BRIEFING

STOA/ESMH Options Brief

Regulating disinformation with artificial intelligence

Effects of disinformation initiatives on freedom of expression and media pluralism

This study examines the consequences of the increasingly prevalent use of artificial intelligence (AI) disinformation initiatives upon freedom of expression, pluralism and the functioning of a democratic polity. It considers the trade-offs in using automated technology to limit the spread of disinformation online and presents options (from self-regulatory to legislative) to regulate automated content recognition (ACR) technologies in this context. Special attention is paid to the opportunities for the European Union as a whole to take the lead in setting the framework for designing these technologies in a way that enhances accountability and transparency and respects free speech.

Options for regulating AI in disinformation

Option/form of regulation	Typology of regulation
0 Status quo	Corporate social responsibility, single-company initiatives. Enforcement of existing legislation would continue and expand, such as the General Data Protection Regulation, ePrivacy Regulation, and Audiovisual Media Services (AVMS) Directive
1 Non-audited self-regulation	Corporate code of practice on algorithmic moderation, with transparency and self-reporting. Example: Santa Clara Principles
2 Audited self-regulation	Open interoperable publicly available industry standard on algorithmic moderation. Example: European code of practice of September 2018
3 Formal self-regulator	Industry standard on algorithmic moderation, with requirement to conform to standard/prove equivalence, powers to expel non-performing members, and dispute resolution ruling/arbitration on cases
4 Co-regulation with statutory powers	Government-approved industry standard on algorithmic moderation. Examples from broadcast and advertising regulation
5 Statutory regulation	National regulatory agencies. Formal regulation – tribunal with judicial review

Several levels of AI disinformation policy options are open to European policymakers. We consider the status quo/no-regulation option (0) and the pair that propose formal regulation (4-5) to be the least likely to be deployed. Option 1 remains the least favourable option throughout. The costs of uncertainty are much higher for the less regulatory options (0-3), and regulatory sustainability and protection of fundamental rights (including freedom of expression/media pluralism) is more strongly supported for the more regulatory options (4-5). However, that is **not** a proposal for any kind of super-regulator or 'Offdata'.[1] Legislation may be premature at this stage: collaboration between different stakeholder groups with public scrutiny (2-3) is preferable, provided protection of fundamental rights is included in the design of the self-regulatory measures and can be independently demonstrated via audit. Importantly, options are interdependent – if regulation is proposed, it sits atop a pyramid of activities including co-regulation, self-regulation, technical standards and company/NGO/academic initiatives. No single option solves disinformation.

Option 0: Status quo

This option would entail permitting both 'natural' technical experiments in moderation, and the legislative responses that already exist, such as that of Germany's Network Enforcement Law (NetzDG). It would also rely on individual corporate efforts to enforce regulation, rather than an industry self-regulation scheme or democratically legitimate institutional oversight. Individual users would continue to rely on companies' Terms of Service enforcement for their own and others' freedom of expression (with widely varying content standards, definitions of abusive/harmful content etc.). Individual companies would continue to pursue disparate aims according to their own judgement of brand interest.

The benefits of no regulation are that this allows the classic United States libertarian 'marketplace of ideas' to combat disinformation. The costs are that the government would only carry out research and evaluation, with no carrot-and-stick threat to regulate. Sustainability would be jeopardised by any political calculation that disinformation has overwhelmed the media ecosystem's own established defences. We conclude that the evidence shows substantial failures in the regulatory ecosystem for the media, notably with regard to bot accounts and unregulated online political advertising. Option zero is only effective if the disinformation problem is held to be capable of self-healing by market actors and individuals without the need for more formal coordination, investment or even direct regulation.

Option 1: Non-audited self-regulation

This option would increase platform activity compared with Option zero in terms of preventing immediate regulatory intervention, with increasing industry-government coordination, but no sanction on those companies choosing not to cooperate. Government and private industry research funding could be increased to encourage machine learning-based and other forms of content moderation. The EU code of practice on disinformation proposed under the aegis of the European Commission would continue to be developed. However, the lack of formalised transparency processes (other than reporting) makes this option ineffective and potentially damaging to the European policy process, and thus an unsatisfactory hybrid option as compared to Option zero or Option 2. Any ethical code that becomes an industry standard for certification, especially affecting fundamental rights like algorithmic content recognition, is likely to lead to a call for legislative enforcement.

