PROMPT

Wikipedia and the challenges of disinformation

Strategies and recommendations for the resilience of the Wikimedia movement

Deliverable D7.1

PROMPT is a pilot project co-financed by the European Union under Grant Agreement CNECT/LC-026293

WIKIMEDIA
FRANCE

Introduction

In 2026, Wikipedia, which is celebrating its twenty-fifth anniversary, has established itself as a critical information infrastructure. According to a survey by the Marsouin laboratory conducted with a representative sample of the French population (November 2025), 85% of French people consult the encyclopedia and **73% of them use the encyclopedia to verify information found elsewhere.** It is highly likely that this high level of trust is found in similar proportions across all European Union countries. This trust among European citizens places the Wikimedia movement at the heart of the fight against information manipulation. This document analyzes how, through its founding principles, technological tools, and partnerships, the movement protects the integrity of free knowledge.

I) The ecosystem of trust: governance and mechanisms of regulation

Wikipedia's effectiveness in combating misinformation does not rely on a central authority or opaque algorithmic moderation, but on a peer governance model. Unlike social media, Wikipedia has built an ecosystem where every edit is traceable, contestable, and correctable. This section details the technical, normative, and institutional pillars that guarantee the encyclopedia's integrity.

1. A gradual and proactive moderation architecture

Wikipedia's open principle ("Anyone can edit") is its main strength: the multitude of eyes on the articles allows for the almost instantaneous detection of anomalies. This ability for everyone to edit the encyclopedia, initially perceived as a sign of a project lacking seriousness, is now a real barrier to manipulation attempts and is also perceived as such by the French population, with 55% believing it to be an improvement.[1] of theFrench people who approve of this model.

The contributor community also has protection levers to reduce the "attack surface" of articles.

[1]Survey conducted by the Marsouin IMT Atlantique laboratory in November 2025 (link)

A. Hybrid filtering: Between AI and human intelligence

Each modification goes through a rigorous verification funnel:

- Automated analysis (Bots and ORES/Machine Learning): Even before human intervention, tools for *Machine Learning* have been developed. They assess the probability that a change will be harmful. The robots instantly revoke crude vandalism (insults, page whitening).
- Coordinated patrolling: Volunteers use tools like the recent changes tracker to review the flow of changes in real time. Each contributor also has a personalized watchlist. During election periods, some groups of contributors update their watchlists to specifically monitor the pages of candidates and parties.
- Priority review marking: Edits deemed "suspicious" by AI are flagged to experienced moderators for thorough manual review, ensuring that subtle manipulations do not slip through the net.

B. Technical protection levels of pages

The community can also adjust the open threshold for an article based on the risk level:

- Semi-protection: This limits modifications to users registered for more than 4 days. This blocks most impulsive vandalism from new accounts or people using temporary accounts.
- Extended Semi-Protection: This restricts access to accounts with at least 500 contributions and 90 days of activity. It serves as the primary safeguard against "dormant accounts" or communications agencies attempting to infiltrate the encyclopedia in a coordinated manner over a specific period.
- Total Protection: The article is locked. Only a consensus reached on the talk page allows an administrator to make changes. This is the last resort in the event of an **edit war** (a conflict between several groups of contributors over the content of an article).

Administrators on Wikipedia are volunteers elected by the community of contributors and who thus have additional tools such as page protection or blocking users who do not respect the rules of the encyclopedia and the normative framework of the Wikimedia movement.

2. The regulatory framework

The fight against manipulation is also guided by fundamental texts that define the framework for contributing to the encyclopedia.

A. General Terms and Conditions of Use (GTC) and Transparency

THE CGU of the Wikimedia Foundation[2] impose strict obligations:

- Prohibition of deception: “Publishing or modifying any content with the intention of deceiving or leading others in the wrong direction” is not permitted on Wikimedia projects
- Paid contributions: Wikipedia allows sponsorship or professional contributions, but requires mandatory disclosure. Failure to disclose that one is paid to write an article is grounds for immediate removal from the user's account.

B. The Universal Code of Conduct (UCoC): Sanctions and Security

The UCoC[3]provides an essential dimension of human security:

- Combating harassment: By protecting moderators from pressure and threats (frequent in political discussions), the UCoC ensures that the community remains able to moderate without fear. A committee to enforce this code has been established at the Foundation level, supported by a professional team. *Trust and Safety*, which ensures its implementation.
- Dual level of sanction: Code violations can result in local sanctions (decided by the community, such as the removal of administration tools) or global sanctions (banning of all projects by the Foundation).

3. The Foundation's pivotal role: The Transparency Report 2025

[2] https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use/fr
[3] https://foundation.wikimedia.org/wiki/Policy:Universal_Code_of_Conduct

