Special Issue Article

new media & society
2022, Vol. 24(4) 964–981
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14614448221079035
journals.sagepub.com/home/nms
SAGE

Not all who are bots are evil: A cross-platform analysis of automated agent governance

Mykola Makhortykh
University Bern, Switzerland

Aleksandra Urman
University of Zurich, Switzerland

Felix Victor Münch, Amélie Heldt, Stephan Dreyer and Matthias C Kettemann
Leibniz Institute for Media Research | Hans-Bredow-Institut, Germany

Abstract

The growth of online platforms is accompanied by the increasing use of automated agents. Despite being discussed primarily in the context of opinion manipulation, agents play diverse roles within platform ecosystems that raises the need for governance approaches that go beyond policing agents' unwanted behaviour. To provide a more nuanced assessment of agent governance, we introduce an analytical framework that distinguishes between different aspects and forms of governance. We then apply it to explore how agents are governed across nine platforms. Our observations show that despite acknowledging diverse roles of agents, platforms tend to focus on governing selected forms of their misuse. We also observe differences in governance approaches used by platforms, in particular when it comes to the agent rights/obligations and transparency of policing mechanisms. These observations highlight the necessity of advancing the algorithmic governance research agenda and developing a generalizable normative framework for agent governance.

Corresponding author:
Aleksandra Urman, Social Computing Group, University of Zurich, Andreastrasse 15, 8050 Zurich, Switzerland.
Email: urman@ifi.uzh.ch

Keywords

Algorithmic governance, automated agent, automation, bot, platform regulation, policing, transparency

Introduction

Automated agents constitute an important part of the ecosystem of online platforms. Often referred to as (ro)bots,[1] these software products are capable of formulating decisions and acting upon them with little human intervention (Tsvetkova et al., 2017). Such autonomy enables agents' use for automating multiple tasks[2] and makes them an integral component of algorithmic governance both in a sense of using algorithms to govern and governing algorithms themselves (Katzenbach and Ulbricht, 2019). Hence, in this article we look at how platforms condition the use of agents to understand how practices of algorithmic governance – here taking the form of agent governance – are shaped by divergent platform contexts.

The need to govern agents, namely to provide conditions for the 'ordered rule' (Stoker, 1998) of their activities, is attributed to agents' substantive impact on how platforms operate their services and how these services are misused by the users. However, this task is complicated by agents' multi-functionality and often non-deterministic nature, as well as their undefined legal status, in particular concerning whether creators' rights extend to the agents and whether agents can have rights on their own (Fox, 2019). The integration of agents into platform governance (e.g. as community moderators) further complicates matters and raises ethical and legal concerns (Langford, 2020).

Despite its importance to platform ecosystems, no systematic assessment of agent governance has been conducted yet. This can be attributed to the complex nature of the phenomenon, but also to the strong focus of academic and societal debates on agents' misuse for manipulating public opinion. While the importance of agent-driven opinion manipulation is evident, its extensive media coverage led to moral panics (Walsh, 2020) and pushed the research agenda towards studying agents' destructive potential and platform efforts to cull it down. However, such an emphasis is criticized for overestimating agents' manipulative capabilities (Assenmacher et al., 2020), as well as not accounting for other aspects of agent activity that cause changes in digital labour practices (Hukal et al., 2019) and information gate-keeping (Lokot and Diakopoulos, 2016).

Informed by this criticism, we aim to expand the research agenda by acknowledging the diversity of agents' roles and the multiplicity of approaches to their governance. To do so, we introduce an analytical framework that differentiates between different aspects (what is governed) and forms (how it is governed) of agent governance. Then, we apply it to conduct an explorative study of agent governance across nine platforms, where agents are intensively used. Using document analysis, we seek to answer the following research question: How do different platforms govern the use of automated agents?

Conceptualizing automated agent governance

The study of automated agent governance should acknowledge its multilayered nature. For this aim, we propose to differentiate between aspects of governance (specific issues associated with agent activity) and forms of governance (means through which conditions for dealing with these issues are set). Our analytical framework is inspired by research on different approaches to human agent governance in the context of platforms (for forms)[3] and empirical analysis of automated agent governance structures (for aspects). Specifically, we align with critical studies (e.g. Duguay et al., 2020) that stress the importance of taking into consideration not only formal means of governance codified by the platforms but also less formal norms of behaviour arising from user experiences.