Option 2: Audited self-regulation

Under audited self-regulation, the self-regulatory scheme is subject to regular independent audit to ascertain the degree to which members are adhering to the criteria. For instance, the EU code of practice would be subjected to formal published audit by a commonly agreed self-regulator. The cost-benefit of audited self-regulation depends on the level of independence and rigour of the auditor function. It allows for lower costs and flexible regulation, but efficiency depends on industry actors' commitment to the independence and rigour of the auditor in the absence of any penalty for lack of compliance, often a fatal failing. 'Free riders' are very likely to exist, though the scale of the larger platforms and the existing code of practice commitments may ensure greater scrutiny. Jiménez Cruz et al. argue that Option 2 is best suited to the current evidence, for a 'structured process ahead that will document progress made and expose anyone not taking their responsibilities seriously'.[2]

Feasibility and effectiveness depend on the implementation of audit. Sustainability of audited self-regulation is very low, given the possibilities for non-compliance identified above. Human rights challenges will exist even with an independent multi-stakeholder board, so that self-audit is inevitably judged inadequate and may be supplanted by formal regulatory bodies. Risks and future uncertainties are very high, and there is no satisfactory example of audited self-regulation on the internet without the backstop of formal regulation. The Global Network Initiative claims such a function, but annual reports do not give sufficient detail to satisfy these criteria.

Option 3: Formal self-regulator

This regulator would be recognised by the European institutions and ideally operate with funding separated from the industry. Recognition does not signal statutory power to intervene or to direct the regulator, but does indicate that the institutions wish to guide the choice of self-regulatory scheme employed, short of intervention via legislation. Applied to AI and disinformation, this schematic would suggest a multistakeholder dialogue establishing general principles applying to an AI regulator, while the self-regulator would set out details of the scheme design. Such principles may include, for instance, the principle that no account can be suspended without human intervention to correct for false positive identification of a bot account, and the potential for account holder appeal against such a deletion. The UN Special Rapporteur on Freedom of Opinion and Expression has recommended such a body.

Similar to Option 2, the cost-benefit of self-regulation is held in general to allow for lower costs and very flexible regulation, though 'free riders' who fail to conform fully may continue to exist. Feasibility and effectiveness depend on the initial design, as well as the implementation of that design by the self-regulator. A problem can be that failure to conform to standards may not be subject to a robust system of audit and correction. Sustainability of self-regulation is always an issue. Risks and future uncertainties are closely tied to the regulatory commitment to making self-regulation an end state (subject to satisfactory independent audit) rather than an interim measure.

Option 4: Formal co-regulation

Formal co-regulation comprises a regulatory system in which the regulator is independent from government, making regulation subject to prior approval of codes of conduct, systems for funding and independent appeal. In Germany, this is known as 'regulated self-regulation'. This is a hybrid system subject to statutory control. Note that this body would censor citizens directly, so the right to appeal to an independent adjudicator must be built in. The regulator could be associated with and certified/approved by state regulatory bodies, such as the EU Fundamental Rights Agency or European Data Protection Board.

Co-regulation offers the statutory underpinning and legitimacy of parliamentary approval for regulatory systems, together with general principles of good regulation, such as independence from regulatees, appeal processes, audit and governance principles. It also devolves the responsibility for these practices to an independent body, which theoretically gives agility and flexibility to the regulator within these general principles. Co-regulation is a good example of the pyramid of regulation, with a statutory tip of regulatory principles and authorisation for the regulator, a co-regulator layer that sets out regulatory design, and a base of industry-shaped rules and codes to provide the detailed implementation.

Coherence with EU objectives are easier to assess with co-regulation than with self-regulation, because the national statutory criteria establishing the co-regulator must conform to European law principles, and ex-post comparative evaluation can be undertaken more easily given these common criteria. The divergence of regulatory means for areas such as child protection and video on demand in European consumer internet law shows that a level of co-existence of different regulatory schemes is possible with national differences.

