

Reshaping platform governmentality: The curious case of the moderation strike on the Stack Exchange

Damien Renard^{*a}, Salla-Maaria Laaksonen^b, François Lambotte^a, Jukka Huhtamäki^b and Sophie Del Fa^a

^aInstitut Langage & Communication, Université catholique de Louvain, Louvain-La-Neuve;

^bCentre for Consumer Society Research, University of Helsinki, Helsinki, Finland

^{}damien.renard@uclouvain.be*

Reshaping platform governmentality: The curious case of the moderation strike on the Stack Exchange

Abstract: This research explores the power dynamics between platform owners and online communities, focusing on the Stack Exchange moderation strike against the company's new policy on AI-generated content. Using Foucault's concept of governmentality, we show how the moderation of AI-generated content revealed tensions and resulted in collective action that made platform governmentality visible. In their strike, moderators resisted the policy by repurposing platform features, mobilizing users, and negotiating for greater autonomy. Our analysis indicates that platform governmentality is a combination of the disciplinary and security regimes and that the same mechanisms that support them can also be appropriated for resistance. We show how the strike and subsequent negotiations reshaped the platform's disciplinary and security mechanisms, essentially by reasserting the role of disciplinary mechanisms through new confidential heuristics and revised procedures that also resulted in changing the platform technology. This study also highlights the complex role and power position of volunteer moderators in the organizing processes of platform capitalism.

Keywords: moderators; social media platforms; resistance; power; governmentality

Introduction

On June 4, 2023, moderators of the world's largest Q&A forum, the Stack Exchange (SE), initiated a strike after the platform company prohibited them from removing content generated using generative artificial intelligence (GenAI). The conflict brought to the fore a continued tension between volunteer moderators and platform owners, where the norms of content moderation and the roles of both actor groups were questioned. In this study, we use the strike as our point of entry to interrogate platform governmentality. Governmentality, deemed central to the analysis of power by Michel Foucault (1982), 'consists in guiding the possibility of conduct and putting in order the possible outcome' (p. 789). In this study, we use *platform governmentality* to refer to the rationality and techniques through which social media platforms seek to govern users' conduct. These techniques materialize in such mechanisms as design choices, incentive structures, or the delegating of responsibilities. We follow how these mechanisms were reshaped during the strike.

Previous studies have discussed how social media platforms act as infrastructures that shape communication, power relations, and social action through design, policies, and technological structures like algorithms (Van Dijck et al., 2018; Dolata & Schrape, 2025; Gehl & Zulli, 2023; Gorwa et al., 2020; Bucher, 2012; Duffy & Meisner, 2023). In this study, we theorize the complex power dynamics related to the social, technological, and economic forces that shape platformed organizing. While communities on digital platforms may emerge organically, their maintenance relies on the voluntary labor of content creation and moderation. In digital environments, moderators, often working voluntarily (Matias, 2019; Postigo, 2009), play a key role in curating content and regulating contributors to ensure an adherence to community standards and platform policies (Massanari, 2017). They embody a hybrid role characterized simultaneously by power and subordination: they have strong abilities to govern action on the platform, but this role is afforded by the platform

itself. Thus, they act as mediators of platform power. This role highlights their critical, yet sometimes precarious, position in platform governmentality. In relation to growing tensions in the platform economy, there have been notable moments of collective action and strikes in recent years by community moderators, often as a response to grievances pertaining to platform policies, administrative decisions, and working conditions. For example, moderators on Reddit went on strike in 2015 to defend their working conditions (Matias, 2019).

Despite extensive research into moderation activities on social media platforms, most of the studies focus on the actions of moderators instead of adopting a broader perspective to interrogate the intricate technological and organizational power relations between users, moderators, and platform owners in the context of platform capitalism. In this context, the platforms' business models and operating logics are based around exploiting the data generation efforts and volunteer labor of the users and moderators (Srnicsek, 2017). Moreover, we argue that the capitalistic dynamics related to AI introduce tensions in what we call platform governmentality. First, the speed of GenAI-facilitated content production challenges the current control mechanisms applied by human moderators. Second, the increasing use of GenAI on digital platforms reconfigures power relations by transforming user-generated data into training material for AI and large language models (LLMs). Related conflicts have emerged in connection with data originating from Reddit and Wikipedia. Both companies have engaged in actions to protect their data assets.^{1,2} Third, to survive in the GenAI age, platform companies have proposed new AI-based services³ in a context of declining participation and knowledge production

¹ See <https://www.forbes.com/sites/anishasircar/2025/10/23/would-be-bank-robbers-reddit-escalates-ai-data-wars-with-perplexity-lawsuit/>.

² See <https://wikimediafoundation.org/news/2025/11/10/in-the-ai-era-wikipedia-has-never-been-more-valuable/>.

³ See <https://stackoverflow.blog/2023/07/27/announcing-overflowai/>.

following the launch of ChatGPT (Burtch et al., 2024). Regarding our focus on SE as a case study, we argue that the external trigger of GenAI reshapes platform governmentality and brings forth questions related to the dialectic of power/resistance on digital platforms. Thus, we ask: how did the introduction of GenAI make power relations on the platform visible, and how was platform governmentality shaped as a consequence?

To answer both aspects of this question, we examine the moderation strike on the SE network, which temporarily transformed the platform into a conflicted space. Established in 2009, SE is a collaborative knowledge sharing platform composed of Q&A communities, where volunteer moderators play a central role in structuring, curating, and governing content. Through their moderation work, they exercise authority that is essential to the platform's functioning (Matias, 2019). We adopt a Foucauldian approach to unpack the conflict, including the ways in which moderators attempt to reassert their power as platform workers. By analyzing SE moderators' resistance, this study reveals how mechanisms of governmentality are renegotiated as AI introduces new tensions into platform governmentality.

