The 18th Annual Conference of the Association of Internet Researchers
Tartu, Estonia / 18-21 October 2017

PLATFORM GOVERNANCE AS REFLEXIVE COORDINATION – MEDIATING NUDITY, HATE SPEECH AND FAKE NEWS ON FACEBOOK

Christian Katzenbach
Kirsten Gollatz
Alexander von Humboldt Institute for Internet and Society

Background

Digital platforms have become dominant players in contemporary societies by positioning themselves as key sites for social communication and transactions (van Dijck, 2013). With platform governance becoming a major concern, scholars have increasingly turned their attention to the ways how platforms perform this intermediary role – most notably to algorithms, practices and policies (Gillespie, 2014; Roberts, 2016; DeNardis & Hackl, 2015). What is still scarce, is (a) research that frames and explains these phenomena with key concepts of social theory, and (b) longitudinal studies that track these developments over time. This paper contributes to the debate theoretically by establishing a concept of platform governance as reflexive coordination based on institutional theory, and empirically by presenting a longitudinal study (2005-2016) that combines the analysis of Facebook's evolving policies and practices on controversial content (nudity, hate speech, fake news) with a policy and discourse analysis.

Theorizing Platform Governance as Reflexive Coordination

Governance is widely used as an umbrella term for all sorts of ordering and regulation processes. Yet, it is a notoriously slippery term that often remains vague and difficult to operationalize, repeatedly conflated with the term *regulation*.

This conceptual weakness is reflected in the literature on platform governance. Governance *of* platforms usually refers to public policy measures that try to steer platforms dynamics towards common good (safe harbor, tax and competition policy); governance *by* platforms addresses companies' own measures (policies, algorithms) to influence behavior on their platform

But this conflation of governance and regulation diminishes the concepts' analytical value. Neither the side effects of actions and processes pursuing non-regulatory goals (such as Facebook optimizing its platform for engagement), nor the role of public

debates and user complaints can be adequately captured with a regulatory perspective. Thus, following recent theoretical contributions we apply a broader concept of governance as *ordering*. Specifically, we propose to understand platform governance as *reflexive coordination* (Hofmann et al., 2016) – integrating diverse modes of ordering (terms of service, public debate, algorithms) but focusing on controversies and those critical moments when routine activities become object of contestation and need to be revised. This enables us to understand platform governance as an encompassing social process that is not only exercised by platforms and regulatory agencies.

Methods

Bringing together a *longitudinal analysis of the platforms evolving policies* and community guidelines with an *analysis of the public discourse* on Facebook's handling of controversial content, we investigate empirically how content rules on platforms evolve as a subject of public conflict and controversy. The analysis of Facebook's terms of service and community guidelines is based on a corpus of 31 documents collected through the Internet Archive's "Wayback Machine" ranging from 2005 to 2016. For the discourse analysis, we used the Dow Jones Factiva news database to extract all english-language articles of major news and business sources containing "facebook", and "terms of service" or "community standards" as well as respective synonyms from 2004 to 2016. Applying thematic filtering and coding tools in order to identify relevant actors, statements and critiques resulted in three separate sets of documents: 257 on nudity, 240 on hate speech, and 29 on fake news.

Results

Despite recent spikes in public attention, our analysis reveals for all three cases a long-term processes, in which the distinction between legitimate and illegitimate content on the platform is negotiated, institutionalized and contested. In the case of *nudity*, a long-standing debate on the appropriateness of breastfeeding pictures constitutes a critical moment: Facebook's routine of deleting photos displaying parts of female breasts regardless of context became contested for the first time in 2007 when more and more families started noticing the ban of breastfeeding pictures. Following public discourse turning the issue back and forth for years, fueled by repeated public outcries and online petitions, Facebook in 2013 introduced an explicit exemption for its content rules (with updated wording 2015) that "photos of women actively engaged in breastfeeding" are always allowed. Since then, our data shows a clear decrease in public debate. Finding adequate criteria for handling nudity on Facebook only became an issue again when Facebook removed an iconic Vietnam War photo ("Napalm Girl") in 2016. This time, Facebook responded quickly to the international outcry, reinstated the photo, and later announced a change to its internal content moderation process allowing "more items that people find newsworthy […] — even if they might otherwise violate our standards."

