enforcement mechanisms[15]. However, the Code is seen as a precursor to the DSA, as it informs measures that platforms can adopt to address the issue of disinformation. In the DSA, there are general obligations to combat disinformation (Article 35), which is defined as a type of systemic risk (Article 31, DSA). In this aspect, compliance with the CPD is understood as a basis for an online platform addressing combat measures to mitigate the systemic risk of disinformation.

Article 104 of the DSA suggests that platforms refusing to participate in codes of practice without justifiable reasons may be deemed non-compliant with the DSA. In this sense, although voluntary, the CPD assumes a central role for supervisory authorities regarding enforcing the Digital Service Act. The voluntary nature of the CPD did not impede its adoption and practice, with its impact noted in online platforms' formal and practical aspects[16]. Although platforms have the option to adhere to the Code of Disinformation Practice, this soft law instrument contributes to shaping the online space.

4 Codes of conduct: promises and challenges

The debate surrounding Codes of Conduct has driven certain promises and challenges associated with their implementation in various organizational contexts.

[15] QUINTEL, T. ULLRICH, C (2018). Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond. **Fundamental Rights Protection Online**: The Future Regulation Of Intermediaries. P. 12-14. Available at SSRN: https://ssrn.com/abstract=3298719. However, the voluntary nature of the Code is apparent considering the lack of enforcement actions following the announcement in May 2023 that the social network X would no longer adhere to it, while measures requested under the Code of Practice(such as measures taken to combat the spread of misinformation and access to data) played a role in the opening of a compliance investigation of X pursuant to the DSA. See https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709.

[16] BORZ, Gabriela et al. The EU soft regulation of digital campaigning: regulatory effectiveness through platform compliance to the code of practice on disinformation. Policy Studies, p. 1-21, 2024.

Promises:

1. **Flexibility and Agility:** Codes of conduct offer greater adaptability than government regulations, allowing for regular updates (reducing the risk of outdated detailed standards) to address emerging risks and evolving standards and enabling targeting specific platforms or technical features. They provide coordination and governance in areas where public law is insufficient or silent, offering predictability and clarity.

2. **Multi Sectoral Representation:** The development of codes of conduct can involve representatives from a diversity of sectors industry, academia, governmental, and non-governmental organizations. This ensures that codes of conduct are tailored to specific needs and challenges, enhancing their relevance and effectiveness.

3. **Applicability across Platforms:** Codes of conduct can be adopted across different platforms and contents, encouraging proactive measures to prevent the proliferation of harmful content and practices in different contexts, and enabling mutual learning and cross-pollination.

4. **Enhance effectiveness and accountability.** Codes could significantly enhance the law's effectiveness by strengthening accountability and promoting effective risk mitigation measures in areas not adequately addressed by the law. In addition, codes can introduce specific commitments and develop stronger institutional accountability mechanisms and oversight structures. This may significantly enhance accountability by providing information to regulators, users, researchers and society overall, thereby facilitating independent scrutiny.

Challengees:

1. **Voluntary approach:** The voluntary nature of codes of conduct may lead to inconsistent adoption and enforcement, especially if compliance lacks sufficient incentives or imposes significant costs on firms.

2. **Time and resources.** The parties involved may invest significant time and resources in negotiations, because reporting and participating is a continuous commitment. Thus, developing future codes in several areas simultaneously may be infeasible, leading to difficult decisions on which areas to prioritize.

3. **Questions about legitimacy.** Questions can be raised about the legitimacy of regulating via soft law.

   a. **Corporate capture.** With regard to the throughput legitimacy, Codes can be criticized for not including diverse stakeholders. As participation capacities inevitably reflect existing disparities of power and resources, stakeholder participation offers further potential for corporate capture.
   b. **Private actors as regulators.** On input legitimacy, industry-led codes inherently delegate regulatory authority from democratically-legitimate institutions to private actors. However, using codes to fill gaps indicates a lack of democratic input, suggesting that it was inadequately considered in the legislative process.
   c. **Dominance of Multinational Corporations:** Local stakeholders may have difficulty participating in Code negotiations, which could be dominated by multinational corporations, influencing outcomes to their advantage, and are unlikely to prioritize local issues. This dominance raises concerns about regulatory capture and the undue influence of corporate interests in code development. In addition, Codes may create disproportionate compliance costs for smaller companies, undermining competition and pluralism.

4. **Effectiveness in Changing Business Practices.** It can be questioned how effectively Codes will achieve legitimate outputs, i.e., meaningful and socially beneficial changes in business practices. Without appropriate incentives, codes may reformulate the limited and low-cost measures that the parties involved already have in place, rather than promoting more far-reaching and innovative interventions.

Conclusions and Recommendations:

1. **Interaction between law and codes**. The incorporation of these codes into laws and regulations presents a practical solution for governments seeking to streamline regulatory processes and leverage existing expertise. However, "incorporation by reference"[17], where only the name of the code is referenced, can obscure the specific requirements of the law (particularly if these standards are subject to copyright license).

