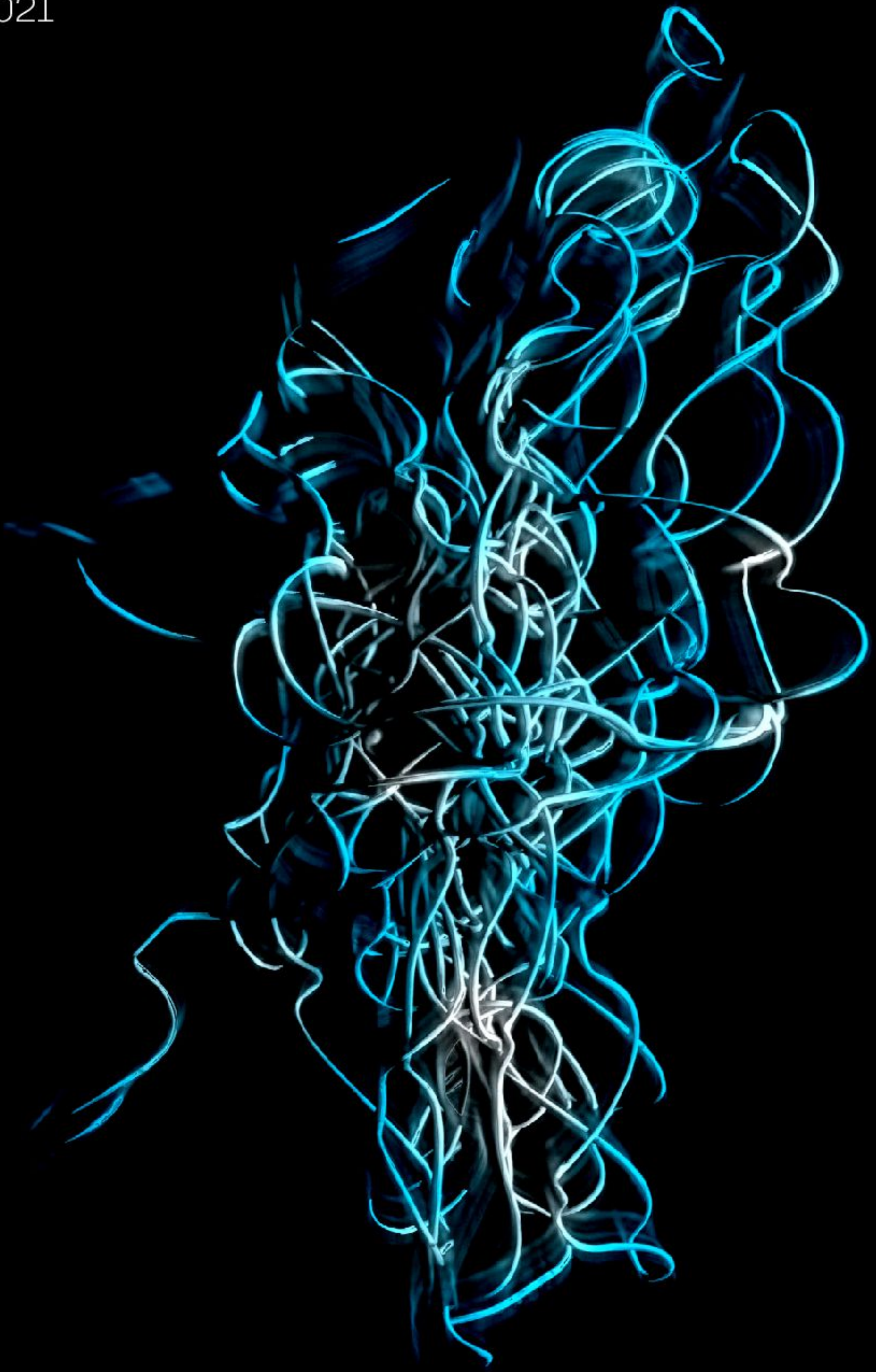
POLICY PAPER

on Artificial Intelligence's (AI) Impact on Freedom of Expression in Political Campaign and Elections

April 2021



OSCE
The Representative on Freedom of the Media

SPOTLIGHT ON
Artificial Intelligence & Freedom of Expression

ELECTION-WATCH.EU

POLICY PAPER

on Artificial Intelligence's (AI) Impact on Freedom of Expression in Political Campaign and Elections

April 2021

Armin Rabitsch[*], Rania Wazir[‡], and Thomas Treml[‡]

[*]Election-Watch.EU

[‡]data4good Initiative of the Vienna Data Science Group (VDSG)

This policy paper was supported by the OSCE Representative on Freedom of the Media Spotlight on Artificial Intelligence and Freedom of Expression (#SAIFE) project.