Aspects of agent governance

We differentiate between four aspects of agent governance: definitions; rights and obligations; scope of forbidden actions; and sanctions. The first aspect is the *definition of the agent* that determines how it is understood and treated (Lior, 2020). Definitions have a substantial impact on how automated agents are treated and what aspects of their activity are regulated. Discrepancies between agent definitions can lead to their inconsistent treatment by different forms of governance that can decrease transparency of governance structures and enable procedural loopholes. An example of it is Twitter, where the reliance on a rather vague definition of automated activity (i.e. the one relying on frequency of posting) occasionally resulted in its mechanisms misclassifying human users as non-humans (Martineau, 2019).

To our knowledge, no study has systematically investigated the agent definitions different platforms use, although there is recognition of 'an incredible breadth of terminology' (Gorwa and Guilbeault, 2020: 2) used to describe agents. Besides general terms, such as bots or sybils, scholars have suggested more fine-grained typologies differentiating, for instance, between web robots and social bots (Gorwa and Guilbeault, 2020) or fixers and advisors (Zheng et al., 2019). These typologies emphasize the multitude of roles performed by agents, but also raise questions about how governance structures accommodate these diverse roles.

The breadth of agent definitions is partially related to the second aspect of governance: *agent rights and obligations*. By specifying the rules agents are expected to follow and activities they are allowed to conduct, platforms outline principles that circumscribe agents' roles within platform ecosystems. However, considering the diversity of agent roles, defining the rules to ensure the effective governance of agents with a broad range of functionalities is often a non-trivial task both conceptually and practically.

The subject of agent rights is gaining prominence in academic scholarship as it increasingly acknowledges that agents' roles go beyond manipulating public opinion. Recent studies explore both functional rights, namely, activities that agents are allowed to pursue, such as communicating with users (Lokot and Diakopoulos, 2016) or conducting financial operations (Roshchinskaya, 2020), and general rights, such as the freedom of expression (Fox, 2019). The subject of general rights remains particularly debated, in line with the broader discussion on whether robots/artificial intelligence (AI) can or should have rights (Gunkel, 2018). Similarly, there are a growing number of debates on agent obligations, but often they focus on applicability of rules for human agent to non-human ones, whereas more specific obligations (e.g. obligatory self-disclosure of automated nature; Lamo and Calo, 2019) remain less studied.

The aspect which features most prominently in academic and societal debates is *the scope of actions forbidden for agents*. To counter potential agent misuses, platforms indicate which agent actions are not allowed and introduce mechanisms to enforce these prohibitions. Thereby, they inversely determine which uses of agents are acceptable or at least not punishable that further specifies the role of agents in platform ecosystems.

Existing research investigates different types of forbidden activity, varying from the manipulation of public opinion (Ferrara, 2017) to harassment (Uyheng and Carley, 2020) and spamming (Maréchal, 2016). However, most studies do not place the discussion of agent misuse into the larger context of governance. Even in cases where it is evident that a particular misuse violates platform policies, the measures taken to counter it often remain obscure (in particular, misuse detection and prevention that is also the case with other forms of automated content moderation; Gorwa et al., 2020). Such obscurity can mask a gap between normative definitions of forbidden activities and mechanisms used to prevent them, which is detrimental for the rights of both automated and human agents. Furthermore, the few comparative studies (e.g. Maréchal, 2016) indicate substantial differences in what different platforms classify as agent misuse.

The *sanctions* used to punish agents/developers for violating platforms' rules is another key aspect of agent governance. Together with agent rights, sanctions define the relationship between the developer and the agent by establishing whether the former is accountable for the latter's actions. Information about sanction mechanisms is essential for assessing the possibility of unfair punishment (e.g. a human moderator's mistake). Both fair and unfair sanctions impact the rights of human (and, possibly, automated) users and are at issue when such cases are brought before courts.