We contribute to studies on platforms and social media in two ways. First, we theorize platform governmentality and demonstrate that it is shaped by techno-economic logics and operationalized through mechanisms rooted in both disciplinary and security regimes. We further show that while some of these mechanisms appear hardcoded in the platform's design, they remain flexible in times of conflict, enabling resistance to emerge within the very structures of governmentality. Second, in analyzing the strike, we highlight how moderators seek to protect the power of the community over the platform's company-driven economic logic and their own role as intermediaries between the community and the company. In doing so, we show how they participate in reconfiguring power dynamics in the era of platform capitalism and the changing dynamics introduced by AI tools.

Literature review

Online communities and power relations in platform capitalism

Online communities have been a prominent feature of online social action since the early days of the Internet. While seemingly spontaneous, online communities have typically shared values or ideals that bring them together (Kiesler et al., 2012), and social reasons, such as a sense of community, are important in assembling people online (Postigo, 2009). Scholars have suggested that, as organizations, online communities are less bureaucratic and hierarchical and more participatory than formal organizations due to the open and distributed forms of communication afforded by the platform technologies (Dolata & Schrape, 2016; Ashuri & Bar-Ilan, 2016). However, as the communities grow, the complexity and potential costs of managing the communities are becoming more evident and require formalizing practices. For example, Rijshouwer and colleagues (2023) showed how organizing on Wikipedia evolves through power dynamics and struggles, eventually creating a form of bureaucratization within the self-organizing community.

One classical practice of organizing content and communities on social media platforms is moderation, a practice that controls the types of content and behaviors deemed suitable for a given online community or platform. Moderation is a combination of industrial, community-reliant, and artisanal approaches (Caplan, 2018), with the typical trajectory of development moving towards more standardized and automated methods as the platform grows. Recent discussions on moderation have focused heavily on content moderation by the platforms, that is, on the removal of unwanted content typically via automated or machine-assisted means (e.g., Gillespie, 2018). However, several online communities or entire sites rely instead on peer-level strategies of human moderation conducted by volunteers or the entire community (Seering et al., 2019; Postigo, 2009), often afforded by platform features designed to support and scale up the activity. Studies have

shown that human moderators play a key role in shaping the users' experiences and in the development of meaningful communities by overseeing and nurturing discussion, not only judging the suitability of content (Seering et al., 2019; Ruckenstein & Turunen, 2020; Grimmelmann, 2015; Postigo, 2009; Laaksonen & Porttikivi, 2021).

According to Dolata and Schrape (2025), platforms are places where social action meets significant market interest by companies. However, their organizational structures are centered around the platform companies that manage them, which implement various control mechanisms to regulate the action taking place on the platform (see also Srnicek, 2017; Etter & Albu, 2020). Studies highlight the role of platforms as mediators, facilitators, agents, or brokers in supporting collective activities (Beverungen et al., 2019; Bennett & Segerberg, 2013; Lee et al., 2021; Milan, 2015). As Bucher (2018) argues, platforms do not merely host interactions but also actively structure them through algorithmic and interface design. In this sense, while platforms enable collective organizing, they also constrain and shape it, rendering the process of organizing inherently *platformed*. In short, platforms are not neutral mediators; they actively intervene in the interactions between platform users by enabling or constraining certain types of actions through various mechanisms, one of which is the concept of platform affordances (Bucher & Helmond, 2018; Davis, 2020; Lee et al., 2021). Therefore, as digital platforms control and discipline behaviors, they can simultaneously serve as power devices (Badouard et al., 2016; Barry, 2019).

Platform governmentality in the context of digital platform

Foucault coined the term *governmentality* to describe the processes of power by which subjects are governed, a concept that, in the context of platform capitalism, is particularly relevant for analyzing the control techniques that make subjects governable within the techno-socio-economic infrastructure of the platform (Badouard et al., 2016; Barry, 2019;

Rieder, 2016). By studying, in his work on governmentality, how governments exert control over their populations through policy instruments, Foucault identifies a shift from the regime of sovereignty to regulating through norms via the regime of discipline and managing population through the regime of security. Discipline focuses on correcting deviations and optimizing individual performance within a normative system. Security, by contrast, aims to manage uncertainty: predicting risks, accepting that some errors will occur, and creating conditions to maintain the circulation of people, goods, or information. We posit that governmentality on digital platforms operates through both mechanisms of discipline and security, targeting the different stakeholder groups of the platform.

Studies have explored the technological mechanisms of platform governmentality, such as the use of algorithms, moderation policies, and surveillance practices (e.g., Badouard et al., 2016; Barry, 2019), but less is known about how those mechanisms become visible and contested in moments of conflict. In particular, the role of moderators, positioned both as governed actors and as governing agents, remains undertheorized. Previous research has nevertheless shown that moderators may engage in collective action to contest platform policies (Matias, 2019; Schmitz & Samory, 2025). The studies suggest that strikes constitute critical moments in which platform power relations become visible. In expanding upon these foundational perspectives, our analysis of the SE strike aims to gain an understanding of the power relations within digital platforms, with particular attention given to how the conflicts involving AI-generated content revealed the dynamics through which platform governmentality is negotiated and transformed.

Material and Methods

Research field: the conflict on the Stack Exchange Network (SE)

The SE is an online communication platform, launched in 2009, that hosts more than 180

question-and-answer (Q&A) sites on diverse topics (e.g., the sciences, technologies, engineering, and mathematics). Participants, often professionals, engage voluntarily and autonomously, collaboratively contributing to the development and continuous improvement of knowledge. The most well-known and oldest subsite is the programming-focused Stack Overflow, launched in 2008. The SE platform is owned by a single company, Stack Exchange, Inc. (hereinafter INC), which was acquired in 2021 by the technology investment group Prosus for USD 1.8 billion, underscoring the platform's growing economic significance.