*Cover designed by Elda Fejzo; cover page image from https://pixabay.com/photos/wire-zoom-sculpture-tangled-art-1990953/

Introduction[1]

Artificial Intelligence (AI), in particular machine-learning and other forms of automated decision-making technologies have a rapidly increasing impact on how citizens seek, receive, impart and access information in political campaigns and elections, especially through online platforms and social media networks. These platforms use AI as part of highly complex and opaque systems to curate content, resulting in a significant impact on freedom of expression that is, however, poorly understood. "There is a genuine risk that such technologies could have a detrimental impact on fundamental freedoms, especially when driven by commercial or political interests. The use of AI, especially algorithmic content curation could seriously jeopardize the enjoyment of our human rights, in particular the freedom of expression. Moreover, given that most AI-powered tools lack transparency and accountability, as well as effective remedies, their increasing use risks exacerbating existing challenges to free speech, access to information and media pluralism."[2]

This Policy Paper captures the current impact of AI on freedom of expression in political campaign and elections, and recommends a human rights based approach to policies and regulatory measures in support of upholding the freedom of expression, the right to political participation as well as an effective civil society access to data for an independent social media monitoring.

While AI also generates chances for participative democracy, like enhanced voter education and – mobilisation, some of the vulnerabilities that have been identified in AI and political competition and elections that ought to be addressed include: misinformation and disinformation[3] campaigns including deep fakes, the amplification and weaponization of hate speech, micro-targeting of voters, racial and gender stereotyping, AI-driven campaigning and the possibility of aspects of electoral processes being malintentionally targeted through automated messaging such as political bots and chatbots.[4] On the one hand these new developments have largely been left to industry self-regulation, meaning regulation by the internet service providers, online platforms and social media networks that manage and host the content themselves. On the other hand, governments increasingly see the role of the state in providing access, and the ability to block, filter or shutdown internet access altogether. Moreover, states often mandate social media platforms to remove specific content, regularly within very short timelines which further incentivizes the use of AI.

The future European public and policy discourse around AI will centre on the European Commission's (EC) proposed two legislative initiatives: the Digital Services Act (DSA) and the Digital Markets Act (DMA). Both aim to encompass a single set of new rules applicable across the whole EU, to create a safer and more open digital space, with European values at its centre and the stated goal of creating a digital space in which the fundamental rights of all users of digital services are protected.[5] Until the DSA/DMA pass the European Parliament and European Council (including Member States) the primary

[1] A Glossary of Terms used is annexed to this text.

[2] OSCE RFoM 2020, Spotlight on Artificial Intelligence and Freedom of Expression; Foreword AI and freedom of speech RFoM; p.7; https://www.osce.org/files/f/documents/9/f/456319_0.pdf

[3] Disinformation is false or misleading information that is created or disseminated with the intent to cause harm or to benefit the perpetrator. The intent to cause harm may be directed toward individuals, groups, institutions, or processes. Misinformation is false or misleading information that is shared without the intent to cause harm or realization that it is incorrect. In some cases, actors may unknowingly perpetuate the spread of disinformation by sharing content they believe to be accurate among their networks.

[4] See the 2016 United Kingdom referendum to leave the EU, the 2016 presidential election in the United States, the 2016 Colombian Peace Referendum, the 2017 French presidential election, the 2018 Brazilian presidential election, and the 2019 general elections in India.

[5] https://ec.europa.eu/digital-single-market/en/digital-services-act-package

legal act governing online platforms in the EU is the E-Commerce Directive which considers that online platforms are not responsible for content, but merely "passing on" content.[6] The tabled new legislations, which stipulate certain safeguards (including the possibility to challenge platforms' content moderation decisions), provides transparency measures (including on the algorithms used for recommendations), includes obligations (like independent audits of their risk management systems), provides access for researchers to key data as well as strengthens oversight by creating a new European Board for Digital Services.[7]

The DSA focusses its attention on platforms' content moderation activities – as defined by Article 2 lit (p), these involve the blocking, demotion or removal of illegal content. However, the platforms' content promotion and micro-targeting activities also deserve scrutiny. As described in the Joint Research Council's Report "Technology and Democracy: Understanding the influence of online technologies on political behaviour and decision-making", these systems have been linked to the spreading of misinformation and conspiracy theories, and given rise to echo chambers and filter bubbles.[8]

International standards[9]

Several resolutions by UN bodies have reaffirmed that "*the same rights people have offline must be protected online*".[10] Comment No. 34 by the Human Rights Committee – which although not legally binding provides interpretative guidance of Article 19 of the International Covenant on Civil and Political Rights (ICCPR) – recognizes that freedom of expression includes "all forms of audio-visual as well as electronic and internet-based modes of expression"[11]. However, the right to political participation not only requires freedom of expression but, as stated by the UN Human Rights Committee, it also presupposes that "(v)oters should be able to form opinions independently, free of violence or threat of violence, compulsion, inducement or manipulative interference of any kind".[12] Furthermore, as in the offline sphere, restrictions and regulatory responses to online freedom of expression should meet the three-part test (should be prescribed by law, pursue a legitimate aim, be necessary and proportionate) outlined in Article 19 of the ICCPR.