The cost-benefit of such co-regulation is held in general to allow for lower costs, and more efficient and flexible regulation. However, feasibility and effectiveness depend on the initial statutory design, as well as the the co-regulator’s implementation of that design. Furthermore, similar to the self-regulatory options, sustainability remains an issue. A frequent failing of co-regulation is that it is eventually supplanted by state regulatory bodies, as for instance with video on demand under the Audiovisual Media Services Directive. Although the transition from self-regulation to state regulation is not inevitable, it can result due to pressure from both government and from regulators seeking regulatory certainty. In such situations, the costs of co-regulation can escalate as the scheme attempts to shadow state regulation. Risks and future uncertainties are closely tied to the regulatory commitment to making co-regulation an end state rather than an interim measure.

Potential ethical, social and regulatory impacts revolve around the media pluralism dilemma, where increasing pluralism and diversity with regulation risks the danger that the regulated diversity does not satisfy the users' needs in a free society. The fundamental rights issues with co-regulation are similar to those for less direct regulatory interventions – freedom of expression as a fundamental right may be held inappropriate for anything but state regulation, a constant issue in internet regulation.

Option 5: Statutory regulation

Option 5 consists of a formal regulator tasked to combat disinformation directly by licensing of content providers and their systems for content moderation. Existing electoral and broadcast regulators perform this function for offline media. A merger of many regulators is not necessary to combine the functions via coordinated federated networks of those regulators. Best practice from the various Member States could be collated, analysed and disseminated, ideally by the European Parliament with assistance from the EU Fundamental Rights Agency.[3] The Digital Rights Clearinghouse set up by the EU Data Protection Supervisor with data protection, consumer protection and competition authorities is another example. It may be that in some instances the reform of existing legislation is a more effective and sustainable form of regulation: for instance, where it is not already the case, electoral advertising rules for online media should be brought within the ambit of existing regulators. Incremental improvements may be more compatible with controlling disinformation via AI, than a more disruptive change at this stage.

Focus on freedom of expression and media pluralism

1. We advise that options to ensure **independent appeal and audit** of platforms' regulation of their users be introduced as soon as feasible. When technical intermediaries need to moderate content and accounts, detailed and transparent policies, notice and appeal procedures, and regular reports are crucial. This is also valid for automated removals.
2. We advise against regulatory action that would encourage increased use of AI for content moderation purposes, without **strong human review and appeal processes**.
3. There is scope for standardising (the basics of) notice and appeal procedures and reporting, and creating **a self-regulatory multistakeholder body**. This multistakeholder body could have competence to deal with industry-wide appeals and work towards a better understanding and minimisation of the effects of AI on freedom of expression and media pluralism. We believe this would best fit the Option 3 classification.
4. Disinformation is best tackled through **media pluralism and literacy initiatives**, as these allow diversity of expression and choice. From a free speech and media pluralism perspective, **source transparency indicators** are preferable over (de)prioritisation of disinformation, and users need the opportunity to understand how their search results or social media feeds work, and edit their own search results/feeds where desirable.
5. Finally, noting the lack of independent evidence or even **detailed research** in this policy area, the risk of harm remains far too high for any degree of regulatory certainty. We reiterate that **far greater transparency** must be introduced into the variety of AI and disinformation reduction techniques used by online platforms and content providers.

ENDNOTES

[1] C. Marsden Prosumer Law and Network Platform Regulation: The Long View Towards Creating OffData. *Georgetown Law Technology Review, 2 (2)* 376-398, 2018.

[2] C. Jiménez Cruz, A. Mantzarlis, R. K. Nielsen, and C. Wardle, 'Six Points from the EU Commission's New Report on Disinformation', *Medium*, 12 March 2018.

[3] For FRA activities in this area, see EU Agency for Fundamental Rights, Enabling Human Rights and Democratic Space in Europe, 2018.

This document is based on a study on 'Regulating disinformation with artificial intelligence' (PE 624.279) financed by the European Science-Media Hub (ESMH) budget and published in February 2019. The study was carried out by Prof. Chris MARSDEN, University of Sussex, and Dr Trisha MEYER, Vrije Universiteit Brussel, in response to a request from the Panel for the Future or Science and Technology (STOA) and managed by the Scientific Foresight Unit within the Directorate-General for Parliamentary Research Services (EPRS) of the European Parliament. STOA administrator responsible: Mihalis Kritikos.

DISCLAIMER AND COPYRIGHT

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament. Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged and the European Parliament is given prior notice and sent a copy.

stoa@ep.europa.eu (contact)
http://www.europarl.europa.eu/stoa/ (STOA website)
www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)