The SE uses several moderation and community-building mechanisms typical of social media platforms: the community is based on questions and answers that the users can upvote and downvote to indicate the quality of the content, and the majority of regular moderation work is done by the users, who, after gaining a good enough reputation, also have the capacity to downvote, flag, and even edit the content. Over time, because of its structured and curated nature, the content has become a key source of the platform's economic value. Since the dataset covering the entire network has been openly available with a Creative Commons license through the so-called Data Dumps file in the Internet Archive, it has served as training data for large language models.⁴ GenAI technologies trained with SE content are well-equipped to solve programming-related problems, as many users of ChatGPT and others have learned. This development has offered SE users a new route to knowledge through ChatGPT and also enabled them to generate answers with an attempt to gain a good reputation on the platform. At the same time, it has reduced human activity on the site (Rio-Chanona et al., 2024), with potential implications for the INC's

⁴ See <https://meta.stackexchange.com/questions/388551/is-se-going-to-be-selling-our-content-for-ai-model-training-and-what-exactly>.

business model. In July 2023, INC announced that it would be integrating generative AI into the platform and developing new AI-based products.⁵

In May 2023, prior to these announcements, INC released a new policy regarding AI-generated content on SE.⁶ The policy instructed moderators to stop suspending AI-generated content, in particular by using GPT detectors that were deemed unreliable. The new policy contradicted earlier moderator-initiated rules that had banned the use of ChatGPT and other LLMs, first on Stack Overflow in December 2022⁷ and later on other subsites. The new policy was not well received by SE moderators, who felt that AI-generated content would now *de facto* be allowed on the network. In this study, we scrutinize the discussions that followed the strike call to understand how it changed the way the platform was governed.

Data collection and analytical strategy

We proceeded iteratively and abductively with a first round of data collection and analysis to generate plausible ‘first suggestions’ (Bamberger, 2019) about the strike and what it triggered in terms of the renegotiating of power relations. In the second round, our aim was to thicken our understanding of the platform’s governmentality and the way in which the mechanisms sustaining it were repurposed to strike, resist, and renegotiate power relations among stakeholders.

In the first round, inspired by Foucault’s (1972) archeology of knowledge, our method drew on the SE Data Dumps file, which constitutes an archive of user interactions (questions, answers, comments). We targeted our query at the SE Q&A sites’ meta sections, which act

⁵ See <https://stackoverflow.blog/2023/07/27/announcing-overflowai/>.

⁶ See <https://meta.stackexchange.com/questions/389582/what-is-the-network-policy-regarding-ai-generated-content>.

⁷ See <https://meta.stackoverflow.com/questions/421831/policy-generative-ai-e-g-chatgpt-is-banned>

as organizing backstages for the SE sites, and included all the threads that mention the keywords ‘moderation strike,’ are tagged with the word ‘discussion,’ and were published after May 30, 2023. The dataset includes 60 threads, 146 answers, and 672 comments (figure 1).

[Insert Figure 1 about here]

When exploring the dataset, we paid special attention to the votes received, reflecting the importance that the community attributes to those threads, and focused on two subsites where strike-related discussions were concentrated: SE Meta (blue color in figure 1) and Stack Overflow Meta (red color in figure 1). This exploration led us to reduce our scope to 22 threads (n=22, with 90 answers and 422 comments in total), which contained thick descriptions of the conflict and the debates surrounding the events and actions undertaken. The remaining 38 threads merely indicated the selected threads. We also accessed the SE community’s Discord server to announce the study and obtain contextual clarifications from active moderators regarding the organization and dynamics of the strike. However, discussions on Discord were not treated as primary data. This first round of data collection helped us describe the sequence of events constituting the strike movement from the letter of declaration to the negotiation of an agreement and revealed that moderators were repurposing moderation tools to resist the proposed policy.

In the second round, we then evaluated the emerging empirical findings commensurate with the Foucault’s conceptualization of regimes of power and governmentality. As a result, we expanded our investigation from the strike itself to the broader system of platform governance on SE, including other documents, to explain how governmentality is achieved on SE through norms, design, and affordances. The other

documents included the Code of Conduct⁸ and the Moderator Agreement.⁹ Using the data, we conducted joint research sessions to analyze how governmentality operates on SE. In doing so, we listed all the rules and features of the platforms based on how Foucault defines discipline and security (see table 1). Finally, we returned to the thread contents to better understand the logics of resistance that the users and moderators were pursuing in relation to the governmentality mechanisms identified. This second round of analysis enabled us to understand that the strike was not just a question of resisting GenAI, but a movement that set in motion a deeper revision of the platform's governmentality.

Given the sensitive and affective nature of the events being studied, particular attention was paid to research ethics. Although the analyzed data is publicly accessible, we anonymized usernames in all excerpts and carefully considered potential harm resulting from reproducing the quotations.

Findings: Platform governmentality on SE

We began by examining how governmentality operates on the SE platform: what kinds of instruments, rules, and practices structure participation and how those mechanisms target different stakeholder groups. In analyzing the platform's apparatus and the ways in which users and moderators described them, we identified nine different mechanisms, which we then mapped based on Foucault's concepts of discipline and security (see table 1).

Disciplinary mechanisms act on individuals through surveillance, corrections, and sanctions, whereas security mechanisms operate at another level: they modify the environment, probabilities, and conditions of action in a way that increases the likelihood of certain behaviors.

⁸ <https://meta.stackexchange.com/conduct>

⁹ <https://stackoverflow.com/legal/moderator-agreement>

Disciplining users and normalizing content

Essentially, SE users are subjected to a set of norms related to quality. The axiomatic disciplinary mechanism on SE is to sanction content positively or negatively based on quality assessment through upvoting and downvoting. The reputation system rewards users with points for valuable contributions, like upvotes on their questions (+5) and answers (+10), or having their answers accepted as the best (+15). A higher reputation unlocks privileges, such as editing posts, voting on content, and moderating. This reputation mechanism acts as a form of disciplining both users and contributors by shaping their behaviors within the framework of platform norms and affords new disciplinary means to the users as their reputation grows.