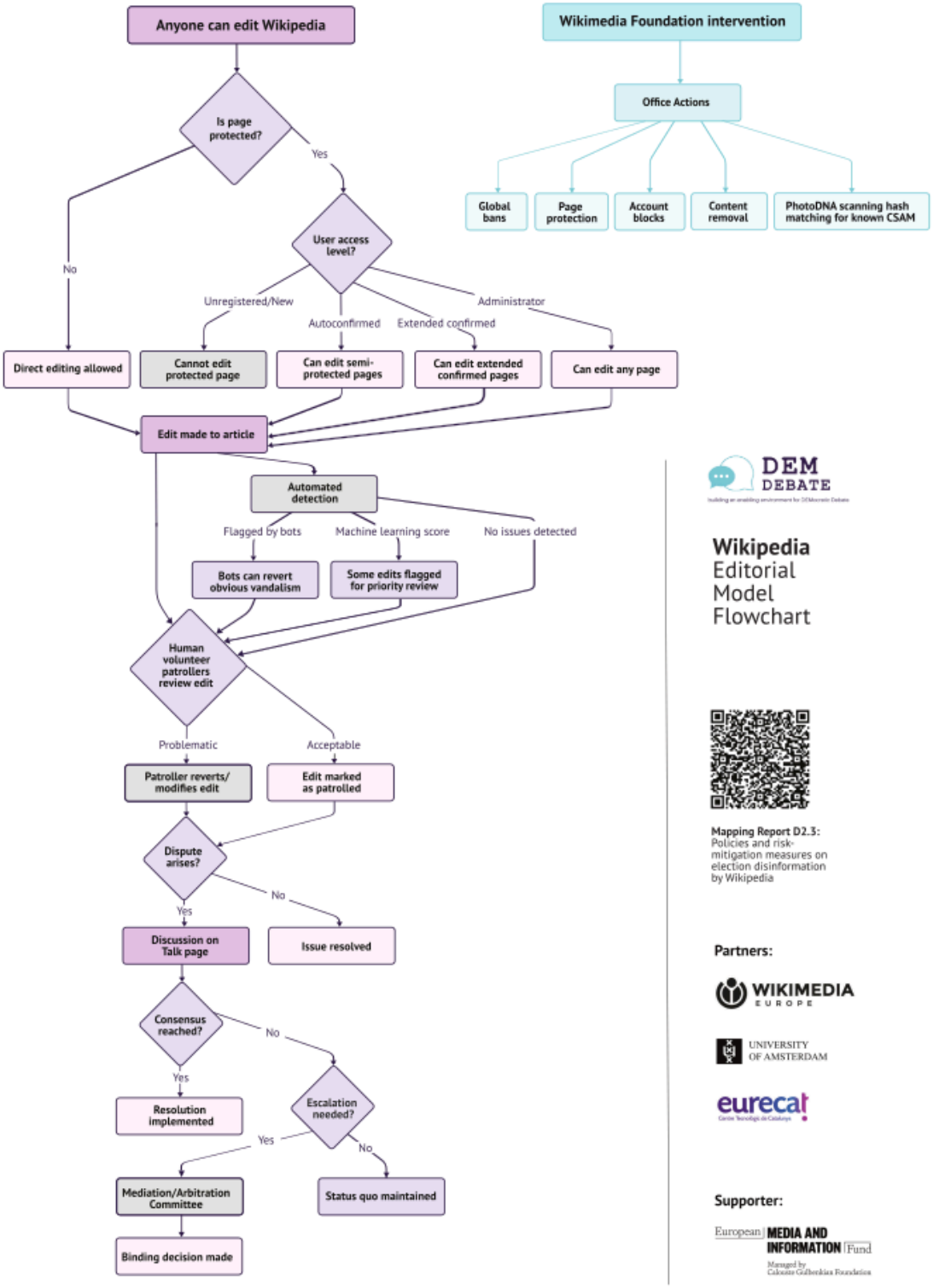
While the community manages the content, the Wikimedia Foundation (WMF) acts as the guarantor of the legal and technical security of the infrastructure. The latest Transparency Report [4]reveals crucial points:

- Resistance to state pressure: Despite hundreds of content removal requests from governments, the Wikimedia Foundation has consistently refused to remove content that did not violate any national law or the platforms' terms of service, in order to preserve its independence and the right to information. This independence is vital to preventing Wikipedia from becoming a tool of state propaganda.
- Global Bans: The Foundation intervenes directly to remove "Sockpuppet" networks (fake accounts) coordinated by influence agencies. The report notes the increasing sophistication of these attacks, requiring technical investigations of metadata (IP addresses, browser fingerprints).
- Human rights protection: The Foundation refuses to disclose contributor data to authorities without a strict legal basis, thus protecting patrollers fighting disinformation in authoritarian regimes.

The current functioning of the community and the various legal and human safeguards within the movement allow it, under normal circumstances, to cope with a wide variety of situations. However, like any system, it is not bulletproof and has several weaknesses identified by the Wikimedia movement.

[4] https://wikimediafoundation.org/who-we-are/transparency/2025-1/

Moderation architecture on English Wikipedia



DEM DEBATE

Wikipedia Editorial Model Flowchart

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

Partners:

WIKIMEDIA EUROPE

UNIVERSITY OF AMSTERDAM

eurecat

Supporter:

European MEDIA AND INFORMATION Fund

Managed by Calouste Gulbenkian Foundation

WIKIMEDIA FRANCE

II) Wikipedia under pressure: Threats, capture attempts and resilience.

In 25 years, Wikipedia has become a major online public space within the global digital ecosystem.It has therefore also become a battleground for the information war and cultural pressures that our societies are subjected to. Wikipedia has never been completely immune to this type of manipulation, but today the threat is no longer just isolated acts of vandalism, but structured strategies aimed at undermining the encyclopedia's legitimacy or restricting access to it. This section analyzes the mechanisms of internal "capture" and external geopolitical pressures.

1. The “Project Capture”

A. The case of the Croatian Wikipedia (2011-2020)

The case of the Croatian Wikipedia (*hr.wikipedia*) remains to this day the most documented example of a systemic failure of moderation through ideological capture.

For nearly ten years, a small group of far-right administrators managed to seize control of the project. By circumventing the community rules, they:

- **Rewriting history:** The articles concerning the Ustaše regime were modified to minimize war crimes and rehabilitate collaborationist figures, in total contradiction with the principle of neutrality.
- **Anti-LGBT propaganda:** Articles concerning abortion or other topics related to LGBT communities have also been manipulated to reflect openly homophobic opinions or those close to very conservative circles at the expense of sourced scientific and medical information.
- **Silencing dissent:** Any contributor attempting to restore neutrality was systematically harassed, blocked or banned under false pretenses.
- **Create an echo chamber:** By discouraging new contributors, the group created an environment where misinformation became the local editorial norm.

