Disinformation, Wikimedia and Alternative Content Moderation Models: Possibilities and Challenges

Abstract

In this research we propose to look at how three different Wikipedia communities deal with the challenge of identifying trustworthy sources in the context of editing entries on topics crossed by dynamics of political polarization in three countries of Latin America. We seek to shed light on how Wikipedia's community-led moderation model, as well as the broader normative commitments of the community itself, and global moderation standards and technology deployed across the platform, work for Wikipedians dealing with such a pressing and challenging task. The findings could have a broader impact on global debates on disinformation and content-moderation challenges.

Introduction

This research proposal seeks to answer a narrow question: how do Wikipedians identify trustworthy sources when discussing controversial Wikipedia articles? This question is connected to a fundamental challenge to our broader information ecosystem, related to the effects of political polarization on the status of shared and fact-based knowledge, a prerequisite for meaningful democratic deliberation. In politically polarized environments, drafting Wikipedia entries on controversial topics must be challenging. Wikipedia editors are themselves subjected to the dynamics of polarization that affects their larger communities. Our hypothesis is that the rules, principles, procedures, and normative commitments of the Wikipedia community helps members break those dynamics and reach some degree of consensus on identifying what a trustworthy source is and which one is not. If so, these findings could help us identify better practices and to better understand ways of combating disinformation in polarized environments.

Start time: July 1, 2023.
End time: June 30, 2024.

Related work

Much has been written about the "epistemic crisis" currently threatening democracies around the globe (Benkler, Faris, and Roberts 2018, chap. 1). A common proposed solution is to highlight and promote trustworthy sources of information. Still, what and how trustworthiness is attributed has been understudied. Wikimedia usually describes itself as a "horizontal" platform within content moderation policy discussions. The community-led moderation model championed by Wikipedia competes with the algorithmic and automated models promoted by social media platforms such as Facebook and Twitter (Caplan 2021, 174) and how problems should be addressed. Both models are ultimately based on different conceptions of what the Internet is and should be (Lessig 2006, chap. 6; McDowell and Vetter 2020). Understanding how Wikipedia deals with issues such as disinformation around polarized issues may produce relevant insights into the strengths and challenges of a model

that is based on a robust conception of community, as opposed to the others', based on advertising, scalability, and lucre.

Our approach is based on previous research on Wikipedia's governance (Lovink 2011; de Laat 2012; Grimmelmann 2015a; R. Stuart Geiger and Halfaker 2016; Caplan 2021; Rijshouwer 2019; McDowell and Vetter 2020; Konieczny 2009). This literature has considered Wikipedia's democratic ethos a *rara avis* in the current Internet ecosystem (de Laat 2012, 124). According to Paul de Laat, in Wikipedia "two equally consistent visions are pitted against each other: Wikipedia as a reliable, encyclopedic institution on the hand, and Wikipedia as a solidary community on the other" (de Laat 2012, 125). These values guide members of the community and inspire the rules and procedures developed to moderate content. These two values of "distributed organization and strong social norms" are the basic commitments of the Wikipedia community (Grimmelmann 2015b, 80). While automatic systems have been deployed to deal with moderation challenges specially related to conduct (R. Stuart Geiger and Halfaker 2016; Caplan 2021, 178) but also to help users combat disinformation (Saez-Trumper 2019), at the end of the day the combination of individual judgements, deliberation between editors and contributors, and hierarchical decision-making in a context of strong central organizing structures (Rijshouwer 2019, 226–33) are the ways in which controversies are solved (even though some of them remain open and users are informed of that circumstance).

Emiel Rijshouwer's study of Wikipedia is at the core of our theoretical approach. We find his theory of "self-organizing bureaucratization" a compelling explanation of Wikipedia's strength, sustainability, and democratic ethos produced by the ongoing tension between committed individuals and bureaucratic structures. Indeed, for Rijshouwer bureaucratization is "something that is not always designed and strictly imposed, effectuated from the top-down, but rather is a way of democratically coping in a complex environment" (Rijshouwer 2019, 236). The core of his argument is captured in the following excerpt.

> "Weber, as many Wikipedians with him, would argue that once there is a comprehensive disciplinary structure consisting of clearly defined goals, well-established and well-documented norms, well-thought out procedures, and objective criteria and while participants commit themselves to use rational arguments, to follow the lead of those with formal and informal positions and clear mandates, and to publicly document and to account for their work, it would be much more efficient to meet certain challenges and to achieve certain objectives in a complex environment, based on the commitment of self-selected participants, then when matters would be organized 'from case to case'. In such an open and transparent and at the same time bureaucratic organization, self-selected volunteers could easily join a project based on their personal motivations, objectives and concerns, without established participants having to step in in each and every case" (Rijshouwer 2019, 234).