Similar to the scope of forbidden actions, the sanctions aspect is visible in the debate about agent governance, but its role in how agent governance is structured remains under-studied. Platforms regularly report banning batches of agents involved in forbidden activity, such as disinformation and propaganda (e.g. Timberg and Dwoskin, 2018), but the use of sanctions for other types of rule infringement is less visible. The same is true for the question of responsibility, namely, whether sanctions can/should be applied to the agent and/or its developer.

Forms of agent governance

Shifting our attention from *what* is being governed, to *how* it is governed, we identify three forms of governance: policies, mechanisms and practices. *Policies* refer to the rules and guidelines regulating platform services. Usually codified as terms of service, they delineate possible agent (ab)uses and countermeasures to address rules' infringements. Policies serve as internal 'normative orders' (Kettemann, 2020) that are often more intricate than non-platform governance structures (e.g. national legislation). This is particularly so for areas where specific regulation is yet to be adopted, as is the case for agent regulation. As a result, platform policies determine most of agent governance aspects, ranging from their obligations and rights to the score of forbidden actions.

Despite policies being the foundation of platform governance, there is little research on them in the contexts of automated agents. In theory, policies are equally applicable to human and non-human agents, but in practice governance of the latter is often treated separately (Fox, 2019). Examples of such treatment include obligatory disclosure of agents' automated nature (Gorwa and Guilbeault, 2020; Pedrazzi and Oehmer, 2020) and specific regulation of agents' use with the latter being quite inconsistent between different platforms (Maréchal, 2016).

The second form of governance is *mechanisms* which include procedures and technically implemented limits setting conditions for agent design and use. Mechanisms, such as platform Application Programming Interfaces (APIs) or coding conventions, enforce policies by determining how agents can be deployed on the platform and what functionalities can be programmatically implemented. Together with policies, mechanisms constitute the 'law of cyberspace' (Lessig, 1999) that shapes the formal structures of agent governance.

Research on mechanisms of agent governance is scarce and often focuses on one specific category, namely, mechanisms differentiating between human and non-human activity (Gorwa and Guilbeault, 2020; Rauchfleisch and Kaiser, 2020). However, even for this category, there is limited understanding of what techniques are used by the platforms and to what degree they approximate those used by scholars (Gallwitz and Kreil, 2021). The major exception is Wikipedia, where governance mechanisms are known to rely on a combination of human moderation and automatically generated inputs (Geiger, 2014). In the case of other platforms, in particular more business-oriented ones, the degree of transparency is low. Similar to mechanisms dealing with human agents, which are equally characterized by opacity (Bloch-Wehba, 2020; Gorwa et al., 2020), non-transparency of technical procedures can amplify issues created by policies that are not in line with legal standards.

The third form of governance is *practices*, which are the social norms developing around policies and mechanisms. Despite not being enforceable in a strict sense, practices guide the design and use of agents by developers (Katzenbach and Ulbricht, 2019), which makes them normatively significant. Similar to the governance of human agents (Duguay et al., 2020), practices influence effectiveness of formal forms of governance by reinforcing or undermining their sustainability (e.g. by promoting compliance with policies or, vice versa, identifying ways to circumvent governance mechanisms).

Despite their importance, the analysis of practices is mostly confined to community-reliant platforms. Most existing research focuses on Wikipedia, where the line between practices and policies/mechanisms is blurry as evidenced by the significant role of community-based governance structures (e.g. Bots Approval Group; Geiger, 2018; Livingstone, 2016) that offers an alternative to top-down policing. However, practices also play an important role on other platforms, such as Facebook (Seering et al., 2019), where they enable a 'multidimensional "entanglement" between algorithms put into practice and the social tactics of users who take them up' (Gillespie, 2014: 183).

