

Mapping Report D2.3: Policies and risk-mitigation measures on election disinformation by Wikipedia

June 2025

Authors: Ronan Fahy (University of Amsterdam), Gabriela Trogrlić (University of Amsterdam), Lydia Koraki (University of Amsterdam), and Joris van Hoboken (University of Amsterdam)

June 2025

1. Introduction

This Mapping Report (D2.3) examines Wikipedia’s policies and risk-mitigation measures that are applicable to disinformation during elections in the EU. It is the second legal mapping report of the DEM-Debate project (Building an Enabling Environment for Democratic Debate: Insights from community-governed platforms to cultivate a resilient election information ecosystem in Europe), an interdisciplinary research project examining how to increase resilience of the online information ecosystem to safeguard informed civic participation. In particular, the DEM-Debate project examines how certain community-governed platforms tackle election disinformation, based on a case study of Wikipedia during the 2024 European Parliament elections. Based on this research, the DEM-Debate project seeks to develop new policy approaches to contribute to ensuring the information ecosystem surrounding elections in the EU is sufficiently insulated from the harmful effects of disinformation, with the research seeking to contribute to building an “enabling environment” for democratic debate.¹

As noted in the first Mapping Report (D2.1), legal research has mainly focused on the role of large commercial online platforms, operating a centralised-governed model (TikTok, Instagram, YouTube, and X), in the dissemination of disinformation during elections, and current European legal policy is mainly focused on these online platforms.² However, there has been less in-depth European legal analysis given to examining (non-commercial) community-governed platforms, and how they tackle disinformation during elections in the EU, and how European disinformation regulation applies to these platforms. Building on work by scholars such as Grimmelmann,³ Rozenshtein,⁴ and Seering,⁵ community-governed platforms broadly encompass those online

¹ See Council of Europe Steering Committee for Media and Information Society, *Guidance Note on countering the spread of online mis- and disinformation through fact-checking and platform design solutions in a human rights compliant manner* (Council of Europe, 2023), p. 35, <https://edoc.coe.int/en/internet/11885-guidance-note-on-countering-the-spread-of-online-mis-and-disinformation-through-fact-checking-and-platform-design-solutions-in-a-human-rights-compliant-manner.html>.

² See, for example, OECD, *Facts not Fakes: Tackling Disinformation, Strengthening Information Integrity* (2024), https://www.oecd.org/en/publications/2024/03/facts-not-fakes-tackling-disinformation-strengthening-information-integrity_ff96d19f.html; M. Husovec, “The Digital Services Act’s red line: what the Commission can and cannot do about disinformation” (2024) *Journal of Media Law*, 16(1), 47; S. Galantino, “How Will the EU Digital Services Act Affect the Regulation of Disinformation?” (2023) *SCRIPTed*, 20(1), 89; A. Strowel and J. De Meyere, “The Digital Services Act: transparency as an efficient tool to curb the spread of disinformation on online platforms?” (2023) 14 *JIPITEC* 66; and J. van Hoboken and R. Ó Fathaigh, “Regulating Disinformation in Europe: Implications for Speech and Privacy” (2021) 6 *UC Irvine Journal of International, Transnational, and Comparative Law* 9.

³ J. Grimmelmann, “The Virtues of Moderation” (2015) 17 *Yale Journal of Law and Technology* 42.

⁴ A. Rozenshtein, “Moderating the Fediverse: Content Moderation on Distributed Social Media” (2023) *Journal of Free Speech Law* 217.

⁵ J. Seering, “Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation”, (2020) *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4, 1. See also, J. Seering & S. Kairam, “Who Moderates on Twitch and What Do They Do?: Quantifying Practices in Community Moderation on Twitch”, *Proceedings of the ACM on Human-Computer Interaction*, Vol. 7, 1.

platforms where content moderation is *generally* not undertaken in a centralised “top-down” approach, but rather, is “user-led moderation” undertaken by a community of users of the platform in a generally decentralised manner.⁶

The DEM-Debate project aims to address this gap by examining one such community-governed platform - Wikipedia – as a case study, and examining how it addresses disinformation during European elections, and how the European legal framework on election disinformation applies to such community-governed platforms. This is an opportune time to engage in this study, given that Wikipedia was recently designated as a so-called Very Large Online Platform (VLOP) under a new landmark piece of EU legislation, which addresses the risks associated with online disinformation, namely, the Digital Services Act (DSA).⁷ Indeed, Wikipedia is the only non-profit community-governed platform designated as a VLOP under the DSA, alongside the main commercial online platforms, such as TikTok, Instagram, YouTube, and X.⁸

The DEM-Debate project’s legal research proceeds in four phases: first, a mapping of the EU regulatory frameworks on online disinformation during elections that are applicable to community-governed platforms (Mapping Report D2.1); second, an examination of Wikipedia’s policies on tackling disinformation during elections in the EU, and its risk mitigation measures targeting election disinformation under the DSA; third, a critical assessment of the application of EU regulatory frameworks on disinformation during elections to community-governed platforms, using Wikipedia as a case study; and fourth, developing policy recommendations for specific legislative and regulatory reforms of the EU regulatory framework to better counter disinformation during elections, building upon the findings of previous phases in relation to community-governed platforms’ approach to disinformation.

As such, the purpose of this second Mapping Report (D2.3) is to provide an overview of the policies and risk-mitigation measures of Wikipedia that are applicable to election disinformation in the EU. It is a crucial step to set out Wikipedia’s main policies that are applicable to election disinformation, which will form the basis for critically examining, in stage three of the research, the application of EU regulatory frameworks on disinformation during elections apply to community-governed platforms, using Wikipedia as a case study. Notably, this second phase of the research is not designed to provide an assessment of whether Wikipedia’s policies applicable

⁶ A. Rozenshtein, “Moderating the Fediverse: Content Moderation on Distributed Social Media” (2023) *Journal of Free Speech Law* 217.

⁷ Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). See generally, J. van Hoboken et al., Putting the DSA into Practice (Verfassungsbooks, 2023), https://www.ivir.nl/publicaties/download/vHoboken-et-al_Putting-the-DSA-into-Practice.pdf.

⁸ See, European Commission, “DSA: Very large online platforms and search engines”, <https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops>.

to election disinformation in the EU are successful or not in tackling election disinformation.⁹ The purpose of this research is to set out the policies Wikipedia has in place that are applicable to election disinformation in the EU, and the manner in which a community-governed platform designated as a VLOP under the DSA addresses election disinformation in the EU.

The Report is structured as follows: first, in Section 2, some introductory points on Wikipedia and the legal framework are set out. In Section 3, Wikipedia’s policies and principles applicable to election disinformation in the EU are examined, including policies on verifiability of facts, prohibition of original research, neutrality policies, policies on political biographies, and policies on elections; the Wikimedia Foundation’s Universal Code of Conduct and its operation,¹⁰ and the Wikimedia Foundation’s Terms of Use.¹¹ Notably, given Wikipedia’s designation as a VLOP under the DSA, this review will have a unique opportunity to examine how designation as a VLOP, and the consequent additional rules under the DSA in relation to systemic risks related to disinformation during elections, are applicable to community-governed platforms.

2. Wikipedia, the Digital Services Act, and election disinformation

Before examining Wikipedia’s policies applicable to election disinformation in the EU, a number of preliminary points need to be made in relation to the EU’s legal framework on election disinformation, and how Wikipedia (and the Wikimedia Foundation) and its policies fit within this framework. First, it is important to mention the distinct legal position of Wikipedia in the EU. And under the EU’s landmark Digital Services Act (DSA) legislation, which is designed to address the “societal risks” that the “dissemination of disinformation” may generate,¹² Wikipedia has been designated as a so-called Very Large Online Platform (VLOP) by the European Commission.¹³ This means that Wikipedia is subject to the special systemic-risk provisions (Article 34 and 35) under the DSA, which require VLOPs to assess any systemic risks stemming from their services, including relating to negative effects on “electoral processes”,¹⁴ which may stem from the dissemination of

⁹ In a US election context, scholars have been examining how the Wikipedia community tackled misinformation during recent US presidential elections: see Formisano et al., “Counter-Misinformation Dynamics: The Case of Wikipedia Editing Communities during the 2024 US Presidential Elections” (2024), <https://ssrn.com/abstract=4990973>. Relatedly, on how Wikipedia was “proving successful in the fight against medical misinformation in a global pandemic”, see B. Avieson, “Editors, sources and the ‘go back’ button: Wikipedia’s framework for beating misinformation” (2022) 27(11) *First Monday* 1, <https://doi.org/10.5210/fm.v27i11.12754>.

¹⁰ Wikimedia Foundation, “Wikimedia Foundation Universal Code of Conduct”, https://foundation.wikimedia.org/wiki/Policy:Universal_Code_of_Conduct.

¹¹ Wikimedia Foundation, “Wikimedia Foundation Terms of Use”, https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use#introduction.

¹² Digital Services Act, Recital 9.

¹³ See European Commission, Commission Decision designating Wikipedia as a very large online platform in accordance with Article 33(4) of Regulation (EU) 2022/2065 of the European Parliament and of the Council, 24 March 2023, C(2023) 2742 final.

¹⁴ Digital Services Act, Article 34(1)(c).

disinformation; and for VLOPs to put in place “mitigation measures” to address these systemic risks.¹⁵ In this regard, VLOPs must publish a risk assessment and risk-mitigation report, which is, importantly, subject to an independent audit, with the auditor publishing an audit report, and an audit implementation report, with recommendations for the VLOP.¹⁶ As such, Wikipedia (and the Wikimedia Foundation) is subject to a unique legal obligation under EU legislation to publicly disclose risks relating to election disinformation stemming from its system, and publicly disclose the risk-mitigation measures it has in place to mitigate these risks; and all subject to an independent audit.

As mentioned in the first Mapping Report (D2.3), Wikipedia is the only non-profit community-governed platform designated as a VLOP under the DSA, alongside the main commercial online platforms, including TikTok, Instagram, Facebook, YouTube and X.¹⁷ This Mapping Report uses the term community-governed platform to broadly encompass those online platforms where content moderation is *generally* not undertaken in a centralised “top-down” approach, but rather, is “user-led moderation” undertaken by a community of users of the platform in a *generally* decentralised manner.¹⁸ Crucially, however, community-governed platforms may not be fully decentralised, in the sense that some content moderation decisions may ultimately be undertaken by a central host or provider of the platform, and as such, content moderation is “generally” community-led. In this sense, Wikipedia is a community-governed platform, and in particular, a user-edited encyclopaedia, which is written, maintained, and moderated by a community of thousands of volunteers.¹⁹ Indeed, Wikipedia is an important source of public interest information in the EU; with even the Council of Europe’s Commissioner for Human Rights, before the European Court of Human Rights, emphasising the “crucial importance” of Wikipedia throughout Europe in terms of the fundamental right of access to information, and a “crucial online resource” for millions of individuals.²⁰ Notably, the Wikimedia Foundation is a non-profit foundation which hosts Wikipedia, and while Wikipedia’s community of volunteers engages in most content moderation, the Wikimedia Foundation can also engage in content moderation, in certain rare instances.²¹

¹⁵ Digital Services Act, Article 35.

¹⁶ Digital Services Act, Article 37.

¹⁷ See, European Commission, “DSA: Very large online platforms and search engines”, <https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops>.

¹⁸ A. Rozenshtein, “Moderating the Fediverse: Content Moderation on Distributed Social Media” (2023) *Journal of Free Speech Law* 217.

¹⁹ Wikipedia, Wikipedia, <https://en.wikipedia.org/wiki/Wikipedia>.

²⁰ Council of Europe Commissioner for Human Rights, “Third party intervention by the Council of Europe Commissioner for Human Rights, Application no. 25479/19, Wikimedia Foundation, INC. v. Turkey, CommDH(2019)28, p. 6, <https://rm.coe.int/third-party-intervention-before-the-european-court-of-human-rights-app/168098e542>. See also, <https://www.ohchr.org/sites/default/files/documents/issues/expression/cfis/sustainable-development/subm-sustainable-development-freedom-of-expression-csos-wikimedia-foundation-51.pdf>.

²¹ Wikimedia Foundation, ‘Office Actions Policy’, https://foundation.wikimedia.org/wiki/Policy:Wikimedia_Foundation_Office_Actions_Policy.

Second, Wikipedia's VLOP designation is also important for how Wikipedia and the Wikimedia Foundation is viewed under the DSA. Importantly, the European Commission in its VLOP designation Decision considered that Wikipedia is a "hosting service" within the meaning of Article 3(g)(iii) DSA, that "stores and disseminates information to the public at the request of recipients of its service," and is therefore an "online platform" within the meaning of Article 3(i) DSA.²² Notably, the Commission considered that the Wikimedia Foundation is the legal entity "which alone operates" Wikipedia, and as such the Commission's designation was addressed to the Wikimedia Foundation.²³

Notably, in late 2024, the Wikimedia Foundation published its first DSA risk assessment,²⁴ which sets out "the Wikimedia Foundation's assessment of systemic risks linked to the use of Wikipedia in the EU", and "require[d] the Foundation (as the hosting provider of Wikipedia) to not only assess risks, but also determine what risk mitigations are appropriate".²⁵ This risk assessment includes important risks and mitigation measures relating to Wikipedia and disinformation and electoral process in the EU, and will be discussed below. Further, the independent audit report,²⁶ and audit implementation report,²⁷ have also been published, which also specifically address mitigation measures for election disinformation risks, and are also discussed below.