Wikipedians commitment to core values—what Rijshouwer calls their "common belief in rationalism and objectivism" (Rijshouwer 2019, 236)—is a likely candidate to be the mechanism through which Wikipedians get out of the polarized environment their larger selves necessarily inhabit, at least when they engage in editing Wikipedia's content. The commitment to core values, including perhaps in a central position the neutral point of view (NPOV), may be the cause to the effect we expect to find in our research. If we manage indeed to produce a positive answer to our research question (Wikipedians *do* manage to find trust-worthy sources and reach consensus around them)

several additional questions emerge as useful to better understand what drives such success: how do the hierarchies within the community operate to settle disputes, how new editors are socialized in the community's values, and so on.

Our proposal seeks to contribute to Wikipedia scholarship by producing three new and narrow case studies that focus on important events in Latin America that were subjected to the all-too-familiar logic of polarization. Our hypothesis is that certain articles on controversial topics might be the object of heated debates between Wikipedians who edit them. Wikipedians, while members of a community committed to certain values and rules, are also individuals who are members of the broader political demos. They may take part in social and political organizations and they may hold certain political sympathies. In the context of a polarized society, it is likely that some of them will hold views strongly, and will suffer the effects of polarization themselves (Leeper 2014). In fact, it is reasonable to expect that they will hold strong views on certain topics and show high levels of rejection for counter arguments or counter narratives.

While polarization has many causes, one of the usual suspects is the balkanization of mass media and—more recently—the Internet and the opportunity it creates for people to easily trap themselves in eco-chambers and filter-bubbles (Sunstein 2017; Pariser 2011). Although this hypothetical causal explanation has not been proven (and in fact, several studies suggest that these mechanisms are not the cause of polarization (Boxell, Gentzkow, and Shapiro 2017; Bakshy, Messing, and Adamic 2015)) it has garnered some traction and is still useful for our proposed research. By looking at Wikipedia editing decisions within its community-led model, but bounded by rules, principles, and procedures, we can lead our inquiry within a space that by design cuts across the trends that allegedly push people into eco-chambers. If people in their daily lives choose media diets that align with their ideological priors, their involvement in the Wikipedia community forces them to engage in a dialogue that, in controversial issues and on polarized contexts, is necessarily bi-directional.

This makes answering the research question on trustworthy sources especially interesting in the context we propose of polarized communities and around controversial topics. Our intuition is that we should see not only how Wikipedians identify those sources, constrained by Wikipedia rules and other normative commitments of the community, but also how they discuss and go about their disagreements. It is likely that Wikipedia's rules and processes of engagement and deliberation forces individuals who hold strong and opposite views on a range of topics to reach some degree of consensus on the sources that can be used to write and source a Wikipedia article. If so, Wikipedia's community-led model may offer an insight into a core issue within the disinformation dilemma: how are sources ranked and vested with authoritative power.

Our choice of case-studies is specially relevant to the objective of our research. To recall, we want to study the following Wikipedia entries:

1. The 2019 political crisis in Bolivia (link).
2. The death of prosecutor Alberto Nisman in Argentina (embedded in his biographical entry) (link).
3. The portuguese entry on the Operação Lava Jato, in Brazil (link).

The three articles fulfill several conditions that are useful for our inquiry, and meet some of the prerequisites we consider necessary to produce good case-studies. First, all three societies (Brazilian, Bolivian, and Argentinean) are crossed by deep and pervasive political polarization. Hence, it is likely that we will find

polarized individuals within the local Wikipedia communities. Second, the three articles cut to the core of the kinds of deep and partisan disagreements that feed polarizing dynamics in the political process. While in Bolivia for some the 2019 outing of president Evo Morales was the outcome of a popular uprising against his attempt to circumvent constitutional limits to reelections, for others it was a *coup d' etat*. In Argentina, prosecutor Alberto Nisman killed himself because his accusation against former president Cristina Fernández was made up; for others, he was murdered precisely because he was moving forward with his investigation. Finally and similarly, in Brazil some see the Operação Lava Jato as one of the most important judicial inquiries into political and corporate corruption, while others see it as a paradigmatic case of the use of courts to persecute popular political leaders. Third, in all the articles discussions between editors took place. In the case of the entry on Alberto Nisman, the disagreement as to the extent to which the neutral POV policy has been respected remains and a relevant label informs users of that disagreement.