Additionally, the ICCPR (Art.2, 25, 26) enshrines the right to non-discrimination and participation of vulnerable groups in public life and requires the prevention of attacks on them.[13] Moreover, the right

[6] DRI, 2021, Tackling Disinformation and Online Hate Speech, p.8, https://democracy-reporting.org/wp-content/uploads/2021/01/Tackling-Disinformation-and-Online-Hate-Speech-DRI.pdf

[7] See: European Commission Digital Services Act Q & A; https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348

[8] https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/technology-and-democracy, pp.48-50

[9] See: Krimmer, R; Rabitsch, A; Kužel, R., Achler, M. Licht, N: 2021, UNESCO/UNDP Elections & Internet, Social Media and Artificial Intelligence (AI): A Guide for Electoral Practitioners (forthcoming)

[10] UN GA resolution of 27 June 2016 on the Promotion, protection and enjoyment of human rights on the Internet", A/HRC/32/L.20, par 1, as well as; UN HRC Resolution 20.8 of 5 July 2012 and 26/13 of 26 June 2014 on the promotion and protection of human rights on the Internet, HRC resolutions 12/6 of 2 October 2009 on freedom of opinion and expression HRC resolution 28/16 of 24 March 2015 on the right to privacy in the digital age, GA resolutions 68/167 of 18 December 2013 and 69/166 of 18 December 2014 on the right to privacy in the digital age and 70/184 of 22 December 2015 on the information and communications technologies for development, amongst others.

[11] CCPR/C/GC/34 General Comment (GC) No. 34 on Article 19 of the ICCPR, par. 12

[12] Human Rights Council General Comment 25, para. 19

[13] See the Report of Special Rapporteur on Freedom of Expression and Opinion, Frank La Rue of 7 September 2012 (A/67/357) and the Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, on hate speech online, 9 October 2019, A/74/486.

to privacy is legally protected by Article 17 of the ICCPR, as "(n)o one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence, nor to unlawful attacks on his honour and reputation."

The UN Guiding Principles on Business and Human Rights, which were endorsed by the Human Rights Council in 2011, reaffirm that governments must guarantee that not only state organs respect human rights, but that companies operating under their territory/jurisdiction do so as well. These principles are also of relevance when considering the role and responsibilities of internet intermediaries like social media companies.[14] A coherent global regulatory framework of social media networks and online platforms would be preferable since national or regional regulations are limited and fall short of covering global and online based activities. However, given the diverse models, like digital authoritarianism systems (e.g., in China, Russia, Iran) the developing of regional legal frameworks of the digital space in liberal democracies like at the EU level are preferable to a plethora of national legislations.

Regulating transparency for algorithmic content curation and AI accountability mechanisms

International obligations of freedom of expression and the right to impart information as well as participate in public life is a prerequisite in any attempt to regulate the cyber sphere. This must always be taken into account, with governments not acting as the arbiters of truth. The distinction between illegal and/or harmful content, in addition to concerns of algorithmic governance and data protection, are necessary to define any regulatory framework. A regulatory framework needs to define the level of accountability, transparency and accessibility requirements for AI and other algorithmic decision systems to uphold and, indeed promote, the freedom of expression in political competition. Transparency and data access are necessary prerequisites for accountability. It is important to highlight that transparency in AI is a *choice*. The term "black box algorithm" is often used as a smoke screen. While the algorithm's functioning may not be understood perfectly[15], it *is* possible to understand what data was put into the algorithm; it *is* possible to observe what decisions an algorithm makes based on that input; and one *can* understand the objective the algorithm is trying to optimize. In fact, while it is possible for humans to hide their intentions, in order to make an algorithm work, the humans developing the algorithm must make their intentions very explicit in the form of a formula or code. That way it can be understood and assessed. Not providing insight into this data/formula/code is a *choice* that is made, not an unalterable fact of nature.

In this context it is required to determine the attribution of rights and responsibilities:

- Who should decide which content should be removed, for which reasons, when and how?
- Who should develop standards for, and implement, content moderation algorithms?
- Who should decide on access to data and code for content-moderation algorithms?
- How do we know what gets deleted, and whether what gets deleted violates laws or not (e.g., on hate speech, violent content, deep fakes, …)? In other words, how do we know that an AI generated content monitor does not mistakenly remove legitimate content, and by removing it thus violates the freedom of expression?