The foregoing paragraphs raise a larger point about the relationship between the Wikimedia Foundation and Wikipedia, and in particular, the "volunteer editor community of Wikipedia",²⁸ and the rules and policies of the Wikimedia Foundation, and the Wikipedia community rules and policies. The auditors have noted that a "clear distinction" needs to be drawn between the Wikimedia Foundation as an "independent organisation," and the "volunteer editor community of Wikipedia, which operates in a decentralised manner" separate from the Wikimedia Foundation.²⁹ Thus, the Wikimedia Foundation "establishes the governance mechanisms within which the community acts", such through the Universal Code of Conduct, "which detail a baseline of

²² European Commission, Commission Decision designating Wikipedia as a very large online platform in accordance with Article 33(4) of Regulation (EU) 2022/2065 of the European Parliament and of the Council, 24 March 2023, C(2023) 2742 final, section 1.

²³ European Commission, Commission Decision designating Wikipedia as a very large online platform in accordance with Article 33(4) of Regulation (EU) 2022/2065 of the European Parliament and of the Council, 24 March 2023, C(2023) 2742 final, section 6.

²⁴ Wikimedia Foundation, "WMF 23-24 approved DSA SRAM (Risk Register)", [https://foundation.wikimedia.org/wiki/File:WMF_23-24_approved_DSA_SRAM_\(Risk_Register\).zip](https://foundation.wikimedia.org/wiki/File:WMF_23-24_approved_DSA_SRAM_(Risk_Register).zip).

²⁵ Wikimedia Foundation, "2023 EU Systemic Risk Assessment - Cover note", 31 August 2023, https://upload.wikimedia.org/wikipedia/foundation/5/5a/Wikipedia_DSA_SRA_submission_cover_note_31Aug2023.pdf

²⁶ Wikimedia Foundation, "Wikipedia DSA Audit Report 2023-24 Public.pdf", https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_Audit_Report_2023-24_Public.pdf.

²⁷ Wikimedia Foundation, "Wikipedia DSA Audit Implementation Report 2023-24 Public.pdf", https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_Audit_Implementation_Report_2023-24_Public.pdf.

²⁸ Wikimedia Foundation, "Wikipedia DSA Audit Report 2023-24 Public.pdf", p. 9.

²⁹ Wikimedia Foundation, "Wikipedia DSA Audit Report 2023-24 Public.pdf", p. 9.

behaviour for collaboration on Wikimedia projects worldwide and the types of cases that fall within each of their scopes”; while “each community can go further to ‘develop policies that take account of local and cultural context, while maintaining the criteria listed here as a minimum standard’.”³⁰ Crucially, the auditors note that the only actions within the scope of the audit “are those conducted directly” by the Wikimedia Foundation, and “not the actions of the volunteer community.”³¹ For the purpose of this Mapping Report, and given its link to the DSA’s legal framework for addressing election disinformation, this report also focuses on the policies and risk mitigation measures of the Wikimedia Foundation. However, this Mapping Report also examines the policies of Wikipedia, with a particular focus on English Wikipedia, and its rules on verifiable accuracy, reliable sources, and neutral point of view;³² and gives a somewhat broader perspective. This is essential to fully understand how Wikipedia addresses election disinformation.

Finally, in order to examine Wikipedia’s policies applicable to election disinformation in the EU, it is important to set out a definition of disinformation. And as mentioned in the first Mapping Report, there is no legal definition of disinformation under EU legislation, including the DSA. However, there are specific policy definitions that are used by EU institutions and regulatory bodies, such as the European Commission. And for the purposes of this second Mapping Report, a widely-adopted definition is that of the European Commission, which defines disinformation as “false or misleading content that is spread with an intention to deceive or secure economic or political gain and which may cause public harm”.³³ Notably, this definition is used in the Code of Conduct of Disinformation, which has recently (in early 2025) taken on a new significance, as the Code was officially recognised as a Code of Practice under Article 45 DSA (and becomes part of the DSA regulatory framework).³⁴ A crucial feature of this definition is the element of “intention,” which distinguishes disinformation from misinformation, which lacks the element of intention (where false content is shared without intent to harm, and is shared in good faith). As such, false information or misleading information that is spread with an intention to deceive is classed as disinformation; and this notion is the focus of this Mapping Report.³⁵

³⁰ Wikimedia Foundation, “Wikipedia DSA Audit Report 2023-24 Public.pdf”, p. 9; citing Wikimedia Foundation, Wikimedia Foundation Universal Code of Conduct, Section 1, https://foundation.wikimedia.org/wiki/Policy:Universal_Code_of_Conduct.

³¹ Wikimedia Foundation, “Wikipedia DSA Audit Report 2023-24 Public.pdf”, p. 9.

³² Wikipedia, “Wikipedia:Five pillars”, https://en.wikipedia.org/wiki/Wikipedia:Five_pillars.

³³ European Commission, Communication on the European democracy action plan, COM(2020) 790 final. See also, European Commission, Tackling online disinformation: a European Approach, COM/2018/236 final, (“verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm”).

³⁴ European Commission, “Commission endorses the integration of the voluntary Code of Practice on Disinformation into the Digital Services Act”, 13 February 2025, <https://digital-strategy.ec.europa.eu/en/news/commission-endorses-integration-voluntary-code-practice-disinformation-digital-services-act>.

³⁵ Notably, while there has been critique of this definition in particular, there is also critique of the notion of disinformation as a policy concept generally. See, for example, R. Ó Fathaigh, N. Helberger, & N. Appelmann, “The perils of legally defining disinformation” (2021) 10(4) *Internet Policy Review* 1.

It should be pointed out at this stage that Wikipedia's policies on false information actually overlap with the EU definitions. For example, the Wikimedia Foundation's Terms of Use specifically prohibit "engaging in false statements," including "posting or modifying content with the intent to deceive or mislead others".³⁶ While the Wikimedia Foundation's Universal Code of Conduct also prohibits "deliberately" introducing "false", or "inaccurate content", including "deliberately false rendering of sources and altering the correct way of composing editorial content".³⁷ Notably, English Wikipedia's rules on vandalism and "accidental misinformation" include that a user who, "in good faith," adds content to an article that is "factually inaccurate in the belief that it is accurate," is trying to "contribute to and improve Wikipedia, not vandalize it."³⁸ And where editors "believe inaccurate information has been added to an article in good faith," editors should "remove it once you are certain it is inaccurate, and consider discussing its factuality with the user who has added it".³⁹ Thus, an important preliminary finding is that Wikipedia and Wikimedia Foundation policies also place "intention" as a central notion in policies targeting false information, and very much align with the European Commission's definition of disinformation.

3. Wikipedia's policies applicable to election disinformation

Having set out the introductory issues, this section begins by outlining the main rules and principles that underpin Wikipedia's approach to dealing with election disinformation. It discusses Wikipedia's editorial model and core content policies, as well as community moderation practices and the use of technological tools. Following this, there will be an analysis of the intervention mechanisms available to the Wikimedia Foundation, and a discussion of the relevant rules under the Terms of Use and the Universal Code of Conduct. Subsequently, an examination of Wikipedia's risk assessment and audit obligations under the DSA will follow.

At the outset, it should be emphasised that the Wikimedia Foundation has various policies and initiatives on election disinformation; and general Wikipedia content rules, such as those concerning neutrality, verifiability and original research, are crucial mechanisms for addressing election disinformation in the EU, particularly within the editorial community. Additionally, provisions in Wikimedia Foundation's Terms of Use,⁴⁰ and the Universal Code of Conduct,⁴¹ are

³⁶ Wikimedia Foundation, "Wikimedia Foundation Terms of Use", section 4, https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use.

³⁷ Wikimedia Foundation, "Wikimedia Foundation Universal Code of Conduct", section 3.3, https://foundation.wikimedia.org/wiki/Policy:Universal_Code_of_Conduct.

³⁸ English Wikipedia, "Wikipedia:Vandalism", "<https://en.wikipedia.org/wiki/Wikipedia:Vandalism>."

³⁹ English Wikipedia, "Wikipedia:Vandalism", "<https://en.wikipedia.org/wiki/Wikipedia:Vandalism>."

⁴⁰ Wikimedia Foundation, "Wikimedia Foundation Terms of Use", https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use.

⁴¹ Wikimedia Foundation, "Wikimedia Foundation Universal Code of Conduct", https://foundation.wikimedia.org/wiki/Policy:Universal_Code_of_Conduct.

directly applicable to disinformation, guiding community actions and potential interventions by the Wikimedia Foundation. Alongside practical initiatives including, for example, the Wikimedia Foundation’s Disinformation Response Team, which was set up to support the wider European community in addressing threats of disinformation in the run-up to and during 2024 European Parliament elections,⁴² these frameworks provide important insights into how Wikipedia addresses election-related disinformation. Indeed, the Wikimedia Foundation’s Disinformation Response Team published a report on the European Parliament election 2024, which is also discussed below.⁴³

3.1 Wikipedia’s Editorial Model

To understand how Wikipedia’s policies are applicable to election disinformation, it is helpful to first begin with Wikipedia’s editorial model. In this regard, the content on Wikipedia comes from voluntary contributions worldwide.⁴⁴ Although certain pages can be “protected”, in principle, anyone is able to edit Wikipedia by clicking on the “Edit” tab at the top of an article.⁴⁵ “Protected pages” are used to prevent disruption, such as vandalism or repeated violations of content policies and, in such cases, can only be edited by users with certain access rights depending on the level of protection applied.⁴⁶ Protection may be temporary or permanent, and besides edit protection, there are other forms of protection, such as preventing the creation of pages or protecting the talk pages.⁴⁷ A notice at the top of the page always indicates which form of protection is in place, and the exact protection rules vary between language communities.⁴⁸ Protection is most commonly applied by community administrators, although in some cases the Wikimedia Foundation may also impose it.⁴⁹ Due to its “anyone can edit” policy, facilitating

⁴² Wikimedia Foundation, ‘Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024’ https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

⁴³ Wikimedia Foundation, “European Parliament Election 2024 - DRT Report”, https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

⁴⁴ Brennan Schaffner and others, “Community Guidelines Make This the Best Party on the Internet”: An In-Depth Study of Online Platforms’ Content Moderation Policies’, *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2024) p. 5, <https://dl.acm.org/doi/10.1145/3613904.3642333>.

⁴⁵ Amanda Bertsch and Steven Bethard, ‘Detection of Puffery on the English Wikipedia’ in Wei Xu and others (eds), *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)* (Association for Computational Linguistics 2021) p. 329, <https://aclanthology.org/2021.wnut-1.36/>.

⁴⁶ Wikipedia, ‘Protection Policy’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Protection_policy&oldid=1284331874.

⁴⁷ Wikipedia, ‘Protection Policy’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Protection_policy&oldid=1284331874.

⁴⁸ Wikipedia, ‘Protection Policy’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Protection_policy&oldid=1284331874.

⁴⁹ Wikipedia, ‘Protection Policy’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Protection_policy&oldid=1284331874.

anonymous editing, Wikipedia's relationship with identity is considered complex.⁵⁰ All users can, in principle, edit existing articles (entries), anonymously or not, and comment on them via a corresponding so-called Talk page, and they are able to create new articles, or entries upon registration with a user name.⁵¹ Volunteers dedicate a substantial amount of time to welcoming new editors and correcting erroneous edits.⁵²

Discussion processes involving the use of policies and guidelines are used to ensure the quality of the articles written.⁵³ Examples of such policies and guidelines include the “Notability” guideline, which ensures that only topics worthy of an encyclopaedic entry are included in Wikipedia.⁵⁴ Similarly, the “Neutral Point of View” policy (NPOV) outlines practices for maintaining an unbiased viewpoint, which means “representing fairly, proportionately, and, as far as possible, without editorial bias, all the significant views that have been published by reliable sources on a topic”, and can be enforced through the NPOV tag, serving as an annotation to report possible violations of policies or guidelines.⁵⁵ Other examples include the “Verifiability” policy and the “No original research” policy (NOR), which ensure that information can be verified as coming from reliable sources and that no original research is included in the article.⁵⁶

These principles move the focus of expertise away from the contributors and onto the sources themselves.⁵⁷ Discussion pages and editing summaries often use abbreviations to refer to these policies and guidelines.⁵⁸ With every edit logged, through transparency as an aspirational virtue, Wikipedians could be made to explain and, if necessary defend, their actions in detail.⁵⁹ For each language community (e.g., English Wikipedia, French Wikipedia), however, there are different

⁵⁰ J. Grimmelmann, “The Virtues of Moderation” (2015) 17 *Yale Journal of Law and Technology*, p. 84.