The three articles, then, provide a useful occasion to see how Wikipedia editors deal with sorting trustworthy vs untrustworthy sources in a context of political polarization.

Methods

Within the scope of this research, the narrow focus we selected will increase the likelihood of success of the research method we propose: trace ethnography (Rijshouwer 2019, 40; R. Stuart Geiger and Ribes 2011). Through this method we will look at how these articles evolved and the kind of discussions that they generated among Wikipedians who proposed editions. Furthermore, we want to understand how editors understand their role, how they fight the polarization that affects their communities, and what strategies they develop to deal with the tensions that naturally ensue.

This calls for expanding our understanding of this methodology. Stuart Geiger and David Ribes define trace ethnography as a form of institutional ethnography that takes advantage of the rich documentation produced in "highly technologically-mediated systems" (R. Stuart Geiger and Ribes 2011, 1).

> "Analysis of these detailed and heterogeneous data—which include transaction logs, version histories, institutional records, conversation transcripts, and source code—can provide rich qualitative insight into the interactions of users, allowing us to retroactively reconstruct specific actions at a fine level of granularity. Once decoded, sets of such documentary traces can then be assembled into rich narratives of interaction, allowing researchers to carefully follow coordination practices, information flows, situated routines, and other social and organizational phenomena across a variety of scales" (R. Stuart Geiger and Ribes 2011, 1).

In that sense, our own research proposal will jump into those documents, specially the 'Discussion' tab of the three entries we have identified, as a first object of analysis. We expect to follow the most interesting discussions from the point of view of our research question in detail, and interview editors and contributors who took part in them. (We depend to some extent on the collaboration of local Wikipedia communities in that effort and have already contacted some of them). Towards the final stage of our data gathering efforts, we will likely interview senior members of Wikipedia's bureaucracy as well as Wikimedia Foundation employees supporting these efforts from different sectors (engineering, policy, legal, trust and safety, etc).

Expected output

Outputs to our research will include mainly an academic paper, to be subjected to a peer-review process in order to strengthen its relevance and contribution to the broader literature we have discussed; as well as derivatives of that research paper capable of contribution to relevant policy discussions on disinformation and content moderation. The main audience for this output is other researchers, broadly within the fields of disinformation, content moderation, and polarization. While decisions on where to submit are difficult to predict, fast-track venues such as the *Misinformation Review* or the *Journal of Trust and Safety* are two likely candidates for publishing the final output.

We will also produce derivatives from our research, in order to reach broader audiences through fact-sheets, content designed to disseminate findings in social media, and so on. The goal is to engage the broader Internet governance community, for we feel that Wikipedia's community moderation model is a good alternative to the models followed in other platforms. In that sense, we specifically expect to disseminate our research:

1. In the Global Network Initiative, of which CELE is a member as an academic institution;
2. Within the Latin American digital rights movement, through CELE's membership in the AlSur coalition and through CELE's annual regional workshop;
3. In international conferences. In particular, and because of the timeframe of the project, we aim to present our findings at RightsCon 2024.

Risks

Our case selection has some advantages but also poses some risks and presents some limits.

The main advantage is that these articles cover three important Latin American countries, and thus offer the opportunity for comparative research within a region with many connections across borders. The selection also makes it more likely to find aids with enough knowledge of local context in all three countries, something we believe will strengthen the process of inquiry.

A risk that can be derived from some of the comments received would be that our findings could not be extrapolated to different contexts . However, the generalization of findings can be seen from a different perspective. We will study how three different communities identify trustworthy sources in the context previously described. If these processes are guided by organic procedures and commitments to the normative values of the Wikimedia movement, then we might gain a better understanding of how sources of information are organized under a hierarchical basis, how members of the community process their disagreements and how internal procedures and rules help (or not) to reach that outcome. These findings could produce a set of best practices that might be useful in other contexts.

Community impact plan

Our research will be useful for two general audiences beyond researchers and academics.

In-depth study of how Wikipedians deal with the challenge posed by our research question will be useful for the Wikipedia community at large. Other editors and contributors, also working in contexts of polarized societies, might benefit from seeing how the rules, procedures, and normative commitments of the movement work in practice. As we mentioned in our Stage I

submission, we see a good fit between our proposal and Wikimedia 2030 Movement Strategy Recommendations, in particular in relation to the goals of (a) improving user experience; (b) to identify topics of impact and (c) to ensure equity in decision-making. We plan to work with local Users Groups to find the best ways in which our findings can reach the broader Wikipedia community, specially in Latin America.