In the context of *hate speech*, Facebook's policies have become more and more explicit

and resolute over time: from “inappropriate content” that “simply [does] not belong in a community like Facebook” (2007) to constituting its own category in 2011 defined as a form of inappropriate behaviour that the platform “does not tolerate”. In 2015, the hate speech policy grew to 274 words in 13 paragraphs, most notably by adding language to explain and justify procedures. The discourse analysis shows clear spikes at the end of 2015 due to Europe’s refugee situation, and a push from politicians, above all in Germany, to act more diligently. However, Facebook’s official hate speech policy has remained unchanged; instead responses include public engagement campaigns (e.g. promoting counter-speech), and the adaptation of internal enforcement procedures (eg.“migrants” now constitute a “quasi-protected group” in Facebook’s internal content moderation guidelines).

Albeit not termed *fake news,* false or misleading information has been identified as a problematic issue long before the 2016 US-election. Facebook prohibits users for a long time from using the platform “to do anything unlawful, misleading, malicious, or discriminatory”, and Facebook Pages “must not contain false, misleading, fraudulent, or deceptive claims or content.” Not surprisingly, our data shows an uptake of fake news discourse towards the end of 2016, displaying a strong frame that compares Facebook’s standards for handling fake news with the professional standards and regulations of journalism. While (yet) not present in its policies, Facebook has adapted its routines by integrating fake news into the categories that users tick to justify flagging of objectionable content. In the US and Germany, the company is partnering with journalistic organizations to fact-check flagged content. If there is no evidence for the facts presented, it will be flagged as “disputed”, reducing possibilities for further sharing and monetizing this content.

Discussion

Bringing together a new conceptual approach to governance and a longitudinal empirical study we were able to characterize platform governance as an evolving negotiation process. This process oscillates between Facebook’s unilateral provisions, user engagement, public discourse, and public policy measures – jointly and sometimes antagonistically institutionalizing distinctions between legitimate and illegitimate content. Critical moments occur when the parties involved disagree; in the controversies all stakeholders need to legitimize their perspectives and practices – abandoning their every-day routines. Thus, this process is not merely “governance by shock” (Annany & Gillespie, 2016), the changes in policies cannot always directly attributed to public demands. The critical moments highlighted in our empirical study are rather temporary tips of the iceberg that is the long-lasting ongoing negotiation process. Taken together, our analysis shows that attention in this negotiation process has shifted from platform policies to the practices of enforcement.

References

Annany, M., & Gillespie, T. (2016). Public Platforms: Beyond the Cycle of Shocks and Exceptions, Paper presented at The Internet, Policy & Politics Conference, 22-23 September 2016. Oxford.

DeNardis, L., & Hackl, A. (2015). Internet governance by social media platforms. *Telecommunications Policy, 39*(9), 761-770.

Gillespie, T. (2014). The Relevance of Algorithms. In T. Gillespie, P. J. Boczkowski & K. A. Foot (Eds.), *Media technologies: essays on communication, materiality, and society* (pp. 167-194). Cambridge, MA: MIT Press.

Hofmann, J., Katzenbach, C., & Gollatz, K. (2016). Between coordination and regulation: Finding the governance in Internet governance. *New Media & Society,* 1-18. doi:10.1177/1461444816639975

Roberts, S. T. (2016). Commercial Content Moderation: Digital Laborers' Dirty Work. In S. U. Noble, & B. M. Tynes (2016). *The intersectional Internet: race, sex, class and culture online*. New York: Peter Lang.

van Dijck, J. (2013). *The Culture of Connectivity: A Critical History of Social Media*. Oxford ; New York: Oxford University Press.