⁵¹ Paul de Laat, ‘Coercion or empowerment? Moderation of content in Wikipedia as ‘essentially contested’ bureaucratic rules’ (2012) 14 *Ethics and Information Technology* p. 125.

⁵² Amanda Bertsch and Steven Bethard, ‘Detection of Puffery on the English Wikipedia’ in Wei Xu and others (eds), *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)* (Association for Computational Linguistics 2021) p. 330, <https://aclanthology.org/2021.wnut-1.36/>.

⁵³ Lucie-Aimée Kaffee, Arnav Arora and Isabelle Augenstein, ‘Why Should This Article Be Deleted? Transparent Stance Detection in Multilingual Wikipedia Editor Discussions’ in Houda Bouamor, Juan Pino and Kalika Bali (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2023) p. 5891, <https://aclanthology.org/2023.emnlp-main.361/>.

⁵⁴ Wikipedia, ‘Notability’, <https://en.wikipedia.org/w/index.php?title=Wikipedia:Notability&oldid=1282136454>.

⁵⁵ Wikipedia, ‘Neutral Point of View’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Neutral_point_of_view&oldid=1289643337.

⁵⁶ Diego Saez-Trumper, ‘Online Disinformation and the Role of Wikipedia’ (arXiv, 14 October 2019) p. 5, <http://arxiv.org/abs/1910.12596>.

⁵⁷ Bunty Avieson, ‘Editors, Sources and the “go Back” Button: Wikipedia’s Framework for Beating Misinformation’ [2022] *First Monday*, <https://firstmonday.org/ojs/index.php/fm/article/view/12754>.

⁵⁸ Wikipedia, ‘Policies and Guidelines’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Policies_and_guidelines&oldid=1281892125.

⁵⁹ J. Grimmelmann, “The Virtues of Moderation” (2015) 17 *Yale Journal of Law and Technology*, p. 85.

standards for writing style, moderation, and user behaviour.⁶⁰ In general, the distinction between policies and guidelines is that the former describe standards widely accepted by editors that all users should normally follow, while the latter are sets of best practices supported by consensus.⁶¹ Most of Wikipedia's policies and guidelines pages are divided into categories for governing content, conduct, deletion, enforcement, legal issues, and procedures.⁶² Overarching these more specific policies and guidelines are five pillars, which summarise the fundamental principles of Wikipedia.⁶³

3.1.1 The Fundamental Principles of Wikipedia: Five Pillars

Although Wikipedia principles do not explicitly mention “election disinformation”, a number of them are directly applicable to it. This includes the requirement under the “Neutral point of view” pillar that all Wikipedia articles must be accurate and based on reliable sources, as will be further discussed below. The following provides an overview of the five fundamental principles guiding Wikipedia and the Wikimedia Foundation which have “informed how and why Wikipedia grows and develops.”⁶⁴

First, through the “Wikipedia is an encyclopedia” pillar, Wikipedia defines itself by what it is and by what it is not.⁶⁵ While it describes itself as a platform that combines the features of encyclopaedias, almanacs, and gazetteers, it explicitly states that it is “not a soapbox, an advertising platform, a social network, a vanity press, an experiment in anarchy or democracy, an indiscriminate collection of information, nor a web directory”.⁶⁶ Crucially, it is stated that Wikipedia is not a newspaper, meaning that the articles themselves must not contain original reporting or news reports and thus does not constitute a primary source, although topics of historical significance that are currently in the news can be updated with recently verified information.⁶⁷

⁶⁰ Leijie Wang and Haiyi Zhu, ‘How Are ML-Based Online Content Moderation Systems Actually Used? Studying Community Size, Local Activity, and Disparate Treatment’, *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2022) p. 826, <https://dl.acm.org/doi/10.1145/3531146.3533147>.

⁶¹ Wikipedia, ‘Policies and Guidelines’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Policies_and_guidelines&oldid=1281892125.

⁶² Justin Clark and others, ‘Content and Conduct: How English Wikipedia Moderates Harmful Speech’, p. 20, <http://nrs.harvard.edu/urn-3:HUL.InstRepos:41872342>.

⁶³ Wikipedia, ‘Five Pillars’, https://en.wikipedia.org/wiki/Wikipedia:Five_pillars.

⁶⁴ Wikimedia Foundation, ‘2021-2022 Annual Report: Pillars That Inspire’, <https://wikimediafoundation.org/about/annualreport/2022-annual-report/pillars/>.

⁶⁵ Wikipedia, ‘What Wikipedia Is Not’, https://en.wikipedia.org/w/index.php?title=Wikipedia:What_Wikipedia_is_not&oldid=1290476862#Wikipedia_is_not_a_soapbox_or_means_of_promotion.

⁶⁶ Wikipedia, ‘Five Pillars’, https://en.wikipedia.org/wiki/Wikipedia:Five_pillars.

⁶⁷ Wikipedia, ‘Five Pillars’, https://en.wikipedia.org/wiki/Wikipedia:Five_pillars and Wikipedia, ‘What Wikipedia Is Not’, https://en.wikipedia.org/w/index.php?title=Wikipedia:What_Wikipedia_is_not&oldid=1290476862#Wikipedia_is_not_a_soapbox_or_means_of_promotion.

Second, while the idea of a neutral point of view was previously mentioned as a content policy, it is also one of the core pillars, which reads: “Wikipedia is written from a neutral point of view”. This means that all articles should aim to be verifiably accurate by citing reliable sources and should not reflect the personal experiences, interpretations, or opinions of editors.⁶⁸ These reliable sources should come from independent, published sources with a reputation for fact-checking and accuracy, although the reliability of a source still depends on the context.⁶⁹ Some sources are deprecated, meaning they are highly questionable and editors are discouraged from citing them in articles because they almost always fail the reliable sources guideline.⁷⁰ Deprecating a source differs from blacklisting, which is a separate mechanism whereby websites are usually blacklisted if they are involved in spam-related issues, such as external link spamming.⁷¹ Sources can only be deprecated if there is community consensus.⁷² Examples of such sources on English Wikipedia include Russia Today, due to it publishing “false or fabricated information,” and engaging in “propaganda and disinformation;” Sputnik, being an unreliable source, due to “false or fabricated information;” Press TV, being an Iranian “government propaganda outlet that publishes disinformation;” Infowars, Newsmax, Breitbart News, and China Global Television Network, for publishing disinformation.⁷³ The Wikimedia Foundation considers the neutral point of view principle fundamental to safeguarding Wikipedia against misinformation and disinformation, given its demand for verifiable accuracy and impartiality.⁷⁴

Third, the “Wikipedia is free content that anyone can use, edit and distribute” pillar entails that all contributions are openly licensed and can be modified or reused by others, and must comply with copyright rules.⁷⁵ Fourth, the “Wikipedia’s editors should treat each other with respect and civility” pillar emphasises collaborative editing.⁷⁶ It encourages editors to assume good faith, meaning they should not assume that other editors are deliberately trying to harm Wikipedia, even when their actions are harmful, and they should likewise act in good faith.⁷⁷ It also encourages editors to avoid personal attacks or edit wars and to follow dispute resolution

⁶⁸ Wikipedia, ‘Five Pillars’, https://en.wikipedia.org/wiki/Wikipedia:Five_pillars.

⁶⁹ Wikipedia, ‘Reliable Sources’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Reliable_sources&oldid=1290893500.

⁷⁰ Wikipedia, ‘Deprecated Sources’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Deprecated_sources&oldid=1276030532.

⁷¹ Wikipedia, ‘Deprecated Sources’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Deprecated_sources&oldid=1276030532.

⁷² Wikipedia, ‘Deprecated Sources’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Deprecated_sources&oldid=1276030532.

⁷³ Wikipedia, ‘Deprecated Sources’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Deprecated_sources&oldid=1276030532.

⁷⁴ Wikimedia Foundation, ‘2021-2022 Annual Report, Pillar 2: How We Are Fighting Misinformation and Disinformation’, <https://wikimediafoundation.org/about/annualreport/2022-annual-report/pillar2/>.

⁷⁵ Wikipedia, ‘Five Pillars’, https://en.wikipedia.org/wiki/Wikipedia:Five_pillars.

⁷⁶ Wikipedia, ‘Five Pillars’, https://en.wikipedia.org/wiki/Wikipedia:Five_pillars.

⁷⁷ Wikipedia, ‘Five Pillars’, https://en.wikipedia.org/wiki/Wikipedia:Five_pillars.

procedures.⁷⁸ Finally, the “Wikipedia has no firm rules” pillar emphasises that policies and guidelines are adaptable and can change over time.⁷⁹ The important thing is the underlying principles and intentions, rather than strict adherence to the wording.⁸⁰

3.2 Community-based Moderation Practices

Having set out the Wikipedia pillars, it is now important to consider how moderation operates at community level, where collaborative processes are used to make most editorial decisions. This is particularly relevant to how the community tackles disinformation. If an article does not meet the community's standards, editors may recommend its removal.⁸¹ While Wikipedia has a central host, through the Wikimedia Foundation, moderation is assumed to be decentralised.⁸² Accordingly, everyday moderation is primarily handled by users, as seen on other community-governed platforms such as Twitch and Reddit.⁸³ A deletion proposal opens a deletion discussion, where comments on the discussion can be made, the deletion can be supported or rejected, or proposals for merging with other articles can be made.⁸⁴ The volunteers who enforce these decisions have extended privileges that allow them to decide on such content, for instance, known as administrators or functionaries.⁸⁵ When deciding the outcome of a discussion, administrators may also keep discussions open for a longer period of time, or close them with a verdict of “No consensus,” which would mean that the article would not be deleted, as that would require an outcome that is unambiguous.⁸⁶ To measure such consensus, non-binding votes are typically utilised, and in cases where this fails, a mediation committee and arbitration committee exist to address disputes.⁸⁷ In discussing the removal of articles, editors reference policies in

⁷⁸ Wikipedia, ‘Five Pillars’, https://en.wikipedia.org/wiki/Wikipedia:Five_pillars.

⁷⁹ Wikipedia, ‘Five Pillars’, https://en.wikipedia.org/wiki/Wikipedia:Five_pillars.

⁸⁰ Wikipedia, ‘Five Pillars’, https://en.wikipedia.org/wiki/Wikipedia:Five_pillars.

⁸¹ Lucie-Aimée Kaffee, Arnav Arora and Isabelle Augenstein, ‘Why Should This Article Be Deleted? Transparent Stance Detection in Multilingual Wikipedia Editor Discussions’ in Houda Bouamor, Juan Pino and Kalika Bali (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2023) p. 5892, <https://aclanthology.org/2023.emnlp-main.361/>.

⁸² J. Grimmelmann, “The Virtues of Moderation” (2015) 17 *Yale Journal of Law and Technology*, p. 74.

⁸³ Joseph Seering, ‘Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation: Proceedings of the ACM on Human-Computer Interaction: Vol 4, No CSCW2’, p. 8, <https://dl.acm.org/doi/abs/10.1145/3415178>.

⁸⁴ Lucie-Aimée Kaffee, Arnav Arora and Isabelle Augenstein, ‘Why Should This Article Be Deleted? Transparent Stance Detection in Multilingual Wikipedia Editor Discussions’ in Houda Bouamor, Juan Pino and Kalika Bali (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2023) p. 5892, <https://aclanthology.org/2023.emnlp-main.361/>.

⁸⁵ Ryan Merkley, ‘How Wikipedia Is Preparing For The 2020 U.S. Election’ (*Down the Rabbit Hole*, 16 December 2020), <https://medium.com/freely-sharing-the-sum-of-all-knowledge/how-wikipedia-is-preparing-for-the-2020-u-s-election-d2be81ba4bc1>.

⁸⁶ Elijah Mayfield and Alan W Black, ‘Analyzing Wikipedia Deletion Debates with a Group Decision-Making Forecast Model’ (2019), *Proceedings of the ACM on Human-Computer Interaction* 1, p. 4.

⁸⁷ J. Grimmelmann, “The Virtues of Moderation” (2015) 17 *Yale Journal of Law and Technology*, p. 86.

order to support their arguments.⁸⁸ Wikipedia relies on a comprehensive set of community rules that have been translated into practical procedures for content moderation, which can vary per language version.⁸⁹ These policies, established by the community and covering either a single Project edition or multiple Project editions, can be modified by each community according to its own procedures.⁹⁰ In addition, administrators are the ones who can protect pages, block other editors, and also undo these actions.⁹¹

Wikipedia's norms and policies are enforced more strictly when editing “contentious topics,” and administrators hold additional authority to minimise disruption and are authorised to implement protections at any level.⁹² A contentious topic is one that has attracted frequent and persistent disruptive editing and has been designated as such by the Arbitration Committee.⁹³ All pages broadly related to a contentious topic, as well as parts of other pages related to the topic, are subject to this procedure. A particularly relevant example of a contentious topic includes the results of any national or sub-national election.⁹⁴ In such cases, different levels of page protection can be implemented. For example, semi-protection prevents unregistered users (users with only their IP addresses visible) and accounts that are not confirmed or autoconfirmed (accounts that are at least four days old and have made at least 10 edits on English Wikipedia) from editing.⁹⁵ Alternatively, for example an extended confirmed protection level can be implemented, which only allows edits by users who have been active for at least 30 days and have made at least 500 edits, as well as by administrators and bots.⁹⁶ Examples of semi-protected election-related pages in the EU include those of the Dutch politician Geert Wilders, Polish president Andrzej Duda, and

⁸⁸ Lucie-Aimée Kaffee, Arnav Arora and Isabelle Augenstein, ‘Why Should This Article Be Deleted? Transparent Stance Detection in Multilingual Wikipedia Editor Discussions’ in Houda Bouamor, Juan Pino and Kalika Bali (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2023) p. 5892, <https://aclanthology.org/2023.emnlp-main.361/>.