Elected moderators concentrate on enforcing disciplinary measures directed at regular users rather than the content. Moderators' primary responsibility lies in ensuring compliance with the rules of the SE network and adjudicating individual cases. Moderators act as 'human exception handlers' who only intervene in abnormal situations when the self-regulation of the community fails. Moderators have the authority to ban users who break the norms, while users themselves can flag inappropriate content for moderators to review. However, moderators are also subject to precise rules. First, they are organized hierarchically, embedding their actions within a broader chain of power. Second, their role is dictated by the Moderator Agreement and a series of documents that help them to discipline users (e.g., Code of Conduct). These structures, dictated by the platform-implemented rules and founding documents constantly referred to in the discussion, reveal a disciplining of users and a standardization of their conduct on the platform.

Self-regulation and stability maintenance

The managing of moderators and regular users is based on a set of mechanisms and

knowledge that aims to ensure the production of high-quality content and enable self-organization. As company co-founder Jeff Atwood stated in 2009: ‘We designed the Stack Exchange network engine to be mostly self-regulating, in that we amortize the overall moderation cost of the system across thousands of teeny-tiny slices of effort contributed by regular, everyday users.’¹⁰ This self-regulation process is achieved through the implementation of reputation scores, which serve a dual role: they encourage compliance with platform norms and encourage a sense of meritocracy, where users’ privileges and responsibilities are tied to their reputation. Thereby, the governmentality logic avoids direct control and promotes comparison and differentiation among users. In addition, moderators have access to resources, such as a library of knowledge, that guide their actions and reduce the need for direct intervention. Soft pressure is also applied through formalized principles, such as a blog post entitled ‘A Theory of Moderation,’ which outlines principles for predicting and mitigating risks in the moderation process. Finally, the platform’s technical architecture supports the continuous refinement of policies. Subsite-specific meta sections, the main SE meta subsite, as well as private chatrooms (where INC representatives are also present) serve as spaces for collaboration and decision-making.

The two regimes of discipline and security structure everyday governance on SE, often reinforcing each other but also generating tensions. As Foucault (2004) notes, discipline seeks to optimize the performance of individuals within a normative system, while security aims to manage the environment, predicting and avoiding risks, creating conditions to encourage desired behavior. The arrival of GenAI destabilized this interplay, bringing into sharp relief the tensions between discipline and security and triggering the conflict that led to the moderation strike.

[Insert Table 1 about here]

¹⁰ See <https://stackoverflow.blog/2009/05/18/a-theory-of-moderation/>.

AI-generated content controversy

The earliest threads related to the strike, initiated on June 4 and 5, 2023,^{11,12} offer insights into the evolution of events leading to the moderation strike. The community of moderators observed an increase in the use of AI-generated responses, as expressed in the Stack Overflow site moderators' AI policy, published on December 5, 2022 (see figure 2). Some users were even accused of gaming the system by posting ChatGPT-generated answers to rapidly improve their reputation score. In response to the worrying increase in content suspected of being generated with the help of AI, moderators implemented a series of procedures. They used a technical tool to detect if a post was GPT-generated. If so, the post was subject to processing: reporting, warning, and potentially banning. This procedure, while sometimes expeditious, suffered from several weaknesses by the moderators' own admission. Some of the decisions were considered abusive by the users and were the subject of complaints to INC.

[Insert Figure 2 about here]

From the company's perspective, a large amount of user-generated content was deleted without sufficient justification or with ambiguous heuristics by the moderators. Additionally, INC claimed that the moderators' actions resulted in bias toward non-native English speakers, who used GPT tools for proofreading. Following these novel concerns, the company issued two unilateral decisions—one shared only among moderators and one issued publicly—requesting moderators to stop deleting AI-generated content. In so doing, INC failed to respect the Moderator Agreement, which required advance communication of such changes, leaving moderators with insufficient time to react and disregarding their

¹¹ See <https://meta.stackoverflow.com/questions/424979>.

¹² See <https://meta.stackexchange.com/questions/389811/moderation-strike-stack-overflow-inc-cannot-consistently-ignore-mistreat-an>.

authority to legislate on issues that threaten knowledge generation on SE. By permitting AI-generated content without first consulting the moderators, INC challenged the standards that moderators had helped establish, thereby undermining their power. INC's decisions reflect a move toward security: rather than controlling each user, the company seeks to configure an environment that maximizes content production and circulation, optimizing the flows of traffic, engagement, and data, all of which are essential to its economic model.¹³ This controversy was at the heart of the moderators' strike movement.

The moderation strike as collective action

Drawing on classic repertoires of contention and networked mobilization, the moderation strike can be viewed as a form of collective action. As a response to the controversy, the moderators collectively decided to withdraw from their moderating responsibilities, as detailed in an open letter to INC (appendix 1). The letter blamed INC for the strike. By limiting moderators' ability to identify and filter AI-generated content, INC had broken the loyalty contract with the moderators. As voluntary members, they shared the sentiment that explicit and implicit contracts, including the Moderation Agreement, had been breached and that the company was exhibiting vertical and authoritarian leadership.

Although the issues related to AI-generated content and the prohibitions on moderation certainly contributed to the conflict, the lack of trust and respect for moderators on the part of INC triggered a resistance movement that aimed to revendicate the moderators' role on the platform and put pressure on INC to negotiate. What was at stake was the way in which the platform is governed and power distributed within it. The strike, thus, materialized a power struggle over platform governmentality.

¹³ See <https://stackoverflow.blog/2023/07/27/announcing-overflowai/>.

The open letter not only identified the causes of the strike but also outlined the specific actions that were permissible or prohibited during the strike (figure 3). Moderators defined the boundaries of their collective resistance, encouraging some actions—such as the strategic withdrawal of moderation activities—and discouraging others, including vandalism or any action that might damage the platform’s long-term functionality. Therefore, they still maintained their underlying commitment to quality content.