One of the reasons for this takeover of the Croatian Wikipedia stems from the 2003 decision to create national versions of Wikipedia instead of a single Serbo-Croatian language version encompassing the populations of several Balkan countries. In a region marked by recent conflicts, this initial decision quickly led to the project being hijacked by nationalist movements, something that would have been more difficult in a more

unified version. This also resulted in smaller, more fragmented communities, making it easier for a small, motivated, and organized group to seize control. Despite various warnings, it took time for the Wikimedia Foundation to intervene.

This case forced the movement to rethink its boundaries. After years of reporting, the WMF commissioned an independent external audit in 2021.

- Historical sanctions: The Foundation used its "Office Actions" to remove all the directors involved and issue blanket bans.
- Structural lesson: This case demonstrated that the "law of numbers" within a small community can be overturned. It served as a catalyst for the creation of the Universal Code of Conduct (UCoC) to enable the Foundation to intervene when local mechanisms are paralyzed.

The Foundation has published a full report on this capture of the Croatian Wikipedia.[5]

In this report, the Wikimedia Foundation mentions that this risk is not limited to the Croatian Wikipedia and that the movement must be vigilant, particularly for small linguistic communities where this risk is even greater.

B. The "WikiZédia" affair: Coordinated political infiltration (2022)

In France, and within the French-language Wikipedia, a similar situation did not, however, lead to the same outcome due to the greater diversity of the active community. Revealed during the 2022 French presidential campaign by journalist Vincent Bresson (who had infiltrated the campaign team of far-right presidential candidate Éric Zemmour), this affair highlighted a targeted attempt at "capture."

- The method: A clandestine cell called "WikiZédia" operated on Discord and Telegram to coordinate modifications aimed at "Zemmouring" the encyclopedia. Unlike classic vandalism, the objective was subtle: to smooth over the candidate's biography, remove quotes from historians critical of the French 'Vichy regime' (which collaborated with the Nazi regime during the Second World War), and increase the number of links to his page from other articles to boost his search engine optimization (SEO).
- The internal danger: The most alarming element was the participation of highly experienced contributors (including one member who was among the top 70

[5] The Case of Croatian Wikipedia, June 14, 2021, https://meta.wikimedia.org/wiki/File:Croatian_WP_Disinformation_Assessment_-_Final_Report_EN.pdf

contributors to the French-language version). This "pillar" status allowed the group to bypass standard safeguards and influence community discussions from within. Threats were also made against contributors who participated in discussions on topics targeted by this group.

- The answer: Thanks to an internal investigation conducted by volunteer administrators, seven accounts were banned and the changes made were reversed.

2. Interference strategies: Coordinated manipulation and "information laundering"

Current threats obviously extend beyond internal processes and internal data capture. They originate from state actors or private firms using cutting-edge techniques to manipulate content without triggering detection systems.

A. "Sockpuppeting" and troll farms

Manipulation no longer involves massive modifications, but rather the slow infusion of biases via networks of multiple accounts (*sockpuppets*) simulating a false consensus. The **Transparency Report 2025-1** details how the Foundation identifies and bans these networks through the analysis of metadata and coordinated voting behavior.

B. “Information Laundering” and the Use of AI (Portal Kombat Report)

One of the major risks identified by organizations like Viginum is that of AI-assisted "information laundering" to create fake sources to deceive moderators.

- The Network Portal Kombat[6]: As the Viginum report points out, structured propaganda networks use AI to generate thousands of articles on media-looking mirror sites.
- The trap for Wikipedia: These sites serve as fake secondary sources. A contributor, or a manipulation group, can then insert misinformation on Wikipedia by providing a reference that appears legitimate but was entirely generated by AI. This forces the community to be more vigilant not only about Wikipedia's content, but about the reliability of the entire media ecosystem. The Prompt project should allow us to better identify this type of content.

[6]Viginum's technical report on the Portal Kombat network (link)

Although some fake sites ended up on Wikipedia, collaboration with Viginum and the responsiveness of the community in France made it possible to quickly identify and remove them.

C. The Avisa Partners case: "Reputation management" by online agencies (2022)

The Avisa Partners case revealed the existence of a veritable commercial "influence factory" targeting Wikipedia.

- The operation: Press investigations (Mediapart, Reflets) have revealed that Avisa Partners (formerly iStrat) used dozens of fake accounts to modify the pages of its clients (CAC 40 companies, executives, foreign governments).
- The "Ripolinage": The goal was to erase controversies (financial scandals, convictions, environmental debates) and replace them with flattering narratives. The edits were made during office hours, with near-military discipline, sometimes using self-generated sources to validate biases.
- The sanction: The French Wikipedia community has carried out one of the largest cleanups in its history, identifying and deleting dozens of accounts linked to the agency. This affair has forced the Wikimedia Foundation to tighten its rules on undeclared paid contributions, now subject to immediate and permanent banishment.

3. The War for Access: Blockades and State Alternatives

Wikipedia's model of unlimited access to knowledge based on a set of reliable and verifiable sources often clashes with the national narratives of authoritarian regimes, leading to direct confrontation.