⁸⁹ Justin Clark and others, ‘Content and Conduct: How English Wikipedia Moderates Harmful Speech’, p. 19, <http://nrs.harvard.edu/urn-3:HUL.InstRepos:41872342>.

⁹⁰ Wikimedia Foundation, ‘Terms of Use’, https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use.

⁹¹ Brian Butler, Elisabeth Joyce and Jacqueline Pike, ‘Don’t Look Now, but We’ve Created a Bureaucracy: The Nature and Roles of Policies and Rules in Wikipedia’, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2008) p. 1107, <https://dl.acm.org/doi/10.1145/1357054.1357227>.

⁹² Wikipedia, ‘Protection Policy’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Protection_policy&oldid=1284331874.

⁹³ Wikipedia, ‘Contentious topics’, https://en.wikipedia.org/wiki/Wikipedia:Contentious_topics.

⁹⁴ Wikipedia, ‘Contentious topics/Historical elections’, https://en.wikipedia.org/wiki/Wikipedia:Contentious_topics/Historical_elections.

⁹⁵ Wikipedia, ‘Protection Policy’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Protection_policy&oldid=1284331874.

⁹⁶ Wikipedia, ‘Protection Policy’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Protection_policy&oldid=1284331874.

the German political party Alternative for Germany (Alternative für Deutschland).⁹⁷ The page of the Romanian politician Călin Georgescu also falls under extended confirmed protection.⁹⁸

3.2.1 Technological Support

The following section provides an overview of how technological tools and automated features can be used to detect and address issues such as election disinformation. In the context of moderation involving both manual and automated processes, scholars such as Clark and others describe automation as the first layer of moderation, with manual review serving as “the second line of defense”.⁹⁹ As Wikipedia has grown, technologies have been implemented to help deal with issues such as vandalism and incorrect spelling.¹⁰⁰ Since 2017, there has been a team dedicated to machine learning at the Wikimedia Foundation.¹⁰¹ These tools can be implemented top-down from the server-side code, but can also be developed via and by the community.¹⁰² In some cases, for example vandalism, edits can be automatically reverted without human intervention, while in many other cases these tools would complement human editors by flagging edits.¹⁰³ For example, of the thousands of bots developed by the community, ClueBot NG is the most prominent in terms of content removal.¹⁰⁴ Another example is ORES, which, unlike ClueBot NG, does not delete content but generates edit scores to help moderators prioritise revisions.¹⁰⁵ With regard to combatting disinformation specifically, the Wikimedia Foundation has reported that its Research team has, in collaboration with the Machine Learning team, been developing a new spambot detection system to support this effort by helping to maintain Wikipedia’s standards

⁹⁷ See, https://en.wikipedia.org/wiki/Geert_Wilders, https://en.wikipedia.org/wiki/Andrzej_Duda and https://en.wikipedia.org/wiki/Alternative_for_Germany.

⁹⁸ See, https://en.wikipedia.org/wiki/C%C4%83lin_Georgescu.

⁹⁹ Justin Clark, Robert Faris, Urs Gasser, Adam Holland, Hilary Ross, and Casey Tilton, ‘Content and Conduct: How English Wikipedia Moderates Harmful Speech’, p. 23, <http://nrs.harvard.edu/urn-3:HUL.InstRepos:41872342>.

¹⁰⁰ Amanda Bertsch and Steven Bethard, ‘Detection of Puffery on the English Wikipedia’ in Wei Xu and others (eds), *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)* (Association for Computational Linguistics 2021) p. 329, <https://aclanthology.org/2021.wnut-1.36/>.

¹⁰¹ Mina Haq, ‘The Wikimedia Foundation’s Crucial Spot on the Frontlines of the Disinformation War’ (*PEN America*, 23 October 2024), <https://pen.org/the-wikimedia-foundations-crucial-spot-on-the-frontlines-of-the-disinformation-war/>.

¹⁰² Aaron Halfaker and R Stuart Geiger, ‘ORES: Lowering Barriers with Participatory Machine Learning In’, *Proceedings of the ACM on Human-Computer Interaction* (2020) p. 6 <<https://dl.acm.org/doi/abs/10.1145/3415219>.

¹⁰³ Amanda Bertsch and Steven Bethard, ‘Detection of Puffery on the English Wikipedia’ in Wei Xu and others (eds), *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)* (Association for Computational Linguistics 2021) p. 329, <https://aclanthology.org/2021.wnut-1.36/>.

¹⁰⁴ Justin Clark and others, ‘Content and Conduct: How English Wikipedia Moderates Harmful Speech’, p. 22, <http://nrs.harvard.edu/urn-3:HUL.InstRepos:41872342>.

¹⁰⁵ Justin Clark and others, ‘Content and Conduct: How English Wikipedia Moderates Harmful Speech’, p. 22, <http://nrs.harvard.edu/urn-3:HUL.InstRepos:41872342>.

of accuracy and neutrality.¹⁰⁶ This project focuses on building models to help editors detect and prevent inaccurate or misleading revisions.¹⁰⁷

3.3 Interventions by the Wikimedia Foundation

In addition to community-led efforts, the role, or more importantly the potential role, of the Wikimedia Foundation itself is important to examine, particularly in cases where it may be necessary to intervene to address issues of election disinformation, which the community may not be able to address independently. As discussed above, it is generally understood that the staff of the non-profit Wikimedia Foundation are not functionally tasked with making content editing decisions.¹⁰⁸ The Wikimedia Foundation confirms this, by stating: “The Foundation does not hold editorial or supervisory control over content and conduct in the Wikimedia projects; this work is done by a largely autonomous community of volunteers who, in accordance with our Terms of Use, create their own policies meant to uphold the educational goals of our movement”.¹⁰⁹ While generally attempting to avoid interference, they nevertheless have powers to intervene in specific cases and such interventions are generally known as “office actions”.¹¹⁰ If either community action has proven ineffective, or legal considerations demand intervention by the Wikimedia Foundation, they could take appropriate action.¹¹¹ The Wikimedia Foundation Office Actions Policy outlines the specifics of these measures required for intervention going beyond the scope of community governance mechanisms.¹¹² The Wikimedia Foundation policy starts with reaffirming that: “Local policies remain primary on all Wikimedia projects, as explained in the Terms of Use, and office actions are complementary to those local policies”.¹¹³

3.3.1 Office Actions

The purpose of the Office Actions policy is to “improve the actual and perceived safety of Wikimedia community members, the movement itself, and the public in circumstances where actions on local community governance level are either insufficient or not possible”.¹¹⁴ It distinguishes between primary office actions, secondary office actions, and the use of advanced

¹⁰⁶ Wikimedia Foundation, ‘2021-2022 Annual Report, Pillar 2: How We Are Fighting Misinformation and Disinformation’, <https://wikimediafoundation.org/about/annualreport/2022-annual-report/pillar2/>.

¹⁰⁷ Wikimedia Foundation, ‘2021-2022 Annual Report, Pillar 2: How We Are Fighting Misinformation and Disinformation’, <https://wikimediafoundation.org/about/annualreport/2022-annual-report/pillar2/>.

¹⁰⁸ Aaron Halfaker and R Stuart Geiger, ‘ORES: Lowering Barriers with Participatory Machine Learning In’, *Proceedings of the ACM on Human-Computer Interaction* (2020) p. 2, <https://dl.acm.org/doi/abs/10.1145/3415219>.

¹⁰⁹ Wikimedia Foundation, ‘Office Actions Policy’ https://foundation.wikimedia.org/wiki/Policy:Wikimedia_Foundation_Office_Actions_Policy.

¹¹⁰ Wikimedia Foundation, ‘Office Actions Policy’.

¹¹¹ Wikimedia Foundation, ‘Office Actions Policy’.

¹¹² Wikimedia Foundation, ‘Office Actions Policy’.

¹¹³ Wikimedia Foundation, ‘Office Actions Policy’.

¹¹⁴ Wikimedia Foundation, ‘Office Actions Policy’.

rights by Foundation staff.¹¹⁵ While some of the specific actions discussed may be imposed by affiliates or the community, it is important to note that the Office Actions Policy refers only to actions taken by the Wikimedia Foundation.¹¹⁶

3.3.1.1 Primary Office Actions

Primary office actions include so-called Foundation global bans, Foundation event bans, copyright enforcement, and protection of minors. A **Foundation global ban** is described as “one of the most severe actions the Wikimedia Foundation may take”, targeting users who engage in harassment, threats, legal violations, other behaviour that jeopardises the trust or safety of users or employees, or the repeated violation of the Wikimedia Foundation Terms of Use.¹¹⁷ When a global ban is in place it “prohibits individuals, either in their own capacity or as agents of others, from all Wikimedia Foundation websites, platforms and activities”.¹¹⁸ A **Foundation event ban** deals with the actual physical or digital restriction of participation in Foundation-sponsored or related events. In terms of **copyright infringement** the policy first encourages reporting of copyright infringements through community procedures, but also allows for formal copyright infringement requests to be made to the Wikimedia Foundation and discourages the reversal of removals without valid legal grounds.¹¹⁹ Lastly, when minors are at risk and it comes to the **protection of minors**, the Wikimedia Foundation takes immediate action aimed at both stopping behaviour and removing content that promotes child abuse.¹²⁰

3.3.1.2 Secondary Office Actions

Secondary office actions are usually carried out at the Wikimedia Foundation’s discretion. Requests for these actions made directly to the Wikimedia Foundation will generally be deferred to the relevant community governance mechanisms. According to the policy, the Wikimedia Foundation has only undertaken these actions in the past in “extraordinary circumstances”.¹²¹ Actions that fall under this category include **conduct warnings** and the **removal of advanced rights**. The former entails issuing a conduct warning in situations where “a contributor’s online or/and offline behavior is considered borderline abusive, disruptive or otherwise hinders the collaborative process, but does not yet rise to the level of corrective actions”.¹²² The latter entails that, in “extremely rare situations”, the Wikimedia Foundation may revoke advanced user rights

¹¹⁵ Wikimedia Foundation, ‘Office Actions Policy’.

¹¹⁶ Wikimedia Foundation, ‘Office Actions Policy’.

¹¹⁷ Wikimedia Foundation, ‘Office Actions Policy’.

¹¹⁸ Wikimedia Foundation, ‘Office Actions Policy’.

¹¹⁹ Wikimedia Foundation, ‘Office Actions Policy’.

¹²⁰ Wikimedia Foundation, ‘Office Actions Policy’.

¹²¹ Wikimedia Foundation, ‘Office Actions Policy’.

¹²² Wikimedia Foundation, ‘Office Actions Policy’.

without disclosing all details to the community, in cases where there has been a serious breach of trust by Wikimedia contributors or other users with access to advanced tools that cannot be handled through community mechanisms.¹²³

3.3.1.3 Use of Advanced Rights by Foundation Staff

The use of advanced rights by Wikimedia Foundation staff includes administrative measures taken in response to reports of misconduct, in line with the Terms of Use. Again, the Wikimedia Foundation emphasises that it “strives to take the smallest amount of action possible” and generally does not act on community requests for such measures.¹²⁴ These actions are usually carried out by the Trust & Safety team and involve the use of tools such as **CheckUser**, primarily in urgent situations involving threats, legal disputes, or the enforcement of global bans.¹²⁵ In exceptional circumstances, the Wikimedia Foundation may also apply **page protection** when legal orders or significant safety concerns necessitate intervention that goes beyond the scope of standard community governance.¹²⁶ **Range blocks**, which are usually employed by the community to reduce vandalism, may be used to enforce global bans and prevent groups of IP addresses from accessing Wikimedia projects.¹²⁷

3.3.1.4 Requests and Appeals

Requests to consider office actions can be directed to the relevant Wikimedia Foundation team, such as the Trust & Safety or Legal Affairs teams, as office actions are only conducted by Foundation employees or contractors.¹²⁸ Those with the technical ability to reverse or modify office actions, such as Wikimedia administrators, are urged not to do so, as unauthorised changes will be undone and may result in sanctions, including possible revocation of user privileges.¹²⁹ Office actions, as well as decisions not to take action, may be appealed.¹³⁰

3.4 Wikimedia Foundation: Universal Code of Conduct and Terms of Use

As can be seen from the previous section, office action policies primarily refer to the Wikimedia Foundation Terms of Use (Terms of Use). For instance, it explicitly states that “all office actions

¹²³ Wikimedia Foundation, ‘Office Actions Policy’.