[Insert Figure 3 about here]

The moderators shared the call to strike across different communities, stressing the rhetoric of broken trust with the company. As shown in figure 1, the strike movement was the subject of numerous exchanges across the platform. A growing solidarity among moderators and users further strengthened the resistance against INC. Moderators mobilized support through the distribution of an open letter as a rallying cry, highlighting the number of signatories—125 moderators, 20 former moderators and staff, and 1,522 regular users.

This effort underscored the movement’s mass effect and indicated the power dynamics within SE. Of the 525 diamond moderators with special privileges, 25% participated in the strike, predominantly from high-traffic sites like Stack Overflow, which are crucial to daily operations on the platform.

Additionally, the striking moderators used their deep knowledge of the platform to strategically utilize reputation-based and visibility affordances to disseminate strike-related information, generating a network of resistance. They systematically informed members about strike actions, explained the underlying reasons, referenced the open letter, announced and justified the strike through authoritative messages, identified participating moderators, and explained how others could participate. The spread of information not only demonstrated strong support among moderators but also highlighted significant endorsements from the broader user community. The mass effect resulted in a reduction of

moderation activities, which significantly impacted the platform. For example, INC employees had to take on moderation tasks in addition to their regular duties.

Moreover, the moderators suspended some automated tools that support their disciplinary power to strike, thus repurposing the moderation infrastructure to resist. This adaptive repurposing of digital affordances illustrates their dual role in both governance and resistance, underscoring power dynamics within socio-technical systems. A detailed analysis by active users¹⁴ revealed that the average resolution time in the review queue increased. Normally, a low-quality question would be closed if it received multiple flags. The strike resulted in fewer posts on Stack Overflow being closed, leading to an increase in undesirable content visible on the site. As shown in figure 4, the strike did not produce a platform breakdown but led instead to measurable delays in moderation and a deterioration of content quality.

[Insert Figure 4 about here]

From strike to negotiations: reshaping the platform governmentality

The moderation strike on SE served as a critical moment in the reconfiguration of the platform's governmentality, bringing into focus the underlying mechanisms of power. By stopping all forms of moderation, the content quickly deteriorated, and the moderation work was delegated to the company's employees, thus making the power of moderators over the SE community quite visible to INC. The situation presumably helped to accelerate INC's invitation to the moderators to discuss matters at the negotiating table. On June 11, the strike took a new turn as INC asked the strikers to designate representatives for negotiations; the representatives were chosen by popular vote (figure 5). An announcement was made that INC representatives would establish contact with the strikers and start a discussion, for

¹⁴ See <https://meta.stackoverflow.com/questions/425115/>.

confidentiality reasons, on a dedicated Discord channel. The voting process for the moderators' representatives took place during a chat on the SE Meta subsite. Representing the company were the Community Manager and the Vice President for Community, ensuring alignment with the company's strategic goals, while two mediators facilitated the discussions. During this phase, the strike actions developed on two parallel streams of action: the negotiation itself, on the one hand, and the pursuit of collective action on the platform on the other. The coordination of both streams took place on a private Discord server.

[Insert Figure 5 about here]

During the negotiations, the moderators' efforts focused first on making public a series of documents at the root of the conflict: the ban on moderating AI-generated content¹⁵ and an internal study conducted on the reliability of AI detectors. The struggle reached a new level when INC suspended the Data Dump on June 9, a full, Creative Commons-licensed backup of the platform's content, which moderators and the community relied on as a form of 'insurance' to guarantee that the platform's data could not be fully privatized or controlled. The reason advanced by the company was that they wanted to avoid LLMs being trained on their data, which would then make it possible for people to exploit the data for private and economic reasons. This event added to the growing distrust between INC, moderators, and the community.

Eventually, the negotiations between the moderators and INC yielded significant outcomes, aimed at improving the moderation process and addressing key concerns of the moderators. The negotiators agreed on a set of criteria that would guide moderators in making decisions about content moderation. This list of heuristics was not disclosed, as

¹⁵ See <https://meta.stackexchange.com/questions/391626/historical-policy-on-the-use-of-gpt-detectors>.

requested by the moderators, to prevent potential workarounds (see figure 6). It represents a profound change compared to the earlier security regime of transparency.

[Insert Figure 6 about here]

By keeping the heuristics confidential, moderators maintained a level of control over the moderation process, ensuring that their expertise and judgment remained central. This agreement also opened new socio-technical dynamics to support and implement the moderation policy. Technical elements of the platform would be upgraded to enforce the new policies and detect violations effectively. INC announced that it planned to issue a warning to users pasting content into the answer box, notifying them about the prohibition on GenAI-generated content.

Internal processes were improved to facilitate better guidance and communication. The Moderator Agreement (MA) was revised,¹⁶ reshaping the power relations between INC and moderators. The new agreement¹⁷ emphasized moderators' responsibility to protect sensitive corporate information, aligning their role with broader organizational interests (MA part 1 vii). On the other hand, it introduced a protective measure for moderators regarding actions performed in good faith while carrying out their moderation duties (MA part 2, ii). The updated agreement also introduced disciplinary mechanisms allowing moderators to hold INC accountable for actions that might affect their moderation duties. Accordingly, INC is required to provide moderators with opportunities to review and provide feedback on network-wide policies before implementation (MA part 2, vi). It also provides a mechanism by which moderators can vote to determine if INC has violated the Moderator Agreement (MA part 2, viii).

¹⁶ For changes in the Moderator Agreement between 2022 and 2024, see: <https://www.diffchecker.com/dBDxHyxP/>.

¹⁷ See <https://stackoverflow.com/legal/moderator-agreement>.

All these achievements transformed the security regime of SE governmentality. The reinstated Data Dump restored the accessibility of the content and thus its autonomy from the platform owners, while the changes made to the moderator's agreement prevented future issues of communication between moderators and INC and the heuristics framed the power of moderators with a number of criteria guiding content moderation. The publicity of these agreements was also central for the moderators. To remain legitimate in the eyes of the community, moderators need to disclose the norms essential to their power. This need is all the more important since the trust between the community and INC seemed to be permanently weakened (see figure 7).