[14] UN Guiding Principles on Business and Human Rights. They were adopted by the UN Human Rights Council in 2011. Available at: https://www.business-humanrights.org/en/un-guiding-principles

[15] very often this, too, is a choice – it is possible to use algorithms that are easier to understand

For all AI generated content[16], monitoring can and *does* make wrong decisions. The indications are that these errors disproportionately impact marginalized communities - in a political context, minority voices get more often silenced and censored[17], while topics of particular interest to them are at higher risk of deletion or demotion.[18]

The problem is actually two-fold – "false positives" and "false negatives". "False positives" means content gets removed that should not have been removed. This could be countered by informing people that their content was removed, and why. Furthermore, means are required of contesting this decision, preferably to an independent third party and/or the judiciary, and not only to the platform itself; with the right of reversing such decisions if law was not violated. At the same time, independent third party researchers and civil society organizations (CSOs) should have access to deleted content, and be able to perform randomized statistical controls, to determine any incorrectly removed content and the percentage of it. Such data access should be stable, well-documented and standardised, to permit the creation of software tools that can assist researchers in downloading and analysing the data. Additionally, the decision on who has access to which data/deleted content should not be left to the platform itself (gate-keeping). This research should be paired with fact checking, to determine whether the removal of content had, or could have had, an impact on the political campaign.

In the case of "false negatives" – i.e., content that does not get removed, but should have been removed[19] – the problem is much bigger. First, this would require access to all content that is published on the platforms. Random spot checks would not be suitable in this case – the percentage of illegal content is small, so that thousands of comments would have to be randomly sampled and manually checked, in order to find just a few tens of illegal comments. What is additionally necessary in this case, beyond access to data, is also the ability to generate an independent corpus of "training data", and train own algorithms to seek out missed content. This avoids relying entirely on the algorithmic filters of the platform owners, and enables the proliferation of filters and technical expertise among independent researchers and CSOs. This is a good in itself, as otherwise allowing only the filters of platform providers would make them the de facto standard setters for what is permissible, and what is not[20]. In fact, allowing platform providers to set the standard for content moderation is highly problematic, as content is judged differently by different people. "For example, men and women make significantly different assessments of sexual harassment online (…) An algorithmic definition of what constitutes inappropriately sexual communication will inherently be concordant with some views, and discordant with others."[21] The track record of algorithmic decisions leaves ample reason to fear that with a one (secret and proprietary) algorithm-fits-all approach, the views and needs of disadvantaged demographics will be neglected.

Furthermore, while the business model of platforms may require that proprietary content promotion algorithms remain obscured, it is not entirely clear why content moderation algorithms should be

[16] Textual, visual, audio content and its combinations

[17] See YouTube removal of footage of human rights abuses by Syrian activists https://www.nytimes.com/2019/10/23/opinion/syria-youtube-content-moderation.html; and research on bias against African American dialect in online hate speech filters: https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter

[18] Measuring and Mitigating Unintended Bias in Text Classification: https://dl.acm.org/doi/10.1145/3278721.3278729

[19] According to international human rights law (for example Art 20 ICCPR) and not the platform's community guidelines.

[20] This is, for example, the situation that currently prevails for the automated detection and deletion of terrorist content: https://www.eff.org/deeplinks/2020/08/one-database-rule-them-all-invisible-content-cartel-undermines-freedom-1

[21] Bender, Gebru et al: "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" http://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf

veiled in secrecy as well. A more equitable proposal would be to encourage the creation of interoperability standards for content moderation, and permit the development of open source, inclusive content moderation algorithms.

Finally, the access to "all content" published on platforms such as *Twitter* has enabled research that demonstrates bias in automated hate speech detection – for example against African American English. The lack of such access on other platforms means that results obtained from *Twitter* are generalized to other platforms, without a strong scientific basis to even indicate that online behaviour and automated filters behave similarly on other platforms. As previously mentioned, content moderation is not the only high-impact area of application of AI. It is also used extensively to decide which messages to promote (recommendation systems), and whom to target. Harvested and mined meta-data is used to create complete *psychometric profiles* of voters, on the basis of the content they post, share or react to on social media platforms or discuss in private messages, as well as taking into account their search histories.[22] Preferably, transparency requirements on all advertising (and not just political), as recommended in the DSA, is essential. In fact, if only political advertising is required, then the question remains open – who will define political advertising, how is it defined[23], and how to verify that all such ads are in fact being recorded.

Finally, the effects of micro-targeting are very hard to measure – from a technical point of view, this would require access to the feeds of a very large random sample of users. In this case, the question, as posed in the European Commission's Joint Research Centre (JRC)[24] report, is: should it really be possible to send a different political message to different people? Perhaps here more than anywhere else, it is not a technical, but a regulatory solution that is required.