¹²⁴ Wikimedia Foundation, ‘Office Actions Policy’.

¹²⁵ Wikimedia Foundation, ‘Office Actions Policy’.

¹²⁶ Wikimedia Foundation, ‘Office Actions Policy’.

¹²⁷ Wikimedia Foundation, ‘Office Actions Policy’.

¹²⁸ Wikimedia Foundation, ‘Office Actions Policy’.

¹²⁹ Wikimedia Foundation, ‘Office Actions Policy’.

¹³⁰ Wikimedia Foundation, ‘Office Actions Policy’.

are performed pursuant to the Wikimedia Foundation Terms of Use”.¹³¹ As such, it provides the formal basis for Foundation interventions and also explicitly references the Wikimedia Foundation Universal Code of Conduct (Code of Conduct), which establishes a baseline of behaviour applicable across all Wikimedia projects and enforceable by local functionaries and the Wikimedia Foundation itself. The next section outlines specific relevant provisions from the Code of Conduct and the Terms of Use in the context of election disinformation.

3.4.1 Universal Code of Conduct

The Code of Conduct was formally adopted by the Wikimedia Foundation’s Board of Trustees in December 2020 with the aim of establishing a binding set of minimum behavioural standards across all Wikimedia projects and spaces.¹³² In total, more than 1,500 volunteers from 19 different Wikipedia projects, representing five continents and 30 languages, participated in creating it together with the Wikimedia Foundation.¹³³ Furthermore, the Code of Conduct Enforcement Guidelines state that the Wikimedia Foundation is committed to developing and maintaining a centralised reporting and processing tool to handle Code of Conduct violations.¹³⁴ The Code of Conduct establishes a universal behavioural baseline for all Wikimedia projects, enabling communities to develop additional policies tailored to their specific contexts, with violations potentially resulting in sanctions imposed by either local functionaries or the Wikimedia Foundation itself.¹³⁵

3.4.1.1 Relevant Provisions of the Universal Code of Conduct

The Code of Conduct explicitly identifies several forms of unacceptable behaviour, including harassment, the abuse of power, and the abuse of privilege and influence.¹³⁶ A notable aspect is the focus on content-related violations, including “deliberately introducing biased, false, inaccurate or inappropriate content, or hindering, impeding or otherwise hampering the creation (and/or maintenance) of content”.¹³⁷ This includes arbitrarily and unexplainably repeatedly removing content and “systematically manipulating content to favour specific interpretations of facts or points of view”, including by means of “unfaithfully or deliberately falsely rendering

¹³¹ Wikipedia, ‘Office Actions’, https://en.wikipedia.org/w/index.php?title=Wikipedia:Office_actions&oldid=1285434804.

¹³² Wikimedia Foundation, ‘Resolution: Approval of a Universal Code of Conduct’, https://foundation.wikimedia.org/wiki/Resolution:Approval_of_a_Universal_Code_of_Conduct.

¹³³ Wikimedia Foundation, ‘Wikipedia Embraces First-of-Its Kind Universal Code of Conduct, Conceived For The New Internet Era’ (Wikimedia Foundation, 2 February 2021), <https://wikimediafoundation.org/news/2021/02/02/wikipedia-embraces-first-of-its-kind-universal-code-of-conduct/>.

¹³⁴ Wikimedia Foundation, ‘Policy:Universal Code of Conduct/Enforcement Guidelines’ https://foundation.wikimedia.org/wiki/Policy:Universal_Code_of_Conduct/Enforcement_guidelines.

¹³⁵ Wikimedia Foundation, ‘Wikimedia Foundation Universal Code of Conduct’ https://foundation.wikimedia.org/wiki/Policy:Universal_Code_of_Conduct.

¹³⁶ Wikimedia Foundation, ‘Wikimedia Foundation Universal Code of Conduct’, Section 3.

¹³⁷ Wikimedia Foundation, ‘Wikimedia Foundation Universal Code of Conduct’, Section 3.3.

sources and altering the correct way of composing editorial content".¹³⁸ Finally, also mentioned under this, are examples such as hate speech or discriminatory language targeting individuals or groups based on their identity or beliefs and the use of symbols, tags or other content that intimidates or harms others, particularly when used to marginalise or ostracise rather than for encyclopaedic or informational purposes.¹³⁹

3.4.2 Wikimedia Foundation Terms of Use

The Terms of Use commence with an acknowledgement that, in general, the Wikimedia Foundation does not contribute to, monitor or delete content, except in rare cases involving legal compliance or imminent threats of serious harm.¹⁴⁰ The Terms of Use state that, when users report problems, the Wikimedia Foundation will first consider whether the issue can be addressed by existing community-led mechanisms.¹⁴¹ This is substantiated by a reminder that when joining as a contributor, editor or author, the policies governing each independent project edition, including the Code of Conduct, must be adhered to.¹⁴² Subsequently, it outlines the specific services provided, namely the project websites hosted by the Wikimedia Foundation, the technological infrastructure they maintain, and any technical spaces hosted for the maintenance and improvement of the projects.¹⁴³

3.4.2.1 Relevant Provisions of the Wikimedia Foundation, 'Terms of Use'

The Terms of Use explicitly prohibit a range of harmful activities that violate platform rules, notwithstanding their legality under applicable law.¹⁴⁴ The most notable aspect is that users are not permitted to engage in false statements, more specifically specified as "posting or modifying content with the intent to deceive or mislead others".¹⁴⁵ Other prohibited actions include disseminating defamatory material under U.S. law, impersonating individuals, concealing affiliations where disclosure is required by the terms, or using the name or username of another person with the intent to deceive, and engaging in fraudulent behaviour.¹⁴⁶

¹³⁸ Wikimedia Foundation, 'Wikimedia Foundation Universal Code of Conduct', Section 3.3.

¹³⁹ Wikimedia Foundation, 'Wikimedia Foundation Universal Code of Conduct', Section 3.3.

¹⁴⁰ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use',
https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use.

¹⁴¹ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 10.

¹⁴² Wikimedia Foundation, 'Wikimedia Foundation Terms of Use'.

¹⁴³ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 1.

¹⁴⁴ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 4.

¹⁴⁵ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 4.

¹⁴⁶ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 4.

In light of concerns surrounding election disinformation, several additional provisions of the Terms of Use take on particular relevance. These include the prohibition of disruptive and illegal misuse of Wikimedia facilities.¹⁴⁷ More specifically, users are not permitted to carry out automated activity on project websites that is abusive or disruptive, or that has not been approved by the Wikimedia community.¹⁴⁸ This includes flooding the platform with content or traffic without intending to contribute constructively, as well as unauthorised access to non-public areas of Wikimedia's systems.¹⁴⁹ Furthermore, the Terms of Use require full disclosure of paid contributions: any contributor who receives, or expects to receive, compensation for their contributions must clearly state the organisation or individual for whom or on whose behalf work is being done.¹⁵⁰ Such disclosure must be made through a statement on the user page or the relevant talk page accompanying any paid contributions, or in the edit summary.¹⁵¹

A noticeable feature of the enforcement framework is that the Wikimedia Foundation explicitly reserves the right to take action beyond the measures listed in its Office Action Policy. As stated in the Terms of Use, "enforcement of these terms may include actions not listed in the Wikimedia Foundation Office Action Policy", with the commitment to updating that policy within one year to include any new enforcement measures.¹⁵²

The Terms of Use state that the Wikimedia Foundation reserves the right to intervene directly in exceptional cases involving "an especially problematic user or especially problematic content because of significant Project disturbance or dangerous behavior", either at their own discretion or when legally required to do so.¹⁵³ In such cases, the Foundation reserves the right to investigate user behavior, detect and prevent fraud, false or unverifiable information, security or technical issues in order to respond to user support requests, or restrict access to any user's contributions that violate the Terms of Use.¹⁵⁴ The Foundation may also "refuse, revert, disable, or restrict access to contributions of any user who violates these Terms of Use" and "ban a user from editing or contributing or block a user's account or access for actions violating these Terms of Use, including repeated posting of unlawful material under applicable law in line with human rights principles".¹⁵⁵ They may also take legal action against users who violate the Terms of Use. As outlined further by the Terms, these moderation activities may be performed or informed by automated tools, like, for example, traffic flood or Denial of Service protection, with human

¹⁴⁷ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 4.

¹⁴⁸ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 4.

¹⁴⁹ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 4.

¹⁵⁰ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 4.

¹⁵¹ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 4.

¹⁵² Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 4.

¹⁵³ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 10.

¹⁵⁴ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 10.

¹⁵⁵ Wikimedia Foundation, 'Wikimedia Foundation Terms of Use', Section 10.

review generally available upon request.¹⁵⁶ However, the Terms of Use clarify that “the Foundation itself will not ban a user from editing or contributing or block a user’s account or access solely because of good faith criticism that does not result in actions otherwise violating these Terms of Use or community policies”.¹⁵⁷ Finally, users who believe that moderation actions have been taken improperly against them may be entitled to submit an appeal.¹⁵⁸

3.5 Wikipedia Election Disinformation: Practical Observations

Academic studies confirm that the number of Wikipedia edits increases sharply during politically charged moments, especially on pages of parties and candidates, where both the number of edits and the reversal likelihood increase around important political events.¹⁵⁹ In such cases, the main concern would be the way in which voter suppression tactics spread, for example by affecting information about polling station locations, or by undermining confidence in the facts through other topics, rather than through vandalism or biased campaign editing, which is usually noticed and reversed quickly.¹⁶⁰ This is substantiated by the observation that biased language and unsubstantiated claims have been identified as the most common forms of disinformation, followed by misleading information, containing edits with subtle biases that are not explicitly incorrect.¹⁶¹

The Wikimedia Foundation states that Wikipedia’s editorial standards discussed earlier would have played a major role in keeping the platform free of disinformation during the US 2020 elections.¹⁶² Moreover, with regard to the 2020 US presidential election and the Covid-19 pandemic in particular, the Wikimedia Foundation has taken significant steps to protect Wikipedia from election-related disinformation.¹⁶³ For example, a disinformation task force has

¹⁵⁶ Wikimedia Foundation, ‘Wikimedia Foundation Terms of Use’, Section 10.

¹⁵⁷ Wikimedia Foundation, ‘Wikimedia Foundation Terms of Use’, Section 10.

¹⁵⁸ Wikimedia Foundation, ‘Wikimedia Foundation Terms of Use’, Section 10.

¹⁵⁹ Giuliano Formisano and others, ‘Counter-Misinformation Dynamics: The Case of Wikipedia Editing Communities during the 2024 US Presidential Elections’ (Social Science Research Network, 17 October 2024) p. 17, <https://papers.ssrn.com/abstract=4990973>.

¹⁶⁰ Ryan Merkley, ‘What Wikipedia Saw during Election Week in the U.S., and What We’re Doing next’ (Wikimedia Foundation, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

¹⁶¹ Giuliano Formisano and others, ‘Counter-Misinformation Dynamics: The Case of Wikipedia Editing Communities during the 2024 US Presidential Elections’ (Social Science Research Network, 17 October 2024) p. 18, <https://papers.ssrn.com/abstract=4990973>.