[Insert Figure 7 about here]

Discussion

This study investigated the struggles over governmentality on the SE platform. As the angle of analysis was the moderation strike, we sought to explore how the mechanisms of governmentality, anchored in discipline and security (see table 1), were contested and renegotiated in response to the platform's unilateral policy changes. We interpreted the case within the context of platform capitalism (Srnicek, 2017), which seeks to balance autonomy and control. While content control previously relied mainly on disciplinary mechanisms aimed at correcting and sanctioning deviations, the scale and uncertainty associated with AI have weakened those forms of regulation. Platforms therefore increasingly rely on security-oriented mechanisms that tolerate greater deviations and primarily seek to maintain content circulation, reflecting the growing importance of data production and monetization in platform capitalism. From INC's perspective, tolerating AI-generated content reflects an economic rationality that prioritizes traffic, data circulation, and platform value over strict content discipline.

The role of power mechanisms within platform governmentality

The strike sheds light on the role of the technical, contractual, and symbolic mechanisms on which platform governmentality is built, and it shows how such mechanisms—between discipline and security—are themselves embedded in power dynamics. Prior to the conflict, disciplinary tools—automated detectors, deletions, and bans—were central to maintaining content quality. By forbidding tools that moderators had adopted to combat GPT-generated content, INC attempted to strengthen a security logic privileging risk management and content deregulation over the correction of deviations. The negotiations that followed reasserted the role of disciplinary mechanisms, as moderators retained authority through confidential heuristics and revised procedures.

Understanding this reconfiguration also puts the moderators themselves in the limelight. Our analysis shows that moderators occupy a hybrid position. As in other online knowledge communities (Konieczny, 2016), decision-making processes during the struggle were handled by a relatively small number of active contributors. On the one hand, they are governed by platform policies and technical mechanisms. On the other, they enact governance themselves by disciplining users and maintaining platform standards. As Matias (2019) suggests, the activity of volunteer moderators extends beyond merely performing the tasks associated with their role: their actions are embedded in power dynamics, both among themselves and within platform capitalism. At the same time, they are actors with their own incentives and motivations, which might sometimes contradict those of the platform. In our case, the moderators occupy a key position within the platform's organizational structure, performing tasks that are formally assigned to them by platform management and operating within established boundaries and policies. By reinforcing their authority through new heuristics and procedures, moderators also help embed discipline more deeply within the platform's technological infrastructure, exemplifying, perhaps unintentionally, Winner's

(1980) classical idea that ‘technologies are ways of building order in our world’ (p. 127). In the context of platformed organizing, then, our findings confirm the importance of both social capital—reputation, networks, and trust among moderators—and the spatial dimension that has been highlighted in similar strikes (Schmitz & Samory, 2025). However, this strike also highlights a power struggle in which moderators are confronted with a company seeking to optimize its business model in response to the rise of AI. Against the company’s strategy of deregulation, moderators defended a community logic that has prevailed since the creation of the platform, emphasizing the imperative of autonomy and the idea that community knowledge should not be fully enclosed within corporate boundaries. When INC unilaterally restricted moderators’ authority, the moderators interpreted it not only as a breach of formal agreements but also as a violation of this shared custodianship of data, thereby triggering a shift toward a logic of resistance.

The dialectic of power and resistance

According to Foucault, power relations inherently entail possibilities for resistance. When action is mediated by digital platforms, the same technologies and strategies that enable governmentality can also be mobilized for resistance. In the context of platformed organizing, resistance often takes the form of opposing governing structures, algorithmic systems, and the policies of platforms (e.g., Ettlinger 2018; Bonini & Treré, 2024; Charitsis et al., 2024). Foucault’s view of resistance, that it is ‘just as inventive, as mobile, as productive as power itself; (...) like power, it organizes, coalesces, and cements itself’ (1994, p. 267, trans.), echoes in our data and helps us understand how moderators actively reshape the power structures embedded within the platform. In practice, it took the form of repurposing the platform’s features: automated tools were suspended to highlight their absence, reputation and visibility signals were strategically mobilized to legitimize the

strike, and discussion spaces initially designed for coordination became hubs of mobilization and awareness. These practices resonate with tactics observed in data activism (Milioni & Papa, 2022) and algorithmic reconfiguration (Bonini & Treré, 2024). While technology might appear to be a hard-coded structure, it can still be remolded through human use and repurposed for resistance.

While the strikers expressed discontent and presented specific demands to INC, the strike itself also raised questions about the target and meaning of the resistance. In this case, the aim was to maintain the status quo, a stable functioning platform, and a type of governmentality that supported the community. Before the policy change, automated tools were used extensively to help moderators, not to eventually replace them or diminish the content quality but to help them fight against the influx of GPT-generated sub-quality content onto the platform. Maintaining content quality remains a shared objective between the corporation owning the platform and the moderators since it also serves as a critical resource within the platform's data-driven business model. This finding aligns with existing literature emphasizing the centrality of data extraction and accumulation to feed the capitalist logic of platform economies—processes inherently dependent on user participation (Srnicsek, 2017). The moderators' choice to maintain order during the strike reinforced this alignment. Ultimately, the strike sought to preserve a balance of power among moderators, INC, users, and platform infrastructure.