A. National blockades as a tool of censorship

The Transparency Report 2025 documents the Foundation's resistance to withdrawal demands. Some states then opted for total blockade:

- China: Maintains a permanent block on all language versions.
- Russia: It has multiplied threats of blocking and record fines, demanding the removal of information about the Ukrainian conflict. The WMF systematically prioritizes factual integrity, even at the risk of having its website shut down. This

resistance led the Foundation, in cooperation with the Russian-speaking community, to close the local association to protect people living in Russia.

Focus: The blocking of Wikipedia in Türkiye (2017-2020)

For nearly three years (991 days exactly), access to all language versions of Wikipedia was completely blocked in Turkey, marking one of the longest and most extensive periods of censorship for the encyclopedia.

1. Reasons for the blockage (April 2017)

On April 29, 2017, the Turkish Information and Communication Technologies Authority (BTK) ordered the immediate blocking of the site.

- The official reason: The government criticized Wikipedia for not removing two articles (in English) claiming that Turkey supported or collaborated with extremist and terrorist organizations in Syria.
- Accusations of a "Smear Campaign": Ankara accused Wikipedia of becoming a source of information participating in an international campaign to tarnish Turkey's image.
- The technical impossibility of filtering: Because Wikipedia uses the secure HTTPS protocol, Turkish authorities could not block only the offending pages. They therefore chose to block the entire domain.wikipedia.org.

2. The resistance of the Wikimedia Foundation

True to its principles of neutrality and independence, the Foundation refused to give in to the demands of the Turkish government.

- Refusal of censorship: The WMF argued that the articles in question were based on reliable secondary sources and that their removal would constitute a violation of the encyclopedia's editorial integrity.
- The legal battle: The Foundation immediately challenged the decision in Turkish courts, then took the case to the Constitutional Court of Türkiye and the European Court of Human Rights (ECHR) in 2019.

3. The unblocking and victory of the law (January 2020)

On December 26, 2019, the Turkish Constitutional Court issued a landmark ruling:

- Violation of freedom of expression: The Court ruled that the blocking was unconstitutional and constituted a violation of the freedom of expression guaranteed by Article 26 of the Turkish Constitution.
- Restoration of access: The blockade was technically lifted on January 15, 2020. This decision was hailed as a major victory for the right to information and academic freedom in the country.

The Turkish example demonstrates several important aspects. The Foundation's role as a shield was paramount because, without a central organization willing to bear international legal costs and risk complete invisibility in a country, the articles would likely have been censored by contributors worried about the potential personal repercussions. When, in 2013, the French DGSI arrested the president of Wikimedia France to force him to delete a Wikipedia article, he complied under pressure, but the ability of Quebec Wikipedians to republish the article prevented censorship. Despite the blockade, Turkish contributors continued to edit Wikipedia via VPNs and mirror websites, proving that the community can survive the shutdown of its official infrastructure. Finally, this victory strengthened Wikipedia's legitimacy in the face of other regimes that use the pretext of "fighting terrorism" to censor inconvenient historical or political facts.

B. The creation of competing "State Wikipedias"

To counter the influence of Wikipedia, some states are developing their own controlled projects:

- Ruwiki (Russia) **:** A Wikipedia clone purged of all criticism of the government.
- Baidu Baike (Chine)**:** An encyclopedia subjected to strict algorithmic and ideological censorship. These projects aim to isolate populations from global knowledge.

1. China and Russia

These two countries illustrate the strategy of controlling the information space where Wikipedia is perceived as a tool of Western influence.

- China (Total blockade since 2019) **:**
    - The mechanism**:** After intermittently blocking the Chinese version, Beijing has generalized the blocking of all languages in 2019, just before the 30th anniversary of the Tiananmen Square events.

    - The imposed alternative: China has invested massively in Baidu Baike, an encyclopedia where every edit is approved by state censors before publication. It is the absolute antithesis of the Wikipedia model.
    - Despite this, the Mandarin Wikipedia continues to exist and function internationally. The regime continues to monitor this language version and intervenes through agents in other language versions. In 2021, a group pressuring the Mandarin community was uncovered and banned by the Wikimedia Foundation.[7]
    -
- Russia (Record threats and fines):
    - Information warfare: Since the invasion of Ukraine in 2022, the Russian regulator (Roskomnadzor) has imposed millions of euros in fines on the Wikimedia Foundation for its refusal to remove articles documenting war crimes or using the term "war" instead of "special operation".[8]
    - The "Ruwiki" clone: To prepare for a possible permanent lockdown without depriving the population, Russia launched Ruwiki, a carbon-copy of Wikipedia from which critical Kremlin elements have been purged before switching to an even more restrictive version, Ruviki. This creates a "parallel truth" ideologically opposed to Wikipedia.[9]

2. Democracies: Regulations and Pressures

In democratic countries, the threats are more subtle but just as important for the future of the project.

- **India:**

(2024-2025) :The Indian government threatened to block access to English Wikipedia following an article about a news agency that was officially independent but accused of being a government mouthpiece. After an initial ruling ordering the removal of the article and the disclosure of the contributors' identities—a legal obligation to which the Foundation was bound—the Indian Supreme Court ultimately overturned the first ruling and sided with the Wikimedia Foundation. Concurrently, though it's impossible to definitively say whether the two situations are related, the local organization

[7] https://www.wikimedia.fr/wikipedia-en-chinois-intervention-en-urgence-de-la-fondation-wikimedia/
[8] https://www.bfmtv.com/tech/la-russie-inflige-une-amende-a-wikipedia-pour-des-articles-sur-l-invasion-de-l-ukraine_AV-202303010309.html)
[9] https://www.clubic.com/actualite-525624-apres-avoir-clone-wikipedia-la-russie-la-censure-et-la-remplace-avec-des-articles-de-propagande.html

promoting Wikimedia projects, the Center for Science and Internet, had its funding for foreign projects renewed, cutting it off from the movement's resources and leading to its closure and the unemployment of its staff.