¹⁶² Ryan Merkley, ‘What Wikipedia Saw during Election Week in the U.S., and What We’re Doing next’ (Wikimedia Foundation, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

¹⁶³ Ryan Merkley, ‘What Wikipedia Saw during Election Week in the U.S., and What We’re Doing next’ (Wikimedia Foundation, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

been established to protect the integrity of the elections.¹⁶⁴ The approach during the 2020 US elections provides an indication of the process: around 2,000 election-related pages were protected, meaning they could only be edited by users with certain access rights, depending on the level of protection applied, including the main page on the 2020 US presidential election.¹⁶⁵ In addition, these pages are said to be monitored in real time by over 56,000 volunteer editors.¹⁶⁶ Furthermore, the Wikimedia Foundation's disinformation task force recorded and evaluated 18 relevant events, and in addition, the community reversed an estimated 800 edits on election-related pages between 3 and 7 November 2020.¹⁶⁷ Notably, neither the task force members nor the Wikipedia administrators detected any large-scale state-sponsored disinformation during that period.¹⁶⁸

In addition to establishing the Internal Disinformation Task Force, efforts have been made to invest in research and development. More specifically, the Wikimedia Foundation's research team has been collaborating with international universities to enhance the community's ability to detect disinformation and vandalism by focusing on various technical tools mentioned earlier as well, such as ORES, source verification algorithms, and monitoring tools.¹⁶⁹ At the community level, specific guidelines were established and debated during that period, including the requirement of a minimum of three reliable sources to determine the winner in each state, a waiting period of at least 12 hours after the polls closed before publishing results, and an absolute prohibition of original research sources.¹⁷⁰

A more recent report by the Wikimedia Foundation mentions the establishment of a Disinformation Response Team, similar to the internal task force discussed earlier, providing

¹⁶⁴ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

¹⁶⁵ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

¹⁶⁶ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

¹⁶⁷ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

¹⁶⁸ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

¹⁶⁹ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

¹⁷⁰ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

support to the European community before the 2024 European Parliament elections.¹⁷¹ As reported, no large-scale disinformation was identified during the 2024 European Parliament elections.¹⁷² Communities expressed concern mainly about vandalism and promotional editing, which through experience they knew how to deal with independently, without interference from the Wikimedia Foundation.¹⁷³ In addition, the use of AI for edits was mentioned, including spammers using AI to generate articles, although no AI use for disinformation was identified.¹⁷⁴ The absence of major incidents put forward is the fact that misinformation manifested itself mainly on social media, sources that Wikipedia usually considers unreliable.¹⁷⁵ These findings are consistent with previous election rounds, such as in India and Poland.¹⁷⁶

Similar recommendations on approaches emerge from academic literature, including enhanced monitoring during critical political events and the deployment of targeted fact-checking for polarising figures, as well as a focus on strengthened source verification amid AI-generated misinformation.¹⁷⁷ Furthermore, it has been found that higher-risk edits are often associated with vague or absent edit summaries, which make them more difficult to address and hinder transparency in the editing process.¹⁷⁸

3.5 Conclusion

Although disinformation is not explicitly mentioned, the content rules apply directly to it. The foundational content policies, such as Neutral Viewpoint, Verifiability, and No Original Research, govern editorial activities and support the aim of providing accurate, reliable, and unbiased

¹⁷¹ Wikimedia Foundation, 'Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024' https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

¹⁷² Wikimedia Foundation, 'Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024' https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

¹⁷³ Wikimedia Foundation, 'Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024' https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

¹⁷⁴ Wikimedia Foundation, 'Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024' https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

¹⁷⁵ Wikimedia Foundation, 'Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024' https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

¹⁷⁶ Wikimedia Foundation, 'Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024' https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

¹⁷⁷ Giuliano Formisano and others, 'Counter-Misinformation Dynamics: The Case of Wikipedia Editing Communities during the 2024 US Presidential Elections' (Social Science Research Network, 17 October 2024) p. 19, 20, 21, <https://papers.ssrn.com/abstract=4990973>.

¹⁷⁸ Giuliano Formisano and others, 'Counter-Misinformation Dynamics: The Case of Wikipedia Editing Communities during the 2024 US Presidential Elections' (Social Science Research Network, 17 October 2024) p. 18, <https://papers.ssrn.com/abstract=4990973>.

information on Wikipedia. These principles are mostly enforced through moderation practices guided by the community, involving transparent editing processes, discussion pages and deletion procedures. Such moderation is supported by automated tools, which are thought of as functioning as a first layer of moderation in addition to manual review. These include bots, which can reverse obvious vandalism (including false content) without human intervention, and systems, which generate edit scores to assist in prioritising for manual review.

In addition to community-led and technical moderation efforts, Wikipedia's approach to election disinformation is supported by behavioural standards and rules established by the Wikimedia Foundation, most notably the Universal Code of Conduct and the Terms of Use. The Terms of Use and the Code of Conduct provide the formal basis for a range of intervention options by the Wikimedia Foundation, also known as office actions, which the Wikimedia Foundation could deploy if for example community action has proven ineffective, or if legal considerations demand intervention by the Wikimedia Foundation, such as for example page protection and range blocks. Most of these actions can also be carried out by the community.

Regarding unacceptable behaviour, the Universal Code of Conduct identifies a range of unacceptable behaviour relevant for election disinformation. This includes deliberately introducing “biased, false, inaccurate or inappropriate content”, and the hindrance, impedance, or other obstruction of content creation as well as its maintenance. Including arbitrarily and unexplainably repeatedly removing content and systematically manipulating content to favour specific interpretations of facts or points of view, including by means of unfaithfully or deliberately falsely rendering sources and altering the correct way of composing editorial content. Similarly, the Terms of Use prohibit engaging in false statements, specified as “posting or modifying content with the intent to deceive or mislead others”. Other prohibited actions include impersonating individuals, concealing affiliations where disclosure is required by the terms, or using the name or username of another person with the intent to deceive, and engaging in fraudulent behaviour.

In addition, as promotional editing was identified as a concern during the 2024 European Parliament elections, the provisions in the Terms of Use on paid contributions are also relevant in this context of election disinformation. These provisions require contributors who receive, or expect to receive, compensation for their edits to disclose the organisation or individual on whose behalf they are contributing. This disclosure may be made on the user page, the relevant discussion page or in the edit summary. These measures are intended to promote transparency and enable the community to understand the context in which edits are made, particularly in politically sensitive areas. In addition, promotional editing is incompatible with Wikipedia's own definition as “not an advertising platform”, as set out in its Five Pillars.

To conclude, this section has demonstrated that Wikipedia relies on a system of multifaceted community moderation, content rules, and guidelines and intervention possibilities at the level of the Wikimedia Foundation in order to address election-related disinformation.

4. Wikipedia's risk assessment and audit under the DSA

Having examined Wikipedia and the Wikimedia Foundation's policies applicable to election disinformation in the EU, it is now helpful to turn to the EU's DSA legislation.¹⁷⁹ This is because the DSA is specifically designed to address the "societal risks" that the "dissemination of disinformation" may generate,¹⁸⁰ and as discussed in the first Mapping Report, Wikipedia has been designated as a so-called Very Large Online Platform (VLOP) under the DSA by the European Commission.¹⁸¹ This means that Wikipedia is subject to the special systemic risk provisions (Article 34 and 35) under the DSA, which require VLOPs to assess any systemic risks stemming from their services, including relating to negative effects on "electoral processes",¹⁸² which may stem from the dissemination of disinformation;¹⁸³ and for VLOPs to put in place "mitigation measures" to address these systemic risks.¹⁸⁴ In this regard, VLOPs, including Wikipedia, must publish a risk assessment, and put in place risk mitigation measures, which are subject to an independent audit, with the auditor publishing an audit report, and an audit implementation report, with recommendations for the VLOP.¹⁸⁵

Second, and as mentioned above, Wikipedia's VLOP designation is also important on how Wikipedia and the Wikimedia Foundation is viewed under the DSA. Importantly, the European Commission in its designation Decision considered that Wikipedia was an "online platform" within the meaning of Article 3(i) DSA;¹⁸⁶ with the Commission considering that the Wikimedia Foundation is the legal entity "which alone operates" Wikipedia, and as such the Commission's designation was addressed to the Wikimedia Foundation.¹⁸⁷

¹⁷⁹ Digital Services Act.

¹⁸⁰ Digital Services Act, Recital 9.

¹⁸¹ See European Commission, Commission Decision designating Wikipedia as a very large online platform in accordance with Article 33(4) of Regulation (EU) 2022/2065 of the European Parliament and of the Council, 24 March 2023, C(2023) 2742 final.

¹⁸² Digital Services Act, Article 34(1)(c).

¹⁸³ Digital Services Act, Recital 84.

¹⁸⁴ Digital Services Act, Article 35.

¹⁸⁵ Digital Services Act, Article 37.

¹⁸⁶ European Commission, Commission Decision designating Wikipedia as a very large online platform in accordance with Article 33(4) of Regulation (EU) 2022/2065 of the European Parliament and of the Council, 24 March 2023, C(2023) 2742 final, section 1.

¹⁸⁷ European Commission, Commission Decision designating Wikipedia as a very large online platform in accordance with Article 33(4) of Regulation (EU) 2022/2065 of the European Parliament and of the Council, 24 March 2023, C(2023) 2742 final, section 6.

Notably, in late 2024, the Wikimedia Foundation published its first DSA risk assessment,¹⁸⁸ which set out “the Wikimedia Foundation’s assessment of systemic risks linked to the use of Wikipedia in the EU”, and “require[d] the Foundation (as the hosting provider of Wikipedia) to not only assess risks, but also determine what risk mitigations are appropriate”.¹⁸⁹ This risk assessment includes important risks and mitigation measures relating to Wikipedia and election disinformation in the EU. Further, the independent audit report¹⁹⁰ and audit implementation report¹⁹¹ have also been published, which also specifically address mitigation measures for election disinformation risks in the EU.

4.1 Wikipedia’s risk assessment report

Wikipedia’s inaugural risk assessment was published in the form of a spreadsheet (“systemic risk and mitigation measures” register (SRAM)),¹⁹² a methodological explanation,¹⁹³ and a Cover Note.¹⁹⁴ Regarding the methodology followed in Wikipedia’s SRAM register, it is noted that following the definitions provided in Articles 34 and 35 DSA, the Wikimedia Foundation focused on the risks of systemic nature and scale in the EU, stemming from Wikipedia’s design, functioning or use. The methodology used prioritised proportionality, in the selection and assessment of risks focusing only on risks that are severe, probable, and specific to Wikipedia’s operations in the EU taking into account the Wikimedia Foundation’s overall societal presence and impact. Therefore, the Wikimedia Foundation clarified that if an in-scope risk is relatively low in severity and probability in this specific context, it does not need to receive as much attention in this DSA-specific exercise, regardless of how much attention they might otherwise give that risk.¹⁹⁵

More specifically, if a given risk could be substantially lower for Wikipedia than the same or similar risk on third party platforms – such as “large, commercial, highly personalised, paid-amplification, virality-driven social media” – then this might also be considered in the

¹⁸⁸ Wikimedia Foundation, “WMF 23-24 approved DSA SRAM (Risk Register)”, [https://foundation.wikimedia.org/wiki/File:WMF_23-24_approved_DSA_SRAM_\(Risk_Register\).zip](https://foundation.wikimedia.org/wiki/File:WMF_23-24_approved_DSA_SRAM_(Risk_Register).zip).

¹⁸⁹ Wikimedia Foundation, “2023 EU Systemic Risk Assessment - Cover note”, 31 August 2023, https://upload.wikimedia.org/wikipedia/foundation/5/5a/Wikipedia_DSA_SRA_submission_cover_note_31Aug2023.pdf

¹⁹⁰ Wikimedia Foundation, “Wikipedia DSA Audit Report 2023-24 Public.pdf”, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_Audit_Report_2023-24_Public.pdf.

¹⁹¹ Wikimedia Foundation, “Wikipedia DSA Audit Implementation Report 2023-24 Public.pdf”, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_Audit_Implementation_Report_2023-24_Public.pdf.

¹⁹² Wikimedia Foundation, ‘DSA SRAM Register’ (HTML Spreadsheet, Q2 2023) [https://foundation.wikimedia.org/wiki/File:WMF_23-24_approved_DSA_SRAM_\(Risk_Register\).zip](https://foundation.wikimedia.org/wiki/File:WMF_23-24_approved_DSA_SRAM_(Risk_Register).zip).

¹⁹³ Wikimedia Foundation, ‘Methodology: EU DSA Systemic Risk Assessments and Mitigations Register’, https://foundation.wikimedia.org/wiki/Legal:Wikimedia_Foundation_EU_Compliance/DSA_Publication_Archive.

¹⁹⁴ Wikimedia Foundation, ‘EU Systemic Risk Assessment – Cover Note’, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_SRA_submission_cover_note_31Aug2023.pdf.

¹⁹⁵ Wikimedia Foundation, ‘Methodology: EU DSA Systemic Risk Assessments and Mitigations Register’, part A, https://foundation.wikimedia.org/wiki/Legal:Wikimedia_Foundation_EU_Compliance/DSA_Publication_Archive.

assessment of the Wikipedia-specific risk.¹⁹⁶ Given this, the Wikimedia Foundation's risk selections, assessments and mitigations should not be compared with those of other VLOPs. This approach was particularly important in the assessment of the risk of disinformation taking into account that a global community of committed volunteers has created a unique repository of encyclopedic data, available and culturally relevant to more than 300 language communities. This volunteer-led, self-governing model of Wikipedia allows advancing knowledge equity and combating disinformation through collective contributions and open debate about content moderation, rapidly removing content of coordinated disinformation campaigns.¹⁹⁷

Moreover, the methodological description particularly mentioned that Wikipedia's intra-EU disinformation risks (and mitigations) would be different than for-profit services on which extremist or political actors can pay for preferable treatment (e.g. content placement), and whose content will not be as susceptible to editing or erasure by a diverse, global public working for the public interest. For this reason, the Wikimedia Foundation declared that their choice to make the fight against disinformation a priority (as it will be seen in the mitigation measures description below) for the year ahead did not mean that the risk was as large, for Wikipedia, as it was for other platforms that similarly focus their immediate efforts on combatting disinformation.¹⁹⁸

The Wikimedia Foundation's DSA-specific register is grounded in the Wikimedia Foundation's existing Human Rights work, including past Human Rights Impact Assessment exercises, ongoing Human Rights Due Diligence for new/evolving projects. After risks are raised as part of the human rights work, a second-level DSA-specific assessment is therefore conducted, that considers the following factors: (1) To what extent does that global risk apply to Wikipedia, versus other Wikimedia projects and initiatives? (2) To what extent is the risk "systemic"? (3) To what extent is that risk systemic in the EU specifically, taking into account its relatively strong rule of law, fundamental rights protections, digital regulation, and its various "safety nets" for vulnerable individuals? (4) To what extent is the risk due to the "design or functioning" of Wikipedia and related systems, and their use?¹⁹⁹ After the selection of risks to be included in SRAM based on these criteria, the assessment was conducted examining: (1) the likelihood of a risk manifesting, and (2) the impact a risk is likely to have, if it manifests, taking into account the mitigations already in place.²⁰⁰

¹⁹⁶ Wikimedia Foundation, 'Methodology: EU DSA Systemic Risk Assessments and Mitigations Register', (n 139) part A, [https://foundation.wikimedia.org/wiki/File:WMF_23-24_approved_DSA_SRAM_\(Risk_Register\).zip](https://foundation.wikimedia.org/wiki/File:WMF_23-24_approved_DSA_SRAM_(Risk_Register).zip).