Practical implications

Civil society and research institutes can play a vital role as third-party investigators and verifiers of codes of conduct, achieving a balance between complete hegemony over online discourse by private platforms versus complete State control. However, in order to fulfil this mission, access to data, infrastructure, expertise, legal clarity and funding for academic researchers, journalists, and civil society watch dogs is a requirement. These capabilities are *all* needed in order to:

- Assess and research the impact of AI in social media on freedom of speech;
- Independently verify claims made by platforms under their transparency obligations;
- Make research findings reproducible – a key element in increasing the validity and trustworthiness of such results;
- Create the tools and code that allow researchers to work with AI and social media data (to understand the effect of access restrictions, see, for example, the disappearance of many off-the-shelf tools that accompanied the Facebook restrictions on API access[25]);
- Create algorithmic tools that can help detect illegal or harmful content.[26] Algorithmic filters are needed to assist in such monitoring work, and hold politicians, political parties and political lobby groups, et.al. accountable for the public content they produce;

[22] Blesik, Murawski, Vurucu, & Bick, 2018

[23] For example, how to account for situations where climate activist posts calling for restrictions on burning of fossil fuels are labelled as political advertising, but the advertisements by oil & coal companies are labelled as private advertising?

[24] https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/technology-and-democracy

[25] https://theconversation.com/facebooks-data-lockdown-is-a-disaster-for-academic-researchers-94533

[26] For example, during the monitoring of the Austrian parliamentary elections 2019, more than 1 million of interactions/posts by/with politicians were collected, a volume that is difficult to check by traditional manual-only means.

- Investigate possible biases in existing algorithmic content filters regarding disinformation (fake news), terrorist content, etc.;
- Prevent the creation of algorithmic "de-facto standards" that are developed by private (commercial interest driven) entities, with no outside oversight, but with the power to decide what is acceptable content.

Any activities to investigate the effect of social media on freedom of speech require data. Be it the effects of hate speech on silencing minority voices; the algorithmically mandated deletion of content; the creation of filter bubbles/echo chambers; the spread of disinformation ("fake news"); the propagation and scaling up of gender and ethnic stereotypes and biases through content moderation/content promotion – the extent of these problems threatening fair political campaigns and genuine elections can best be investigated via access to data – and large quantities of it. In order for results to be scientifically valid, the researchers need to have access to a large enough sample, be able to assess its quality (e.g., are there missing or incorrect values, outliers), and determine how the sample is collected[27] (e.g., via random sampling from the appropriate population). Furthermore, if available to the platform provider, the sample should contain all necessary features to answer a research question, for example:

- When investigating the difference in hate speech content between Austrian provinces, if the data obtained through API only contains country location (even though the provider has more precise location information), then this would be useless for the research.
- When a "vetted researcher"[28] is assessing bias/discrimination on social media, access to "sensitive" data such as gender, race, age, may be required, and should be made available, under compliance with the EU General Data Protection Regulation (GDPR), if the platform has this data.

However, CSOs that have traditionally served as watchdogs to preserve fundamental rights face unprecedented challenges in acquiring and processing the requisite data on social media platforms. This has led, for example, to an over-reliance on *Twitter* data for investigating social media behaviour in general[29], and to difficulties in proving that certain content moderation/promotion behaviours even occur[30]. Furthermore, lack of legal clarity restricts the willingness of researchers to share data for making results reproducible, or for undertaking certain kinds of research[31].

Social media data access for civil society and research organizations presents a series of problems, many of which become visible only in practice. Election-Watch.EU and VDSG/data4good have documented how to request API access to *Twitter* and *Facebook* (see technical guidance documentation), but many platforms (like Telegram) do not even provide API access or restrict it severely (like *Discord*, or *Instagram*) for research purposes. In particular, *TikTok*, *WhatsApp*, *Discord* and Instagram do not allow public pages access. So even if a page is public, and the comments clearly visible to anyone surfing the web, data is only available for page owners/administrators (and of course,

[27] Algorithm Watch https://algorithmwatch.org/en/story/research-data-quality/
[28] DSA, Article 31 (4)
[29] Algorithm Watch https://algorithmwatch.org/en/story/data-access-researchers-left-on-read/
[30] Algorithm Watch https://algorithmwatch.org/en/story/instagram-algorithm-nudity/; see also https://algorithmwatch.org/en/governing-platforms-final-recommendations/
[31] Wahlbeobachtung.org, 2020, Social Media Monitoring Early Parliamentary Election Campaign Austria 2019 Final Report, & sentiment analysis, https://www.wahlbeobachtung.org/wp-content/uploads/2020/02/smm-austria-wahlbeobachtung.org-final-report-030220.pdf

also the platform providers themselves). This means data can only be obtained by much more challenging scraping techniques, whose legality is questionable. Additionally, there exists a problem with the relatively new feature of ephemeral content (such as *SnapChat*, or *Instagram Stories*), which is completely opaque to researchers even with API access. Such features are not only short-lived, but often also combine images or videos with texts and sometimes even music, which suggests that besides attracting viewers' attention their potential for opinion formation is rather high. On the other hand, they can be an opportunity for democratic opinion formation too, if users are enabled to reference content of other users in their posts[32].