- **Portugal :**

The César de Paço v. Wikimedia Foundation case**:** The Portuguese courts ordered the Wikimedia Foundation to remove information from the Wikipedia article about a Portuguese businessman and politician. All of this information was sourced from reputable national media outlets. The case took a new turn when the court also ordered the Foundation to disclose the personal information of the contributors who had worked on the article in question. The Wikimedia Foundation had to comply with the court order but also decided to appeal to the European Court of Human Rights with the following arguments:

- The defense of anonymity : This is the most sensitive point. For the first time in such a direct manner, a European court has forced Wikimedia to hand over the IP addresses and identifying data of eight volunteer editors. The Wikimedia movement maintains that anonymity is a prerequisite *sine qua non* of freedom of expression on the Web. If a contributor fears personal prosecution for documenting sourced facts about a politician, they will stop contributing. The Wikimedia movement denounces this as a violation of the right to privacy and freedom of association.
- The fight against "SLAPPs" (Strategic Lawsuits Against Public Participation): The Foundation presents this case as the perfect example of a SLAPP (*Strategic Lawsuit Against Public Participation*) because César do Paço uses his financial power to intimidate the foundation and volunteers in order to erase information of public interest (including his links with a far-right party). The aim is for the ECHR to rule that the right to be forgotten cannot be used by public figures to rewrite their own political or legal history.
- The Right to Information: The Foundation emphasizes that the suppressed information (such as the €10,000 donation to the Chega party or his previous indictment) was based on investigations by reputable media outlets. By ordering its suppression, the Portuguese justice system deprives the public of access to verifiable information essential for democratic debate.

4. Threats to freedom of expression: The American context (Trump Administration)

On the other side of the Atlantic, Donald Trump's inauguration in January 2025 opened a period of major uncertainty for the San Francisco-based Wikimedia Foundation.

- Legal insecurity : The Trump administration's threats to reform or repeal Section 230 pose an existential risk. This law protects Wikipedia from legal liability for content written by its volunteers. Without it, the free contribution model could become legally and financially unsustainable. Several people close to the Trump administration have already brandished this threat against the Wikimedia Foundation. Others have also questioned the Foundation's tax-exempt status in the United States as a non-profit organization.
- Pressures on moderation: Political discourse denouncing alleged "censorship" of conservative opinions on online platforms sometimes openly targets Wikipedia. The Wikimedia Foundation is attempting to respond to these various attacks and pressures. Conservative lobbies, close to the Trump administration, have openly made their fight against Wikipedia and its contributors a public objective.
- The example of Texas/Florida: Some states have gone further with dLocal laws aimed at preventing platforms from "censoring" political (often conservative) opinions, which thus seek to force Wikipedia to accept unsourced or conspiratorial viewpoints in the name of "freedom of expression," would directly conflict with the principle ofNeutrality of Viewpoint (NVP).
- Anti-Wikipedias: The United States is also seeing a proliferation of anti-Wikipedia projects; the most recent example is Elon Musk's launch of Grokipedia, an encyclopedic project powered by its AI, Grok, whose answers on certain topics have shown a pronounced ideological bias. Other projects also exist, such as Conservapedia and Libertapedia, which aim to develop politically oriented encyclopedias and further contribute to the fragmentation of the digital space.

III) Wikipedia in the European Union - the case of the 2024 European elections

The 2024 European elections served as a laboratory for observing the resilience of community-managed platforms. According to the mapping report D2.3 (DEM Debate Project) published in June 2025, Wikipedia has demonstrated a superior capacity for self-regulation compared to centralized platforms in the face of election disinformation.[10]

1. Systemic resilience to disinformation

The report highlights that Wikipedia functioned as an effective "filter" throughout the election period. The report's main findings point to three key success factors:

- Rejecting unreliable sources: Most of the disinformation narratives during the European elections circulated on social media. However, Wikipedia's policies consider these platforms unreliable sources by default. This barrier prevented the importation of conspiracy theories or fake news into articles.
- The absence of AI-related incidents: Despite widespread concerns, the D2.3 report confirms that no massive use of AI for disinformation was identified on Wikipedia during the vote. The human control structure discouraged attempts at pollution with synthetic content.
- The speed of community response: On the biographies of the lead candidates and articles dealing with EU issues, vandalism or partisan bias was removed in record time (often in minutes), making the manipulation effort costly and ineffective for the attackers.

2. Analysis of specific mitigation measures

[10] https://wikimedia.brussels/wp-content/uploads/2025/09/D2.3-Mapping-Report-Mapping-Wikipedia-policies-and-risk-mitigation-measures-on-election-disinformation.docx-1.pdf

The Wikimedia movement has deployed specific technical and human measures, detailed in the Wikimedia Europe audit:

A. Proactive protection of sensitive pages

To limit the "attack surface," some communities have widely adopted the use of extended semi-protection. This measure, which restricts publishing to accounts with at least 500 contributions and 90 days of activity, was applied to articles by major candidates. This helped eliminate "dormant accounts" or accounts hastily created by lobbying groups. Each community made decisions based on its specific context and ability to react.

B. The moratorium on the publication of the results

The report mentions a critical rule applied during election week:the absolute ban on publishing results before a 12-hour period after the last polling stations closed in Europe. This waiting period prevented Wikipedia from becoming a source of confusion by relaying exit polls or partial, unconfirmed results.