¹⁹⁷ Wikimedia Foundation, 'EU Systemic Risk Assessment – Cover Note', p.4, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_SRA_submission_cover_note_31Aug2023.pdf.

¹⁹⁸ Wikimedia Foundation, 'EU Systemic Risk Assessment – Cover Note', part C, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_SRA_submission_cover_note_31Aug2023.pdf.

¹⁹⁹ Wikimedia Foundation, 'EU Systemic Risk Assessment – Cover Note', part B, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_SRA_submission_cover_note_31Aug2023.pdf.

²⁰⁰ Wikimedia Foundation, 'EU Systemic Risk Assessment – Cover Note', p.2, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_SRA_submission_cover_note_31Aug2023.pdf.

Based on the underlying materials for a global risk assessment, the necessity and appropriateness of mitigations as specified in the DSA were then evaluated against the risks listed in the SRAM Register. In deciding on mitigations, the risks were balanced against “US Constitutional” and “EU/international fundamental rights” (e.g. avoiding undue interference with freedom of expression/information).²⁰¹

The first point to note about the risk assessment is that out of 11 identified risks, two are specifically related to election disinformation. In this regard, first, the risk assessment identifies “disinformation regarding civic and electoral processes” as a specific risk for Wikipedia, and notes that “actors interested in a particular political/electoral/civic outcome” could launch “coordinated campaigns to insert misleading content into Wikipedia”, reducing the “broader reliability of content, misleading readers, and spreading disinformation”.²⁰² As such, these risks “interfere with and diminish users’ freedoms of expression and thought, right to participate in civic and political life, and right to good administration”.²⁰³

Crucially, the risk assessment also identifies a number of “mitigation measures” relating to disinformation regarding election processes implemented. The first mitigation measure is that “volunteer moderators maintain editorial policies, and remove biased or otherwise inappropriate content in accordance with applicable Terms of Use and policies”.²⁰⁴ Second, there was the development of an “institutional strategy to counter disinformation”. The third mitigation measure was the establishment of “temporary task forces (Disinformation Response Teams),” composed of “both staff and community members”, to “support volunteer communities in countering coordinated disinformation campaigns during significant events, such as elections, that may be the target of malicious actors”.²⁰⁵ This was to be “Implemented as needed”. Fourth, the establishment of a “long-term internal working group to coordinate activities relating to legal, Trust & Safety, communications, partnerships, advocacy, and other activities during significant events, such as elections, that may be the target of malicious actors”.²⁰⁶ Fifth, the development of tools to support moderation and content-curation work by volunteers. Sixth, the deployment of an e-learning module to assist volunteer editors in identifying and combatting disinformation.²⁰⁷ Seventh, the establishment of academic partnerships to encourage further research on disinformation on Wikipedia. And finally, “24/7 rapid reaction capabilities for bomb threats targeting polling stations” via a “dedicated reporting channel triggering fast-track triage and response”.²⁰⁸ As such, the mitigation measures were very much a mix of content-related

²⁰¹ Wikimedia Foundation, ‘Methodology: EU DSA Systemic Risk Assessments and Mitigations Register’, part B, https://foundation.wikimedia.org/wiki/Legal:Wikimedia_Foundation_EU_Compliance/DSA_Publication_Archivev.

²⁰² Wikimedia Foundation, “WMF 23-24 approved DSA SRAM (Risk Register)”, section 2.

²⁰³ Wikimedia Foundation, “WMF 23-24 approved DSA SRAM (Risk Register)”, section 2.

²⁰⁴ Wikimedia Foundation, “WMF 23-24 approved DSA SRAM (Risk Register)”, section 2.

²⁰⁵ Wikimedia Foundation, “WMF 23-24 approved DSA SRAM (Risk Register)”, section 2.

²⁰⁶ Wikimedia Foundation, “WMF 23-24 approved DSA SRAM (Risk Register)”, section 2.

²⁰⁷ Wikimedia Foundation, “WMF 23-24 approved DSA SRAM (Risk Register)”, section 2.

²⁰⁸ Wikimedia Foundation, “WMF 23-24 approved DSA SRAM (Risk Register)”, section 2.

practices (i.e., community moderation of disinformation); and establishing and maintaining institutional strategy, task forces, internal working groups, e-learning to support volunteers, and dedicated reporting channels.

Secondly, the risk assessment also noted that disinformation regarding “historical/geographical narratives” was also a risk for Wikipedia, where “actors seeking to advance specific ideologies could try to manipulate encyclopedic entries related to historical or geographical narratives in order to further their agendas”; and in certain cases, “such efforts could run parallel to political disinformation campaigns” and “further inflame existing social and political tensions in some societies”.²⁰⁹ Similarly, these risks interfere with and diminish users’ freedoms of expression and thought and “rights to participate in civic and political life”.²¹⁰

And in terms of mitigation measures, the same mitigation measures would be applied as for election disinformation, including “[v]olunteer moderators maintain editorial policies, and remove biased or otherwise inappropriate content in accordance with applicable Terms of Use and policies;” development of an “institutional strategy to counter disinformation”; and “provision of support” to volunteers and communities in identifying and rectifying cases of disinformation regarding historical narratives.²¹¹

The above reported measures are aligned with the explanation included in the Cover Note that unlike other commercial platforms, the Wikimedia Foundation aims to focus on creating conditions for success by developing practical tools, supporting community structures, providing training, and encouraging discussion “rather than enforcing changes to content policies or replacing efficient community mechanisms (such as for handling complaints or addressing specific categories of problematic content)”.²¹²

4.2 Audit report

Crucially, in addition to carrying out the risk assessment and implementing mitigation measures under Article 34 and 35 DSA, VLOPs, including Wikipedia, are subject, “at their own expense and at least once a year”, to “independent audits” to “assess compliance” with the obligations under Article 34 and 35.²¹³ The auditor issues a report with audit opinion on whether the VLOP “compiled the obligations” under the DSA, including Article 34 and 35, with a possible rating of “positive”, “positive with comments”, or “negative”.²¹⁴

²⁰⁹ Wikimedia Foundation, “WMF 23-24 approved DSA SRAM (Risk Register)”, section 3,

²¹⁰ Wikimedia Foundation, “WMF 23-24 approved DSA SRAM (Risk Register)”, section 3.

²¹¹ Wikimedia Foundation, “WMF 23-24 approved DSA SRAM (Risk Register)”, section 3.

²¹² Wikimedia Foundation, ‘EU Systemic Risk Assessment – Cover Note’, p.5, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_SRA_submission_cover_note_31Aug2023.pdf.

²¹³ Digital Services Act, Article 37(1).

²¹⁴ Digital Services Act, Article 37(4).

First, and as noted above, the auditors noted that a “clear distinction” needs to be drawn between the Wikimedia Foundation as an “independent organisation,” and the “volunteer editor community of Wikipedia, which operates in a decentralised manner” separate from the Wikimedia Foundation.²¹⁵ Thus, the Wikimedia Foundation “establishes the governance mechanisms within which the community acts”, such through the Universal Code of Conduct, “which detail a baseline of behaviour for collaboration on Wikimedia projects worldwide and the types of cases that fall within each of their scopes”; while “each community can go further to ‘develop policies that take account of local and cultural context, while maintaining the criteria listed here as a minimum standard’.”²¹⁶ Crucially, the auditors note that the only actions within the scope of the audit “are those conducted directly” by the Wikimedia Foundation, and “not the actions of the volunteer community.”²¹⁷

Notably, the audit report for Wikipedia was “positive with comments” in relation to its compliance with Article 34 and 35. The audit report first noted that the Wikimedia Foundation had identified “11 systemic risks,”²¹⁸ and moved to examine the “appropriateness” of the systemic risks identified,²¹⁹ and for the purpose of this Mapping Report, the focus is on those concerning election disinformation.

First, the audit report examined the first systematic risk, namely “disinformation regarding civic and electoral processes, and conflicts”. The audit report described the risk as relating to the possibility of having wrong and/or harmful content regarding civic and electoral processes on Wikipedia.²²⁰ The auditor also highlighted that electoral rights are entwined with the establishment and protection of democracy -the system of governance in the majority of countries accessing Wikipedia. The conscientious and appropriate exercise of the right to vote was deemed contingent upon the provision of objective, unbiased, and substantial information to voters.²²¹

Of particular note, the audit report emphasised that the “risk of electoral and civic processes” being targeted by “disinformation” was “increased” due to three main reasons.²²² First, the audit period was an “election season”, with the EU elections conducted in June 2024, snap UK elections

²¹⁵ Wikimedia Foundation, “Wikipedia DSA Audit Report 2023-24 Public.pdf”, p. 9.

²¹⁶ Wikimedia Foundation, “Wikipedia DSA Audit Report 2023-24 Public.pdf”, p. 9; citing Wikimedia Foundation, Wikimedia Foundation Universal Code of Conduct, Section 1, https://foundation.wikimedia.org/wiki/Policy:Universal_Code_of_Conduct.

²¹⁷ Wikimedia Foundation, “Wikipedia DSA Audit Report 2023-24 Public.pdf”, p. 9.

²¹⁸ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 71, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_Audit_Report_2023-24_Public.pdf.

²¹⁹ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 72.

²²⁰ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 72.

²²¹ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_Audit_Report_2023-24_Public.pdf.

²²² Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 73, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_Audit_Report_2023-24_Public.pdf.

in June 2024, and US elections in November 2024. Secondly, given the “open-to-participate nature of editing” on Wikipedia, a disinformation campaign would be “easier”.²²³ Third, due to the “level of reliability” of Wikipedia, “which is framed by WMF and commonly perceived by the internet community” as a “free online encyclopaedia”, a Wikipedia entry about a topic “might be seen as more credible compared to a post about that topic from an account on another platform”.²²⁴ Indeed, the audit report noted that the connection between the “mode of operation of Wikipedia” and the “possibility of being targeted by disinformation” escalated the risk to the level of “being systemic in nature”.²²⁵ The availability of such information, not its inaccessibility, was considered “problematic” due to the risks it posed to the exercise of fundamental rights, and the audit report found the risk fell within the scope of the risks under Article 34(1)(c) DSA as “accurately identified” by the Wikimedia Foundation.²²⁶

Notably, the audit report did highlight that an aspect that must be considered while identifying systemic risks is the “effects of regional and linguistic aspects”, and that the risk assessment did mention these for “some of the identified risks”, particularly under the “risk of disinformation regarding historical/geographical narratives”; but there was “no mention of an aspect that is specific to a Member State”.²²⁷ In this regard, the audit report recommended that the Wikimedia Foundation “provide more information on the effects of regional and linguistic aspects on the risks identified, and Member State-specific assessments, if available”.²²⁸

The audit report then turned to the mitigation measures, and considered the mitigation measures for disinformation risks, including the risk of disinformation regarding civic and electoral processes. The audit report noted that the disinformation risks on Wikipedia “are inextricably linked to the editing and content moderation practices of both the community and WMF, although they are predominantly conducted by the community.”²²⁹ In this regard, the audit report emphasised that the Wikimedia Foundation’s role in content moderation is “limited to the provision of an overarching, functional structure” that interferes “only in certain cases”.²³⁰ Thus, an “efficient measure here could target improving either the community’s capacity to recognise and take action against disinformation or improving the communication channels between users, the community, and WMF on the issues that are to be handled by WMF instead of the community”.²³¹

²²³ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 73.

²²⁴ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 73.

²²⁵ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 73.

²²⁶ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 73.

²²⁷ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 78.

²²⁸ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 80.

²²⁹ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 86.

²³⁰ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 86.

²³¹ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 86.