On the other hand, we hope that our findings will inform ongoing internet governance debates over content moderation and disinformation. Wikipedia`` community-led approach is a model that competes with other approaches (Caplan 2021). Hence, the findings of this research may contribute to inform policy makers and activists working on potential solutions.

Evaluation

We believe the research project should be evaluated according to two main contributions:

1. Its ability to influence the ongoing debate on content moderation in different Interent governance settings. This should measure the impact of the paper, and could use as indicators (a) the amount of other scientific and policy papers that cite it and (b) the amount of activities in which the paper and its findings are actively engaged, both within the broader field of Internet governance and within the Wikimedia community. This impact should be assessed around the narrow field of disinformation studies, but could also influence broader Internet and technological governance debates.

2. Its ability to influence Wikimedia's internal discussion on the best ways to fulfill its mission and realize the community's normative commitments. In that sense, we believe---as we mentioned before---that the research could influence the Wikimedia 2030 Movement Strategy Recommendations.

Budget

Our budget is 40,194 US dollars.

CELE has extensive experience in managing this kind of research and working with different stakeholders to position the research within the broader internet governance ecosystem. To this end, we have permanent staff within the organization and given the regional and international nature of our work, we have an extensive network of people who work with us part time as consultants.

While the budget includes some direct costs associated with dissemination and distribution, the comparative advantages of having this research done by CELE and not an independent researcher is CELE`s extensive networks, locally, regionally and internationally. CELE`s research projects contribute towards our vision and strategic priorities, papers dialogue with each other, and enjoy wide distribution through our website, events that we organize and attend, workshops we lead, etc.

The budget includes a small percentage of direction salaries and a percentage of communications and admin salaries. Additionally, it includes a lead specialist on disinformation and the media ecosystem to serve as the principal researcher to this project. Finally it contemplates hiring two junior researchers, one for each language to be included in the research, for language but also for cultural and contextual background and feedback. Other direct costs include the edition, translation, and correction services (and some funds reserved for printing a small batch of papers for dissemination in certain fora in which we consider this useful). Other publication costs are covered by the University (design, website, etc) We also included a 10 per

cent overhead fee (which we include in all our projects) and fees to transfer funds outside of Argentina to foreign consultants: around 35% of the total funds to be transferred.

While there are alternatives to produce this research, but they are not necessarily less costly. For instance, we could strive to find a full-time senior researcher to conduct the research by herself, and do without junior researchers Still, this would make finding the appropriate candidate for sr. researcher much more difficult and the results less inclusive of the local perspective. The project is also demanding in terms of the necessary time to produce it, specially because we expect to conduct interviews. The amount of time needed to simply analyze and study the Discussion tabs of the three articles is rather short, but desk research will be followed by in-depth interviews with selected contributors who These interviews are very time-consuming, and this explains both the (a) length of the project and the (b) necessary (human) resources we feel are needed to successfully fulfill this research effort.

We should say that even though we believe this proposed budget is sound and takes advantage of embedding the research within an institutional setting such as the one that CELE provides (as opposed e.g. to be a solo project by an independent researcher), we are truly interested in the project and are open to discuss cutting costs if necessary.

References

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. 2015. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348 (6239): 1130–32. doi:10.1126/science.aaa1160.

Benkler, Yochai, Rob Faris, and Hal Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. New York, NY: Oxford University Press.

Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro. 2017. "Greater Internet Use Is Not Associated with Faster Growth in Political Polarization Among US Demographic Groups." *Proceedings of the National Academy of Sciences* 114 (40): 10612–17. doi:10.1073/pnas.1706588114.

Caplan, Robyn. 2021. "6. The Artisan and the Decision Factory: The Organizational Dynamics of Private Speech Governance." In *6. The Artisan and the Decision Factory: The Organizational Dynamics of Private Speech Governance*, 167–90. University of Chicago Press. doi:10.7208/9780226748603-007.

Epstein, Diana, and John D Graham. 2007. "Polarized Politics and Policy Consequences." Occassional Paper. Washington D.C.: Rand Corporation.