3. Wikipedia and the Digital Services Act (DSA)

As a "Very Large Online Platform" (VLOP), the Wikimedia Foundation has had to meet unprecedented transparency obligations.TheReport D2.3 analyzes this transition:

- **Risk audit:**Unlike advertising platforms, the major risk identified for Wikipedia is not algorithmic amplification (since there is none), but harassment of moderators.
- **Community Governance vs. Top-Down Regulation:**The case study shows that European regulation benefits from relying on peer moderation structures rather than imposing centralized moderation models, which are often less responsive to the local and cultural nuances of the 24 languages of the Union.

4. Conclusion of the case study

The results of the 2024 European elections confirm that Wikipedia is a**"island of stability"**The report concludes that the absence of monetization of attention and the requirement for high-quality secondary sources (reference press, official reports) constitute the best defenses against attempts at foreign interference and manipulation of information.

IV) Securing the future. The action of Wikimedia France to strengthen the movement.

The resilience demonstrated during the 2024 European elections should not obscure the structural challenges mentioned above and those yet to come. To protect the long-term integrity of Wikipedia, Wikimedia France is implementing a strategy built around three pillars: technical innovation, community growth, and institutional partnerships.

1. The technical aspect: Equipping the community to face new threats

With an encyclopedic project encompassing nearly 2.7 million articles for the French Wikipedia and a core community of over a thousand volunteers, **It is humanly impossible to place this monitoring task solely on the shoulders of volunteers.** Wikimedia France therefore supports the development of decision support tools for patrol officers.

- Synthetic content detection: Integration of AI-generated text detection tools into patrol interfaces to help identify automated additions.

- The 2nd European Narrative Observatory (PROMPT Project): This ambitious project aims to integrate reliability indicators directly into the contribution interface. The goal is to allow moderators to instantly verify the quality of a cited source and detect "reinformation" or propaganda sites before they permanently pollute an article. It's a powerful tool for supporting the community.

  To support the 2nd European Narrative Observatory, Wikimedia France, with support from the Foundation; and opsci.ai, have developed the Wikipedia Sensitivity Meter. Inspired by the need for proactive monitoring, this tool allows for real-time measurement of an article's vulnerability. By cross-referencing data such as the volume of unusual edits, contributor profiles (recent accounts), and topic (political or societal issues), it generates a "sensitivity score" that alerts monitors to a potential coordinated attack. I

The Wikipedia Sensitivity Meter and Barometer

This tool comprises three main components, each targeting a specific risk category:

HEAT RISK

Objective: Assess the intensity of activity and the sensitivity of the page. This category measures signals of abnormal or sudden activity, often precursors to controversy or manipulation. For example:

- View or edit spikes: A sudden surge in page views or edits may indicate an attempt at manipulation or a sudden media interest.
- Revert probability: Using models like ORES, the tool evaluates whether recent edits are likely to be reverted—a potential sign of vandalism or editorial disagreement.

QUALITY RISK

Objective: Evaluate the reliability and robustness of the content. This category focuses on the intrinsic quality of the article, identifying structural weaknesses or potential biases. For example:

- Suspicious or missing sources: The presence of blacklisted references or a high rate of unsourced citations can undermine the article's credibility.
- Source concentration: Over-reliance on a single source (e.g., 80% of references from one media outlet) may introduce editorial bias.