There is a technical distinction to be made between social media platforms (like *Twitter, Facebook, Instagram, YouTube, TikTok*), and messaging apps (like *SnapChat, WhatsApp, Telegram, Signal*). Messaging apps are, by construction, private. They are intended for the sending of messages between participants, and should be end-to-end encrypted, so that only the sender and the intended recipient can read the message. Serious questions arise, however, when such messaging apps are used to convey content to hundreds of thousands of people: When does the private scope stop, and public sphere begin, and how much reach should a message have, before being considered public content and thus subject to scrutiny? Additionally, on the social media platforms, there are public pages, and private pages. Public pages are those whose content is freely visible to anyone on the internet – no need to log in, or be a friend/follower of that page. Data from these pages should be accessible via API for third party research. However, human rights monitors and election observers face huge constraints in scrutinizing those profiles which are private, even while some or most content is shared with the public. Again, the question of where the private sphere ends, and public data begins, requires some consideration.

Even if API access is available, obtaining it can take a long time, and incur large administrative and bureaucratic hassles[33]; constantly changing rules of access and data handling require an intense and constant maintenance effort; the EU's General Data Protection Regulation (GDPR) and Copyright regulations add to the unstable landscape, and create further confusion and inconsistency across the EU. For example, the recent EU Court of Justice ruling cancelling the Privacy Shield agreement[34], left organizations with sensitive data to store scrambling to find EU-based alternatives. In addition, uncertainty about what is or is not permissible data collection and analysis often prevents civil society organisations from pursuing certain lines of research.[35] Furthermore, little known technical requirements – for example, *Facebo*ok's linkage of permissible data download volume to the number of daily users an app has, effectively limits the usefulness of API access for research purposes, even once obtained[36].

[32] E.g. within the duet feature of TikTok (see Medina Serrano, 2020: p 8)

[33] See our API guidance

[34] The ruling that privacy shield is invalid is seen as a victory for privacy and data protection campaigners and has complex implications for data sharing between the EU and the US; see: https://curia.europa.eu/juris/liste.jsf?num=C-311/18

[35] For example, sentiment analysis on politicians' statements, which, combined with a topic analysis, could lead to the inferral of the politicians' political affiliation, a protected attribute by Austrian privacy law.

[36] In order to collect data from public pages, it is normally necessary to create a server-to-server app, which is an app that does not have users, and whose sole purpose is to collect data through the Facebook APIs. By linking the download volume to the number of users, this forces even server-to-server apps to enjoin a sufficiently large number of users, who will add the app to their personal Facebook account, and will use it daily.

Way forward and recommendations

The objectives of this policy paper on "Artificial Intelligence's (AI) Impact on Freedom of Expression in Political Campaign and Elections" are to:
1) influence the upcoming AI policy regulations at EU level;
2) provide a technical framework to simplify the work of organizations seeking to assess and research the effects of AI and social media on freedom of expression in political competition and elections.

Departing from the experience of conducting a social media monitoring project during the Austrian parliamentary elections 2019[37], and following the webinar on "Artificial Intelligence's (AI) Impact on Freedom of Expression in Political Campaign and Elections"[38] the following recommendations can be summarised:

Policy makers, politicians and legislators[39]

- Establish a clear and strong European legal framework on AI (DSA/DMA) that ensures respect for freedom of expression online and offline, privacy and the right to participate in public affairs, in line with international standards agreed upon in international and regional treaties. Refrain from unduly restricting freedom of expression, taking into account that any limitation of this right should be in accordance with the three-part test of legality, legitimate purpose and necessity. Regulatory responses to these challenges that are worded in a vague manner are not compatible with freedom of expression.
- Regulations should enshrine the key principles of accountability and transparency as well as promulgate the privacy-preserving access for researchers, fact checking initiatives and CSOs[40] to assess AI and freedom of expression in online political campaigns. In particular: a) clear guidelines for "vetted researchers", applicable in *all* EU countries, establishing what data may be collected, how processed and published; b) infrastructure (data storage, compute) to enable researchers – especially journalists and CSOs, to store and process data in strict compliance with GDPR and other privacy regulations; c) together with industry, academics and CSO stakeholders, establish industry standards for API access, to encourage the development of off-the-shelf tools for data research (think of it like a plug for electricity – no matter who your energy provider, the plug always looks the same: this should hold for API access as well) d) together with industry, academics and CSO stakeholders, establish industry standards for content moderation algorithms, as well as ensuring such algorithms are open source, and open to verification and observation by vetted researchers.
- Ensure that not only state organs, but also businesses operating under their territorial jurisdiction respect human rights, as per the Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework.
- Refrain from delegating online content regulation to social media companies and other internet intermediaries beyond what is allowed under international standards. Internet intermediaries