[17] COGLIANESE, C (2023). Private Codes and Standards. U of Penn Law School, Public Law Research Paper, No. 23-34. P. 3. Available at SSRN: https://ssrn.com/abstract=4592696

Therefore, legislators must proactively think about transparency, accountability and equitable access to these codes. Furthermore, more creative forms of interaction between law and codes of conduct ought to be considered, with incentives playing a prominent role in driving adoption and compliance.

2. **Participation and accountability in code design and implementation.** Participation must be built into the design and implementation of codes, in order to ensure that such instruments take into account the interests of diverse stakeholders. Furthermore, it is crucial to consider mechanisms for monitoring and resolution of disputes associated with the implementation of these codes, so as to boost effectiveness in achieving the regulatory objectives.

3. **Need for precise definitions.** A precise definition of what type of content should be categorized as “disinformation” (including its relationship to "misinformation") is essential to drafting an effective regulation. Providing a complete and definitive concept is challenging but it is a problem that should be addressed. Thus, providing a clear definition – even one that can be revised and changed over time - would help to improve users' confidence and clarity about what types of content are being removed from the platforms, for example.

4**. Asymmetry of undertakings**. The Codes could craft differentiated rules, dependent on the particular circumstances of its signatories: for instance, platforms with specific functionalities or a larger user base could be attributed a heightened level of responsibility. By the same token, general principles of the Codes may be applied differently to platforms, in a similar way that Service-Level Indicators analyze the conformity of each platform under the Code of Practice on Disinformation.

5. **Flexibility and adaptability**. The platforms’ terms are transience, they are constantly and without warning changing their design, terms, and conditions. The Government’s Legislative process is typically slow and rife with political and bureaucratic hurdles. In addition, every time that a technological, cultural, or social change happens, it may be necessary to change the law, which requires more expenditure of time and resources. Codes of conduct could be the intermediate option between these options, providing some certainty and stability but without unduly sacrificing the flexibility that is needed to deal with complex challenges.

6. **Availability of several compliance options for signatories**. Codes of conduct may allow different options to platforms concerning how to deal with disinformation online, promoting diversity and generating virtuous competition on effective use of technology and governance practices. For instance, they can focus on remedial approaches, such as partnerships with fact-checking companies. However, in certain situations other instruments can be more effective, such as the discouragement of users to share false content, or controls over the number of users who can see the content of dubious veracity. These specific measures can be used depending on the platform's characteristics, providing more flexibility and effectiveness.

References:

BORZ, Gabriela et al. The EU soft regulation of digital campaigning: regulatory effectiveness through platform compliance to the code of practice on disinformation. Policy Studies, p. 1-21, 2024.

CAVALIERE, Paolo. The Truth in Fake News: How Disinformation Laws are Reframing the Concepts. University of Edinburgh School of Law, Research Paper Series, No. 2022/12. Available at SSRN: https://ssrn.com/abstract=4151908.

COGLIANESE, C (2023). Private Codes and Standards. U of Penn Law School, Public Law Research Paper, No. 23-34. P. 1. Available at SSRN: https://ssrn.com/abstract=4592696.

EUROPEAN COMMISSION. Tackling Online Disinformation: A European Approach. COM(2018) 236 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236.

GRIFFIN, Rachel; MAELEN, Carl Vander. Codes of Conduct in the Digital Services Act: Exploring the Opportunities and Challenges. Law, AI & Regulation Conference, Erasmus University, Rotterdam, 2023. Available at: https://ssrn.com/abstract=4463874.

HELBERGER, N., PIERSON, J., & POELLl, T. (2018). Governing online platforms: From contested to cooperative responsibility. The Information Society, 34(1), 1-14. https://doi.org/10.1080/01972243.2017.1391913.

LEISER, Mark. Reimagining Digital Governance: The EU's Digital Service Act and the Fight Against Disinformation. 2023. Available at SSRN: https://ssrn.com/abstract=4427493.

NENADIC, Iva; BROGI, Elda; BLEYER-SIMON, Konrad. Structural indicators to assess effectiveness of the EU's Code of Practice on Disinformation. Robert Schuman Centre for Advanced Studies. 2023. Available at SSRN: https://www.ssrn.com/abstract=4530344.

QUINTEL, T. ULLRICH, C (2018). Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond. Fundamental Rights Protection Online: The Future Regulation Of Intermediaries. P. 12-14. Available at SSRN: https://ssrn.com/abstract=3298719.

TERRY, Andrew. The Unusual Place of Industry Codes of Conduct in the Regulatory Framework. UNSW Law Journal, volume 45(2), 2022, p. 649-687. Available at: www.unswlawjournal.unsw.edu.au/wp-content/uploads/2022/07/Issue-452-06-Terry.pdf.

TWITTER. Support. Available at: https://twitter.com/Support/status/1270783537667551233.

VERMA, Pranshu. The rise of AI fake news is creating a 'misinformation superspreader'. Available at: https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/.

WHATSAPP. Frequently Asked Questions (FAQ). Available at: https://faq.whatsapp.com/1053543185312573.