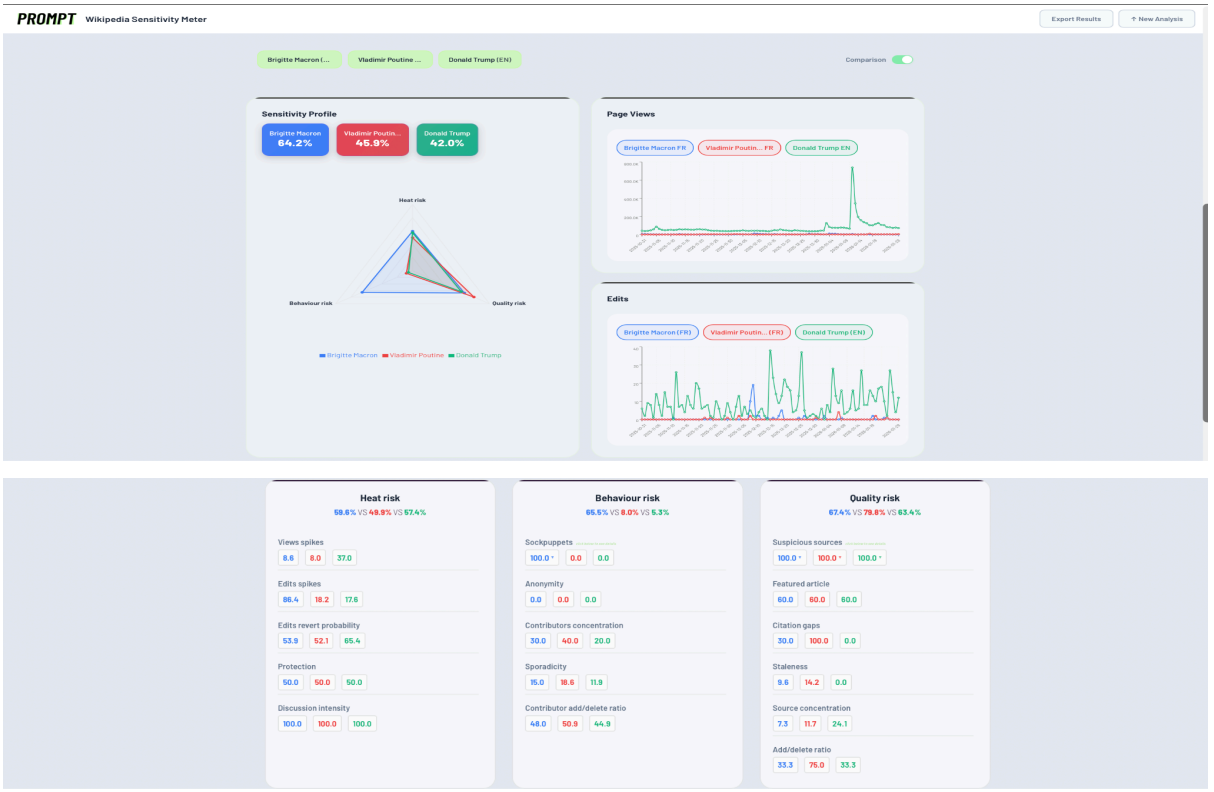
BEHAVIORAL RISK

Objective: Detect manipulation through contributor analysis. This category examines risky editorial behaviors, often linked to coordinated manipulation attempts. For example:

- Anonymous accounts or sockpuppets: An abnormally high proportion of unidentified contributors or fake accounts may signal an organized campaign.
- Contribution concentration: If a small group of contributors dominates edits, it may indicate an attempt to control the narrative.

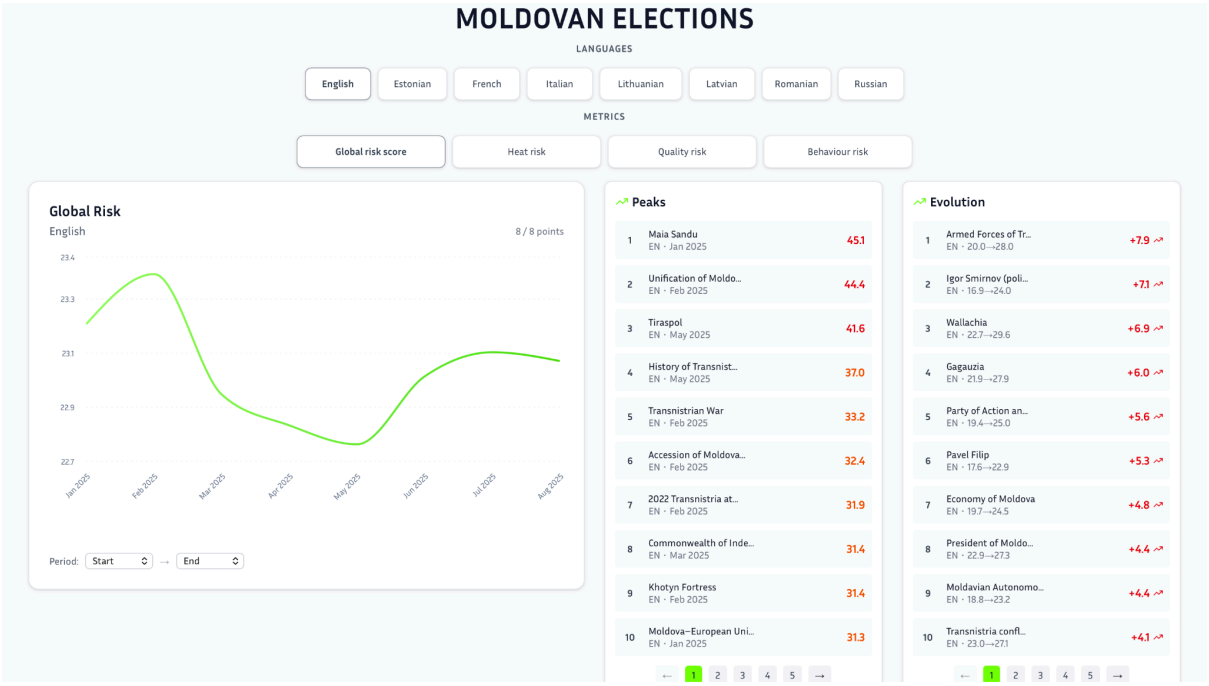
Example: detailed metrics result for the page: https://fr.wikipedia.org/wiki/France

The tool also helps compare sensitivity across multiple pages (up to 5 in 45 languages):



Example: detailed metrics result for the pages 'Brigitte Macron', 'Vladimir Putin' and 'Donald Trump' on French Wikipedia.

The tool can also be easily scaled within the framework of the **Wikipedia Sensitivity Barometer**, which allows for the monitoring of a very large volume of Wikipedia pages on a given topic (a national or local election, a social issue, etc.).



Example: the Wikipedia Sensitivity Barometer designed for the 2025 Moldovan parliamentary elections

2. Strengthening the community: The bulwark against "Project Capture"

The example of the Croatian Wikipedia has proven that a small, isolated community is vulnerable to infiltration. Wikimedia France's response is twofold:

- Diversification and recruitment: To prevent a small group from monopolizing power, the association is increasing its workshops on contributing and moderating. Bringing in new profiles (academics, experts, engaged citizens) guarantees pluralism, which is the best antidote to ideological manipulation.
- Moderation training: It is no longer enough to know how to write an article; one must know how to detect manipulation. Specific programs are being implemented to train volunteers in the techniques of*fact-checking*and the use of protective tools (semi-protection, blocking).

3. Partnership strategy and the fight against disinformation

Wikimedia France acts as an essential "local intermediary" between national authorities, the Foundation and volunteers.