[37] Wahlbeobachtung.org, 2020, Social Media Monitoring Early Parliamentary Election Campaign Austria 2019 Final Report, https://www.wahlbeobachtung.org/wp-content/uploads/2020/02/smm-austria-wahlbeobachtung.org-final-report-030220.pdf

[38] https://www.wahlbeobachtung.org/en/webinar-impact-of-ai-on-freedom-of-expression-in-political-competition-2/

[39] See: Krimmer, R; Rabitsch, A; Kužel, R., Achler, M. Licht, N: 2021, UNESCO/UNDP Elections & Internet, Social Media and Artificial Intelligence (AI): A Guide for Electoral Practitioners (forthcoming)

[40] Where the DSA does not go far enough: Chapter 3, Article 31 – „vetted researchers" includes only academics, but not journalists or CSOs.

should not be held liable for content produced and posted by a third party using their services, except if they intervene in that content[41] or fail to follow a decision requesting its removal emitted by an independent, impartial, oversight body (e.g. a court) despite having the technical capacity to do so. Governments should take into account the risks to freedom of expression involved in imposing social media companies with liability for user-generated content, given that it can incentivise excessive content moderation and removal, including through proprietary and opaque automated mechanisms.

- Ensure that the fundamental right to privacy is respected online. Personal data should not be used without consent, including for the purposes of political advertising and micro-targeting. Default consent to third-party companies, and cross-platform tracking, should be prohibited. Individuals should be informed in a clear and transparent manner why they are seeing certain advertisements and who has paid for them. Thus, all online advertisements should be publicly available and easily accessible and searchable, with detailed information stating who bought them, the source of the funds involved, how much was spent on them, how many users were reached, and the specific targeting parameters that were used. Comprehensive data protection laws must be implemented and enforced and any loopholes that could be exploited by political campaigns should be closed. It is also important to put in place measures to enforce detailed and timely reporting to electoral authorities on campaign financing and advertising.
- Support the use of AI to combat disinformation, for instance through publicly financed, certified and controlled social bots that could undertake automated analysis of online content, accompanied by human curation to verify content. Ensure that the use of these tools is transparent and consistent with human rights obligations, including by putting in place the necessary safeguards so that they are not dependent of any political party, candidate or interest group.

Corporations using AI in Political Campaign and Elections online[42]

- Use algorithms that are based on ethical benchmarks and the respect for fundamental rights; they should be consistent with global standard setting-instruments that are being developed following a multi-stakeholder approach, such as the UNESCO's Recommendation on the Ethics of AI (currently under preparation) and the ongoing work of the Council of Europe's Ad Hoc Committee on Artificial Intelligence towards a legal framework for the design, development and application of AI.
- Facilitate researchers' and independent observers' access to algorithms, especially used for content moderation, to ensure that they meet the requisite standards of ethics and transparent use. Support fact-checking initiatives and independent journalism, as well as research to expand knowledge about disinformation and the effectiveness of responses to it including their own. In regard to this last aspect, further facilitate access to their API. In order to help advance sound research and oversight of electoral/political campaigns, provide better, more precise and more coherent data to accredited election observers and researchers.

[41] Note that EC DSA, Ch.2, Article 3, defines intervention for "mere conduits" as either a) selecting the recipient, or b) selecting/modifying the content, but is much more lenient with "hosts" (Article 5), which includes online platforms. In fact, *only* very large online platforms are required to undertake a risk assessment, and implement risk mitigation measures that *might* include re-design of their content promotion or recommendation systems. However, general online platform liability for user-generated content should be treated separately from platform liability for content that they promote or target. While the first liability scheme bears clear risks of massive content removal by platforms, and consequent user loss of freedom of speech, the second warrants at least a discussion of whether and how it could viably be enforced.