Geiger, R Stuart, and David Ribes. 2011. "Trace Ethnography: Following Coordination Through Documentary Practices." In *2011 44th Hawaii International Conference on System Sciences*, 1–10. Kauai, HI: IEEE. doi:10.1109/HICSS.2011.455.

Geiger, R. Stuart, and Aaron Halfaker. 2016. "Open Algorithmic Systems: Lessons on Opening the Black Box from Wikipedia." *AoIR Selected Papers of Internet Research*, October. https://journals.uic.edu/ojs/index.php/spir/article/view/8772.

Grimmelmann, James. 2015b. "The Virtues of Moderation." *Yale Journal of Law and Technology* 17: 42–109. https://heinonline.org/HOL/P?h=hein.journals/yjolt17&i=42.

———. 2015a. "The Virtues of Moderation." *Yale Journal of Law and Technology* 17: 42–109. https://heinonline.org/HOL/P?h=hein.journals/yjolt17&i=42.

Konieczny, Piotr. 2009. "Wikipedia: Community or Social Movement?" *Interface: A Journal for and about Social Movements* 1 (2): 212–32.

Laat, Paul B. de. 2012. "Coercion or Empowerment? Moderation of Content in Wikipedia as 'Essentially Contested' Bureaucratic Rules." *Ethics and Information Technology* 14 (2): 123–35. doi:10.1007/s10676-012-9289-7.

Leeper, Thomas J. 2014. “The Informational Basis for Mass Polarization.” *Public Opinion Quarterly* 78 (1): 27–46. doi:10.1093/poq/nft045.

Lessig, Lawrence. 2006. *Code*. Version 2.0. New York: Basic Books.

Lovink, Geert, ed. 2011. *A Wikipedia Reader: Critical Point of View*. INC Reader 7. Amsterdam: Institute of Network Cultures.

McDowell, Zachary J., and Matthew A. Vetter. 2020. “It Takes a Village to Combat a Fake News Army: Wikipedia’s Community and Policies for Information Literacy.” *Social Media + Society* 6 (3). SAGE Publications Ltd: 2056305120937309. doi:10.1177/2056305120937309.

Pariser, Eli. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1118322.

Rijshouwer, Alexander. 2019. “Organizing Democracy: Power Concentration and Self-Organizing Bureaucratization in the Evolution of Wikipedia.” Rotterdam: Erasmus University Rotterdam.

Saez-Trumper, Diego. 2019. “Online Disinformation and the Role of Wikipedia.” arXiv. doi:10.48550/arXiv.1910.12596.

Sunstein, Cass R. 2017. *#Republic: Divided Democracy in the Age of Social Media*. Princeton ; Oxford: Princeton University Press.

Annex

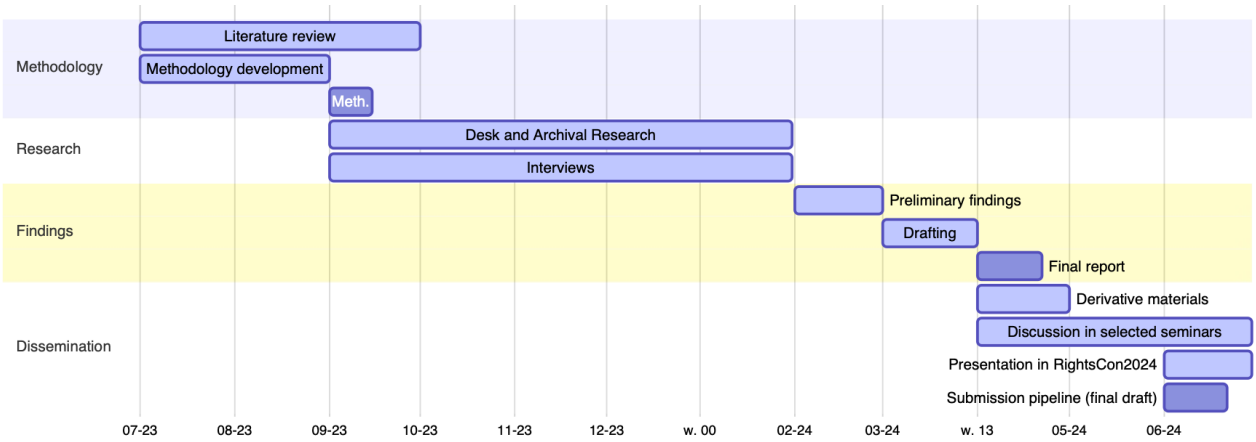

Figure 1: GANNT and proposed timeline