Conclusion

This study has examined how a moderator strike unfolded on the online Q&A Stack Exchange Network site. Our analysis unpacked how mechanisms of governmentality are being renegotiated within platforms in the era of platform capitalism and as a result of the changing dynamics introduced by AI tools. By doing so, we contribute to platform studies

by introducing the concept of platform governmentality to describe how power/resistance mechanisms work in the context of platform capitalism through the regimes of discipline and security. The strike, organized by moderators committed to the transparency of platform governmentality and the value and quality of the content, illustrates how GenAI can trigger a conflict around platform governmentality. In the context of platform capitalism, where both users and quality data are essential components of the platform's business model (Srnicek, 2017; Van Dijck et al., 2018), GenAI-generated content tilts the existing power balance between the moderators, users, and the platform by reducing the cost of content creation and tempting users to game the reputation system. The moderators went on strike to make their power visible and engaged in negotiations with the platform company to solidify their power position as postulated in the Moderator Agreement. Our contributions also relate to the question of resistance in studies on the organizing of collective action. This body of literature often tends to produce typologies of resistance practices (Mumby et al., 2017; Charitsis et al., 2024). We have deliberately chosen not to produce this kind of list and opted instead to foreground the processual and complex dimensions of resistance. Our analysis has shown that resistance can also be a means of maintaining the status quo rather than creating a clean break with a certain way of operating.

This study, therefore, has investigated one of the first concrete cases where actors protest in reaction to the use of GenAI and examined its consequences for platform governmentality. Complementing prior work on how AI imaginaries materialize in policy decisions on data governance (Bareis & Katzenbach, 2022), our research highlights the power dynamics and transformations triggered by new uses of AI for the governmentality of a platform dedicated to producing high-quality information. This upheaval calls into question the primacy of humans in the production and validation of online content. While automation has traditionally played a supporting role in the moderation process, it is now

assuming a more active role by contributing to the content creation process. This change signifies a new era in the integration of AI into information production. Like all studies, this one has limitations. First, our analysis reflects only the voices of active participants and moderators on the SE platform, which leaves the experiences of regular users and the longer-term effects of the strike in the shadows. Second, even the moderators used other communication systems to continue their discussions and strike-organizing efforts outside the spaces offered by the platform, and we had no proper access to those spaces. Therefore, future studies should seek to further understand the virtual spaces of platform governmentality as well as the motivations of the actors by conducting, for example, interviews with moderators and platform administrators. A third limitation is that our investigation focused on the specific setting of SE, which is arguably a very special type of platform. We believe that a similar governmentality framework can be useful for analyzing cases on other platforms as well, especially since many of the features and practices are similar across different platforms (e.g., Reddit). However, the contextual premises are most similar on knowledge-generation platforms with similar collective aims and moderation systems (such as Wikipedia), and perhaps less so on more leisure-oriented social media platforms. Future studies should aim to further theorize the governmentality and power/resistance mechanisms across different platforms.

References

- Ashuri, T., & Bar-Ilan, Y. (2016). How flat organizations filter: organizational gatekeeping in a networked environment. *Information, Communication & Society*, 19(10), 1411–1426.
- Badouard, R., Mabi, C., & Sire, G. (2016). Beyond "Points of Control": logics of digital governmentality. *Internet Policy Review*, 5(3), 1-13.

- Bamberger, P. (2019). On the Replicability of Abductive Research in Management and Organizations: Internal Replication and Its Alternatives. *Academy of Management Discoveries*, 5(2), 103-108.
- Bareis, J., & Katzenbach, C. (2022). Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science Technology and Human Values*, 47(5), 855–881.
- Barry, L. (2019). The rationality of the digital governmentality. *Journal for Cultural Research*, 23(4), 365-380.
- Bennett, L & Segerberg, A. (2013). *The Logic of Connective Action: Digital Media and the Personalization of Contentious Politics*. Cambridge University Press.
- Beverungen, A., Beyes, T., & Conrad, L. (2019). The organizational powers of (digital) media. *Organization*, 26(5), 621–635.
- Bonini, T. & Treré, E. (2024). *Algorithms of Resistance: The Everyday Fight against Platform Power*. MIT Press.
- Bucher, T. (2012). Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society*, 14(7), 1164-1180.
- Bucher, T. (2018). *If ... then: Algorithmic power and politics*. Oxford University Press
- Bucher, T., & Helmond, A. (2018). The affordances of social media platforms. In J. Burgess, A. Marwick, T. Poell (Eds.) *The SAGE Handbook of Social Media* (pp. 233-253). SAGE Publications.
- Burtch, G., Lee, D., & Chen, Z. (2024). The consequences of generative AI for online knowledge communities. *Scientific Reports*, 14(1), 10413.
- Caplan, R. (2018, November 14). *Content or context moderation?* Data & Society Research Institute. <https://datasociety.net/output/content-or-context-moderation/>

- Charitsis, V., Laamanen, M., & Lehtiniemi, T. (2024). Towards Algorithmic Luddism: class politics in data capitalism. *Information, Communication & Society*, 1–18.
<https://doi.org/10.1080/1369118X.2024.2435996>
- Davis J. (2020). *How artifacts afford: The power and politics of everyday things*. MIT Press.
- Dolata, U., & Schrape, J. (2016). Masses, Crowds, Communities, Movements: Collective Action in the Internet Age. *Social Movement Studies*, 15(1), 1–18.
- Dolata, U., & Schrape, J. (2025). Platform companies on the internet as a new organizational form. A sociological perspective. *Innovation: The European Journal of Social Science Research*, 38(3), 1228–1247.
- Duffy, B. E., & Meisner, C. (2023). Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility. *Media, Culture & Society*, 45(2), 285-304.
- Etter, M., & Albu, O. (2021). Activists in the dark: Social media algorithms and collective action in two social movement organizations. *Organization*, 28(1), 68–91.
- Ettlinger, N. (2018). Algorithmic affordances for productive resistance. *Big Data & Society*, 5(1).
- Foucault, M. (2004). *Sécurité, territoire, population. Cours au Collège de France (1977-78)*. Gallimard.
- Foucault, M. (1994). *Dits et écrits (1954-1988), tome III: 1976-1979, texte n°200*. Gallimard.
- Foucault, M. (1972). *The Archaeology of Knowledge and The Discourse on Language*. Pantheon Books.
- Foucault, M. (1982). The Subject and Power. *Critical Inquiry*, 8(4), 777–795.