[42] See: Krimmer, R; Rabitsch, A; Kužel, R., Achler, M. Licht, N: 2021, UNESCO/UNDP Elections & Internet, Social Media and Artificial Intelligence (AI): A Guide for Electoral Practitioners (forthcoming)

- Commit to the creation of open interoperability standards for content moderation, developed jointly with all stakeholders involved that permit the deployment of open source AI-tools to: a) support in the identification of factually wrong content, b) inform users about it, c) create awareness of verified information, and d) diversify political discourse through the infiltration of echo-chambers and filter bubbles. In light of the risks involved in over-relying on automation, notably in connection to freedom of expression, ensure that meaningful and effective human review accompanies the use of any AI tools for content moderation.
- Review online advertising models to ensure that they do not adversely impact on the diversity of opinions and ideas, as well as establish open access political ad libraries featuring structured and transparent information on micro-targeting and digital platforms' advertising regulations, including data on those third parties that try to circumvent the rules and not only those who play by them. This requires different solutions for different platforms, and should provide observers, researchers, media and interested electoral stakeholders with real time information about online political ads.
- Respect minimum due process guarantees when taking content moderation and removal actions: promptly notify users when content that they created, uploaded or host may be subject to an action of this kind and explain the rationale behind the decision; give the user an opportunity to contest it, subject only to legal or reasonable practical constraints; carefully scrutinize claims made under content moderation policies before taking action, and apply the related measures consistently.

Glossary:

Algorithm is a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer. In Machine Learning (ML) or Artificial Intelligence (AI) it is "a procedure that is run on data to create a machine learning "model"."[43]

Artificial Intelligence (AI), the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience.[44]

Bot is a piece of software that can execute commands, reply to messages, or perform routine tasks, as online searches, either automatically or with minimal human intervention. While some bot traffic is from good bots, bad bots can have a negative impact on a website or application.[45]

Chatbot is an artificial intelligence (AI) application that can imitate a real conversation with a user in their natural language. Chatbots enable communication via text or audio on websites, messaging applications, mobile apps, or telephone.[46]

Dis-information is false and misleading information that is created or disseminated to harm or to benefit a person, social group, organization or country."[47] The motivations underlying it could be to make financial profit, to have foreign or domestic political influence, or simply to cause trouble.[48]

[43] https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/
[44] Zuiderveen Borgesius et al.; 2018, Bullock, 2019; Chen et al., 2019 & https://www.britannica.com/technology/artificial-intelligence
[45] https://www.dictionary.com/browse/bot, https://www.cloudflare.com/learning/bots/what-is-a-bot/
[46] https://sendpulse.com/support/glossary/chatbot
[47] https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c. p. 20.
[48] https://firstdraftnews.org/wp-content/uploads/2019/10/Information_Disorder_Digital_AW.pdf?x19182 , p 8

Echo chambers are the (digital) environment in which a person encounters only beliefs or opinions that coincide with their own, so that their existing views are reinforced and alternative ideas are not considered.

Information technology (IT) is the use of computers, storage, networking and other physical devices, infrastructure and processes to create, process, store, secure and exchange all forms of electronic data. The commercial use of IT encompasses both computer technology and telephony.[49]

Machine Learning (ML) is "the process of computers changing the way they carry out tasks by learning from new data, without a human being needing to give instructions in the form of a program"[50].

Malinformation is accurate information that is shared with the intent to cause harm or to benefit the perpetrator, often by moving private information into the public sphere.

Misinformation is false or misleading information that is shared without the intent to cause harm or realization that it is incorrect. In some cases, actors may unknowingly perpetuate the spread of disinformation by sharing content they believe to be accurate among their networks.

Microtargeting can be defined as a "marketing strategy that uses people's data — about what they like, who they're connected to, what their demographics are, what they've purchased, and more — to segment them into small groups for content targeting."[51]

Model is the actual output of a Machine Learning algorithm run on data. It incorporates and represents what was leaned by the algorithm.[52]

Online platforms include a range of services available on the Internet including marketplaces, search engines, social media, creative content outlets, app stores, communications services, payment systems, etc.

Psychometric profiling is a method of collecting data to measure psychological characteristics such as our abilities, attitudes and personality.

Social bot is an agent that communicates more or less autonomously on social media, often with the task of influencing the course of discussion and/or the opinions of its readers.[53]

Social media are web or mobile-based platforms that allow for two-way interactions through user-generated content (UGC) and communication. Social media are therefore not media that originate only from one source or are broadcast from a static website. Rather, they are media on specific platforms designed to allow users to create ("generate") content and to interact with the information and its source. [54]

[49] https://searchdatacenter.techtarget.com/definition/IT
[50] https://dictionary.cambridge.org/dictionary/english/machine-learning
[51] https://blog.mozilla.org/internetcitizen/2018/10/04/microtargeting-dipayan-ghosh/
[52] https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/
[53] Ferrara, Emilio; Varol, Onur; Davis, Clayton; Menczer, Filippo; Flammini, Alessandro (July 2016). "The Rise of Social Bots". Communications of the ACM. 59 (7): 96.
[54] Kaiser, S. (2014). Social Media A Practical Guide for Electoral Management Bodies. (p.11). International Institute for Democracy and Electoral Assistance (IDEA). https://www.idea.int/sites/default/files/publications/social-media-guide-for-electoral-management-bodies.pdf.