Crucially, the audit report focused on the “common mitigation measures”, which included (a) recognising the practical necessity arising from the community-driven content moderation at Wikipedia; (b) maintaining editorial policies, removing biased or otherwise inappropriate content in accordance with the Wikimedia Foundation’s Term of use and applicable policies; and (c) developing an institutional strategy. In particular, the audit report stated that having an “institutional strategy” and “making policies and guidelines accessible for the community” that is “predominantly” doing the editing and content moderation is a “reasonable mitigation measure”.²³² However, the audit report examined how these strategies and policies are put in place and communicated to the community, and noted that the Wikimedia Foundation “does not have a single main policy against disinformation” but a set of “policies and guidelines supporting the volunteer community”.²³³ Further, there are also “guidelines drafted in practical and simple language for inexperienced volunteers”; and crucially, all of these policies and guidelines “refer to the ‘five pillars’ of Wikipedia,” emphasising a “neutral point of view,” “citation of reliable sources,” and “verifiability of information” to be provided when creating or editing content on Wikipedia.²³⁴ The efficiency of such guidelines was “increased with the e-learning module”, which was another measure by the Wikimedia Foundation; while the Wikimedia Foundation planned to develop tools to support the moderation as well as curation of content on Wikipedia, which was a “common measure for all these risks and an action that would contribute to the effectiveness of all other related measures”.²³⁵

Finally, in relation to election-related disinformation, the audit report highlighted how the Wikimedia Foundation had established “a disinformation response team” and was providing administrators with the “information necessary to prepare for possible disinformation campaigns”; while another measure introduced was the establishment of “temporary task forces and internal working groups when needed”, which was a “reasonable and proportionate measure” considering the “temporal and fluctuating effects” that disinformation risks may pose. And given the immediate nature of the threat, having a “separate reaction channel for bomb threats 24/7 is indeed a reasonable, proportionate, and effective measure”. In sum, the audit report attested that the Wikimedia Foundation provided “satisfactorily reasonable, proportionate, and effective mitigation measures” for disinformation risks.

It should also be mentioned that the audit report also discussed mitigation measures under Article 35(1)(h), which include VLOPs “initiating or adjusting cooperation with other providers of online platforms or of online search engines through the codes of conduct and the crisis protocols referred to in Articles 45 and 48 respectively”. The audit report noted that the Wikimedia Foundation “is not a signatory to the 2022 Code of Practice on Disinformation and does not refer to it in relation to the mitigation measures concerning disinformation or collaboration”. Of

²³² Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 86.

²³³ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 86.

²³⁴ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 86.

²³⁵ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 86.

particular note, the audit report recommended that the Wikimedia Foundation “monitor the developments surrounding the 2022 Strengthened Code of Practice on Disinformation and other potential codes of conduct to be issued in accordance with DSA as well as participate, if applicable”.²³⁶

The auditor also recommended that the Wikimedia Foundation reevaluate the SRAM Register and corresponding mitigation sections to comply with the formal and substantial requirements of Article 35 DSA; elaborate further on how a given mitigation measure will be implemented, when it will be completed, and how it will contribute to the mitigation of the risks; include a mapping of their mitigation measures to the measures required under Article 35 (1) DSA.

4.3 Audit implementation report

Since the audit report that was “positive with comments”, pursuant to Article 37(6) DSA an audit implementation report was also published.²³⁷ Two points should be mentioned. First, in relation to the recommendation to monitor the developments surrounding the 2022 Strengthened Code of Practice on Disinformation and other potential codes of conduct to be issued in accordance with DSA as well as participate, if applicable. In this regard, the audit implementation report states that “relevant staff within the Legal department will monitor developments surrounding the 2022 Strengthened Code of Practice on Disinformation and other potential codes of conduct to be issued in accordance with DSA as well as participate, if applicable”. And that “[f]or now, [the Wikimedia Foundation] sees the 2022 Code as part of an evolving body of good practice that informs our approach to a global challenge, and that it therefore merits close monitoring, and ongoing exchanges with its formal adherents”.²³⁸

Second, in relation to the recommendation to revisit whether “more geo-specific detail can be incorporated in any assessments of systemic risks”, the audit implementation report states that systematic risks can “impact linguistic communities represented on Wikipedia according to their unique characteristics”, including the number of “active editors and readers”. As such, the Wikimedia Foundation provided “additional analysis in its second systemic risk assessment regarding how minority linguistic communities may be disproportionately impacted by specific risks”. In particular, language versions of Wikipedia representing minority linguistic communities, “which typically have fewer active editors than those representing larger linguistic communities,” may be “more vulnerable” to risks relating to the propagation of disinformation “due to the inherent challenges of having fewer people contributing to, editing, and patrolling pages”.²³⁹ The

²³⁶ Holistic AI, ‘Wikipedia DSA Audit Report 2023–24 Public’, p. 95.

²³⁷ Wikimedia Foundation Inc, ‘Wikipedia DSA Audit Implementation Report 2023–24: Pursuant to DSA Article 37(6)’, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_Audit_Implementation_Report_2023-24_Public.pdf.

²³⁸ Wikimedia Foundation Inc, ‘Wikipedia DSA Audit Implementation Report 2023–24: Pursuant to DSA Article 37(6)’, p. 34, https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_Audit_Implementation_Report_2023-24_Public.pdf.

²³⁹ Wikimedia Foundation Inc, ‘Wikipedia DSA Audit Implementation Report 2023–24: Pursuant to DSA Article 37(6)’, p. 21.

second risk assessment acknowledges the “unique risks” faced by “linguistic minorities” in the EU and cites a known example of when such risks materialised in an EU member state, namely a report in which this phenomenon was evaluated on Croatian Wikipedia in 2021²⁴⁰ (described as a unique situation in which the contributors to/audience of a single language version of Wikipedia are relatively heavily concentrated in a single country).²⁴¹

5. Concluding remarks

In light of the foregoing, there are a number of concluding points to consider from this mapping of Wikipedia’s policies and risk-mitigation measures that are applicable to disinformation during elections in the EU. First, it should be pointed out that Wikipedia and the Wikimedia Foundation’s policies on false information are very much aligned with the widely-adopted definition of disinformation by the European Commission, and implemented under the EU Code of Conduct on Disinformation (“false or misleading content that is spread with an intention to deceive or secure economic or political gain and which may cause public harm”). A crucial feature of this definition is the element of “intention,” which distinguishes disinformation from misinformation; and Wikipedia and the Wikimedia Foundation policies also place “intention” as a central notion in policies targeting false information (“engaging in false statements,” including “posting or modifying content with the intent to deceive or mislead others”).

Second, Wikipedia relies on a system of multifaceted community moderation, content rules, and guidelines and intervention possibilities at the level of the Wikimedia Foundation in order to address election-related disinformation. Indeed, Wikipedia policies on false and misleading information apply directly to election disinformation. The foundational content policies, such as Neutral Viewpoint, Verifiability, and No Original Research, govern editorial activities and support the aim of providing accurate, reliable, and unbiased information on Wikipedia. These principles are mostly enforced through moderation practices guided by the community, involving transparent editing processes, discussion pages and deletion procedures. Such moderation is supported by automated tools, which are thought of as functioning as a first layer of moderation in addition to manual review. These include bots, which can reverse obvious vandalism (including false content) without human intervention, and systems, which generate edit scores to assist in prioritising for manual review.

Third, in addition to community-led and technical moderation efforts, Wikipedia’s approach to election disinformation is supported by behavioural standards and rules established by the

²⁴⁰ See Croatian WP Disinformation Assessment - Final Report EN, https://en.wikipedia.org/wiki/File:Croatian_WP_Disinformation_Assessment_-_Final_Report_EN.pdf; and Wikipedia, “Croatian Wikipedia”, https://en.wikipedia.org/wiki/Croatian_Wikipedia.

²⁴¹ Wikimedia Foundation Inc, ‘Wikipedia DSA Audit Implementation Report 2023–24: Pursuant to DSA Article 37(6)’, p. 21.

Wikimedia Foundation, most notably the Universal Code of Conduct and the Terms of Use. The Terms of Use and the Code of Conduct provide the formal basis for a range of intervention options by the Wikimedia Foundation, also known as office actions, which the Wikimedia Foundation could deploy if for example community action has proven ineffective, or if legal considerations demand intervention by the Wikimedia Foundation, such as for example page protection and range blocks. Most of these actions can also be carried out by the community.

Fourth, regarding unacceptable behaviour, the Universal Code of Conduct identifies a range of unacceptable behaviour relevant for election disinformation. This includes deliberately introducing “biased, false, inaccurate or inappropriate content”, and the hindrance, impedance, or other obstruction of content creation as well as its maintenance. Including arbitrarily and unexplainably repeatedly removing content and systematically manipulating content to favour specific interpretations of facts or points of view, including by means of unfaithfully or deliberately falsely rendering sources and altering the correct way of composing editorial content. Similarly, the Terms of Use prohibit engaging in false statements, specified as “posting or modifying content with the intent to deceive or mislead others”. Other prohibited actions include, impersonating individuals, concealing affiliations where disclosure is required by the terms, or using the name or username of another person with the intent to deceive, and engaging in fraudulent behaviour.

Fifth, as promotional editing was identified as a concern during the 2024 European Parliament elections, the provisions in the Terms of Use on paid contributions are also relevant in this context of election disinformation. These provisions require contributors who receive, or expect to receive, compensation for their edits to disclose the organisation or individual on whose behalf they are contributing. This disclosure may be made on the user page, the relevant discussion page or in the edit summary. These measures are intended to promote transparency and enable the community to understand the context in which edits are made, particularly in politically sensitive areas. In addition, promotional editing is incompatible with Wikipedia’s own definition as “not an advertising platform”, as set out in its Five Pillars.

Sixth, it seems quite clear that the Wikimedia Foundation has taken significant steps to protect Wikipedia from election-related disinformation.²⁴² This includes a disinformation task force which has been established to protect the integrity of the elections.²⁴³ Indeed, during previous elections, around 2,000 election-related pages were protected, meaning they could only be edited by users

²⁴² Ryan Merkley, ‘What Wikipedia Saw during Election Week in the U.S., and What We’re Doing next’ (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

²⁴³ Ryan Merkley, ‘What Wikipedia Saw during Election Week in the U.S., and What We’re Doing next’ (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

with certain access rights, depending on the level of protection applied.²⁴⁴ In addition, these pages were monitored in real time by over 56,000 volunteer editors.²⁴⁵ Furthermore, the Wikimedia Foundation's disinformation task force recorded and evaluated 18 relevant events, and in addition, the community reversed an estimated 800 edits on election-related pages between 3 and 7 November 2020.²⁴⁶ Notably, neither the task force members nor the Wikipedia administrators detected any large-scale state-sponsored disinformation during that period.²⁴⁷

Seventh, in the EU context, the Wikimedia Foundation established a Disinformation Response Team, similar to the internal task force, providing support to the European community before the 2024 European Parliament elections.²⁴⁸ As reported, no large-scale disinformation was identified during the 2024 European Parliament elections.²⁴⁹ Communities expressed concern mainly about vandalism and promotional editing, which through experience they knew how to deal with independently, without interference from the Wikimedia Foundation.²⁵⁰ In addition, on the use of AI for edits, no AI use for disinformation was identified.²⁵¹ The absence of major incidents put forward is the fact that misinformation manifested itself mainly on social media, sources that

²⁴⁴ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

²⁴⁵ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

²⁴⁶ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

²⁴⁷ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

²⁴⁸ Wikimedia Foundation, 'Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024' https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

²⁴⁹ Wikimedia Foundation, 'Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024' https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

²⁵⁰ Wikimedia Foundation, 'Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024' https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

²⁵¹ Wikimedia Foundation, 'Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024' https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

Wikipedia usually considers unreliable.²⁵² These findings are consistent with previous election rounds, such as Poland.²⁵³

Eighth, in addition to establishing the Internal Disinformation Task Force, efforts have been made to invest in research and development. More specifically, the Wikimedia Foundation's research team has been collaborating with international universities to enhance the community's ability to detect disinformation and vandalism by focusing on various technical tools mentioned earlier as well, such as ORES, source verification algorithms, and monitoring tools.²⁵⁴ At the community level, specific guidelines were established and debated during that period, including the requirement of a minimum of three reliable sources to determine the winner in each state, a waiting period of at least 12 hours after the polls closed before publishing results, and an absolute prohibition of original research sources.²⁵⁵

Finally, the above points will be taken into account as the DEM-Debate project proceeds to critically assess the application of EU regulatory frameworks on disinformation during elections to community-governed platforms, using Wikipedia as a case study. Indeed, these points are also relevant for developing policy recommendations for specific legislative and regulatory reforms of the EU regulatory framework to better counter disinformation during elections, building upon the findings in relation to community-governed platforms' approach to disinformation.

²⁵² Wikimedia Foundation, 'Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024' https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

²⁵³ Wikimedia Foundation, 'Legal/Community Resilience and Sustainability/Trust and Safety/EP Report 2024' https://meta.wikimedia.org/wiki/Wikimedia_Foundation/Legal/Community_Resilience_and_Sustainability/Trust_and_Safety/EP_report_2024.

²⁵⁴ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.

²⁵⁵ Ryan Merkley, 'What Wikipedia Saw during Election Week in the U.S., and What We're Doing next' (*Wikimedia Foundation*, 17 December 2020), <https://wikimediafoundation.org/news/2020/12/17/what-wikipedia-saw-during-election-week/>.