- Gehl, R., & Zulli, D. (2023). The digital covenant: non-centralized platform governance on the mastodon social network. *Information, Communication & Society*, 26(16), 3275–3291.
- Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 205395171989794.
- Grimmelmann, J. (2015). The Virtues of Moderation. *17 Yale Journal of Law & Technology* 42.
- Kiesler, S., Kraut, R., Resnick, P., & Kittur, A. (2012). Regulating behavior in online communities. In Kraut, R.; Resnick, P.; Kiesler, S.; Burke, M.; Chen, Y., Kittur, N. (Eds.), *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press.
- Konieczny P. (2016) Wikipedia in the anti-SOPA protests as a case study of direct, deliberative democracy in cyberspace, *Information, Communication & Society*, 20(2), 167-184.
- Laaksonen, S. M., & Porttikivi, M. (2022). Governing with conversation culture—conditioning organizational interaction in a digital social movement. *Information, Communication & Society*, 25(10), 1456–1474.
- Lee, F., Liang, H., Cheng, E., Tang, G., & Yuen, S. (2021). Affordances, movement dynamics, and a centralized digital communication platform in a networked movement. *Information, Communication & Society*, 25(12), 1699–1716.

- Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3), 329–346.
- Matias, J. N. (2019). The Civic Labor of Volunteer Moderators Online. *Social Media + Society*, 5(2).
- Milan, S. (2015). From social movements to cloud protesting: the evolution of collective identity. *Information, Communication & Society*, 18(8), 887–900.
- Milioni, D., & Papa, V. (2022). The oppositional affordances of data activism. *Media International Australia*, 183(1), 44-59.
- Mumby, D., Thomas, R., Martí, I., & Seidl, D. (2017). Resistance Redux. *Organization Studies*, 38(9), 1157–1183.
- Postigo, H. (2009). America online volunteers: Lessons from an early co-production community. *International Journal of Cultural Studies*, 12(5), 451–469.
- Rieder, B. (2016). Big Data and the Paradox of Diversity. *Digital Culture & Society*, 2(2), 39–54.
- Rijshouwer, E., Uitermark, J., & de Koster, W. (2023). Wikipedia: a self-organizing bureaucracy. *Information, Communication & Society*, 26(7), 1285-1302.
- del Rio-Chanona, R., Laurentsyeve, N., & Wachs, J. (2024). Large language models reduce public knowledge sharing on online Q&A platforms. *PNAS Nexus*, 3(9).
- Ruckenstein, M., & Turunen, L. (2020). Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society*, 22(6), 1026-1042.
- Schmitz, A., & Samory, M. (2025). From Volunteerism to Corporatization: Analyzing Participation in the 2015 and 2023 Reddit Blackouts. *Social Media + Society*, 11(1).

- Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7), 1417-1443.
- Srnicek, N. (2017). *Platform Capitalism*. Polity Press.
- Van Dijck, J., Poell T., & de Waal M. (2018). *The Platform Society*. Oxford University Press.
- Winner, L. 1980. Do artifacts have politics? *Daedalus*, 109(1), 121–36.

June 5, 2023

Stack Overflow, Inc. has decreed a **near-total prohibition on moderating AI-generated content** in the wake of a flood of such content being posted to and subsequently removed from the Stack Exchange network, tacitly allowing the proliferation of incorrect information ("hallucinations") and unfettered plagiarism on the Stack Exchange network. This poses a major threat to the integrity and trustworthiness of the platform and its content.

We, the undersigned, are volunteer moderators, contributors, and users of Stack Overflow and the Stack Exchange network. Effective immediately, we are enacting a **general moderation strike** on Stack Overflow and the Stack Exchange network, in protest of this and other recent and upcoming changes to policy and the platform that are being forced upon us by Stack Overflow, Inc.

Our efforts to effect change through proper channels have been ignored, and our concerns disregarded at every turn. Now, as a last resort, we are striking out of dedication to the platform that we have put over a decade of care and volunteer effort into. We deeply believe in the core mission of the Stack Exchange network: to provide a repository of high-quality information in the form of questions and answers, and the recent actions taken by Stack Overflow, Inc. are **directly harmful to that goal**.

Specifically, moderators are no longer allowed to remove AI-generated answers on the basis of being AI-generated, outside of exceedingly narrow circumstances. This results in effectively permitting nearly all AI-generated answers to be freely posted, regardless of [established community consensus](#) on such content.

In turn, this allows incorrect information (colloquially referred to as "[hallucinations](#)") and plagiarism to proliferate unchecked on the platform. This destroys trust in the platform, as [Stack Overflow, Inc. has previously noted](#).

In addition, the details of the policies issued directly to moderators differ substantially from the guidelines outlined publicly, with moderators barred from publicly sharing the details.

These policies disregard the leeway historically granted to individual Stack Exchange communities to determine their policies, by making changes without the input of the community, overriding community consensus, and outright refusing to reconsider their position.

Until this matter is resolved satisfactorily, we will be pausing activities including, but not limited to:

- Raising and handling flags.
- Running [SmokeDetector](#), the anti-spam bot.
- Closing or voting to close posts.
- Deleting or voting to delete posts.
- [Reviewing tasks](#) in the various review queues.
- Running [various other bots](#) designed to assist in moderation, such as detecting plagiarism, low-quality answers, and rude comments.

Until Stack Overflow, Inc. retracts this policy change to a degree that addresses the concerns of the moderators, and allows moderators to effectively enforce established policies against AI-generated answers, we are calling for a general moderation strike, as a last-resort effort to protect the Stack Exchange platform and users from a total loss in value. We would also like to remind Stack Overflow, Inc. that a network that entirely relies on volunteers for its moderation model cannot then consistently ignore, mistreat, and malign those same volunteers.

Appendix 1. Open letter