- **Partnership with Les Surligneurs:** As members of the PROMPT consortium, the Surligneurs collaborate with the association to contribute their expertise in *legal checking.* This partnership will result in the implementation of an internal newsletter dedicated to the community, deciphering current disinformation and newly identified manipulative narratives.
- **Collaboration with VIGINUM:** The national agency for vigilance and protection against foreign digital interference collaborates with Wikimedia France to identify hybrid threats. This synergy allows them to alert communities about state-sponsored manipulation campaigns detected at the national level.
- **Focus on the 2026 Municipal Elections:** As the election approaches, Wikimedia France is strengthening its pivotal role. The association is preparing rapid liaison mechanisms between the authorities and the Wikimedia Foundation to respond to attempts at local interference, while also training local contributors to monitor the Regional Daily Press (PQR), a vital source for sourcing municipal articles.

4. A hybrid resilience model

Wikipedia's future depends on this balance between global openness and national roots. Thanks to tools like the *Wikipedia Sensitivity Meter* and through strategic partnerships, the Wikimedia movement is no longer only reacting to attacks: it anticipates risks to ensure that the encyclopedia remains this healthy digital public space serving quality information.

Conclusion: Wikipedia, a pillar of the European informational autonomy

Wikipedia's fight against information manipulation constitutes a comprehensive and organic system, a true model of resilience in the face of hybrid threats which target the democratic processes of the Union.

From the technical rigor of the protection tools to the ethical strength of the Universal Code of Conduct, and the proven resilience during the 2024 European Elections, Wikipedia proves that the peer governance model is one of the strongest in the face of hybrid threats.

The future, marked by important elections in Europe and the omnipresence of AI, demands an ambitious legislative response. The role of the Wikimedia movement, through tools such as those developed for the European Narrative Observatory (*PROMPT).* This is crucial, but it must be supported by a political vision that enshrines free knowledge as a critical infrastructure of our sovereignty. The role of national organizations like Wikimedia France is pivotal here: by training new contributors, developing tools like the *Sensitivity Meter*, and forging strategic partnerships with experts like Viginum or Les Surligneurs, the association ensures that Wikipedia will remain this "common good" where 73% of French people come to find factual truth.

Free knowledge is a permanent conquest; it rests on the vigilance of each individual and the solidarity of a united global movement for the neutrality and reliability of knowledge.

Arcep's report on generative AI[11]highlights critical challenges for digital commons:

- Disintermediation and value capture: AI agents (chatbots) are making extensive use of Wikipedia data to train themselves and respond to users without always linking back to the encyclopedia. According to Arcep, this risks drying up traffic to the original site and, by extension, reducing the number of new volunteer contributors.

[11]https://www.arcep.fr/uploads/tx_gspublication/report-generative-AI-challenges-open-internet-january2026.pdf

- Risk of homogenization: AI tends to smooth out nuances. However, Wikipedia's strength lies in presenting a plurality of viewpoints. The proliferation of AI-generated texts could weaken this editorial diversity.
- A plea for transparency: Wikimedia France, alongside other open internet actors, is campaigning for AI models to be forced to be fully transparent about their sources and to respect copyright, in order to preserve the virtuous circle of human contribution.

From these lessons learnt, several recommendations should help policy-makers, activists, and citizens in shaping a better future for information integrity on Wikipedia.

Recommendations: a protective framework for the Digital Commons

1. The "Wikipedia Test": A prerequisite for any new regulation

The European Commission and the Member States must adopt the**"Wikipedia Test"** [12]as a compass for assessing the impact of any future digital legislation.

- The issue: to ensure that no regulation (moderation, copyright, data) inadvertently prevents the emergence or survival of decentralized and non-profit projects.
- Application: Each text must demonstrate that it does not exclusively favour centralised models based on opaque algorithms at the expense of community governance.

2. Implement an ambitious "Digital Knowledge Act"

Based on the proposals of **Common,**[13] the Commission must remove legal barriers that hinder the free flow of knowledge.

- Protecting the Public Domain: To ensure that a work in the public domain remains so after digitization (Article 14 of Directive 2019/790) and to prohibit any appropriation by neighboring rights.
- Freeing up research: Clarify the exceptions for text and data mining (TDM) so that they primarily benefit open knowledge and not solely the capture of commercial value by proprietary AI.
- Mandatory attribution for AI: To constrain AI producers, through the application of theIA Act, to total transparency about their sources and to clear attribution to Wikipedia to maintain the link between the user and the knowledge infrastructure.

3. Protection of contributing citizens: A robust Anti-SLAPP transposition

Although the transposition of the EU Directive 2024/1069 relies on Member States, the Commission must play its role as guardian of the treaties:

[12] https://wikimediafoundation.org/news/2025/06/27/the-wikipedia-test/
[13] https://communia-association.org/publication/digital-knowledge-act-for-europe/

- Appeal to States: The Commission must encourage Member States to adopt ambitious transpositions that explicitly protect contributors to the digital commons.
- Status of “knowledge defender”**:** These citizens participate in public debate and the general interest; they must benefit from mechanisms for the early rejection of abusive complaints (SLAPP suits) so that they are no longer intimidated by economic or political actors.

4. Funding and Sovereignty: The EDIC Digital Commons

Europe has equipped itself with a powerful instrument: the **EDIC (European Digital Infrastructure Consortium) on Digital Commons**.

- **Political ambition:** The Commission must make this EDIC the driving force behind a genuine European public digital infrastructure.
- **Equity of funds:** It is imperative that funding is not directed solely towards industrial consortia, but actively supports the non-profit structures and the volunteer communities that maintain sovereign databases (such as Wikidata or Wikipedia).

5. Safeguarding the Digital Public Space

- Discoverability: Use theDigital Markets Act (DMA) to monitor the "Gatekeepers" and ensure that digital commons are not degraded in search results in favor of sponsored content or in-house AI responses.
- Pursue a full and ambitious application of the DSA, particularly against platforms that do not comply with European rules.