Methodology

Applying the analytical framework outlined above, we explored automated agent governance on nine platforms: Discord, Facebook, Telegram, Twitter, WhatsApp, Wikipedia, Slack, Viber and Stack Overflow. Our selection followed three criteria: first, we wanted to compare governance approaches across platform types, including social networks (Facebook, Twitter), general-purpose messengers (Telegram, WhatsApp, Viber), domain-specific messengers (Discord, Slack) and knowledge platforms (Wikipedia, Stack Overflow). Second, we chose platforms where agents are integrated in the digital ecosystems, unlike platforms such as Instagram or TikTok where agents are primarily employed to abuse content promotion mechanisms. Third, we focused on platforms that are used globally and within geographical and linguistic contexts the authors are able to adequately assess. Consequently, some other relevant platforms (e.g. WeChat) were omitted from the research design.

We collected data for most of the platforms in August–September 2020 with another round of data collection happening in April 2021 to include Viber and Stack Overflow. For information on policies and mechanisms, we used the platforms' official documents (see Table 1 in Online Appendix): privacy policies, terms of service and codes of conduct together with official statements on agents (policies) and documentation of platforms' APIs (mechanisms). For practices, we combined data from platform developer forums (e.g. Twitter Developer Forum) and external platforms (Stack Overflow, Reddit, Medium). Compared with the other two forms of governance, the analysis of practices turned out to be challenging because of the difficulties with locating relevant information,[4] in particular as many practice-related discussions turned out to be technical and narrow-focused.

After identifying documents relevant for the study, we used document analysis to examine documents related to three forms of governance for up to two platforms. To do so, we read the documents and coded parts of them related to the four governance aspects elaborated above. Then, the coded parts were extracted to a shared document, where we discussed them to identify patterns in agent governance within individual platforms and then compared these patterns on the level of specific aspects across the platforms.

There are two important points which have to be accounted for when interpreting our findings. First, our analysis is largely limited to depictions of agent governance on the document level. Because we rely on document analysis and not on experiments or interviews, it is hard to judge how governance forms are actually used to govern agents. While we tried to compensate for it by looking at practices, the difficulties with data collection complicate their comprehensive analysis. At the same time, formal governance forms, such as official platform documents and API standards, still set up conditions for agent activity within individual platforms, thus defining how the agents are governed. However, this limitation is important to address in the future research.

Second, the nature of documents analysed also has implications for our findings. These documents are produced by certain actors (platform personnel and external developers) with a specific audience in mind (assumingly, platform users and developers) and, thus, reflect a certain set of perspectives on agent governance. While these perspectives are arguably essential, they are not the only ones which matter (e.g. there are practices which are not codified via developer forums as well as internal platform guidelines which are not publicly accessible). There is also a broad range of possible differences (e.g. in professional background or worldviews) between platform actors and audiences that can affect differences in governance approaches together with the recognized tendency of platform documents to lack transparency on matters related to governance (Gorwa et al., 2020). Consequently, while it is still possible to identify differences in platform approaches, understanding the reasons behind them is a non-trivial task that requires a more digital ethnography-like study design.

Findings

Agent definitions

Our analysis shows that despite agents being mentioned in many official policy and internal developer documents, only a few platforms formally define them, whereas practice-related discussions omit agent definitions almost completely (Table 2 in Online Appendix). Most examined platforms refer to agents in general terms, such as 'an automated account – nothing more or less' (Twitter; Roth and Pickles, 2020) or 'code, programs or other interfaces which connect to the API Services' (Stack Overflow, n.d.-b), or treat agents as a self-explanatory concept. The absence of formal definitions results in the lack of differentiation between agent roles that decreases the transparency of agent governance and facilitates rather arbitrary decisions on what agents can (not) be.

If an agent definition is provided, the characteristic of automation is usually emphasized. Agents are referred to as 'a separate type of user account dedicated to automation' (Discord, n.d.-c), 'automated scripts used to provide automation' (Wikipedia, n.d.-b) or something that 'automatically posts content into groups' (Facebook for Developers, n.d.). Some platforms also draw parallels between automated agents and humans in their policy documents. Telegram refers to agents as 'special Telegram users' (Telegram, n.d.-c), attributing certain agency to them, whereas Wikipedia claims that agents interact with the platform 'as though they were human editors' (Wikipedia, n.d.-b). Slack goes even further by referring to places where agents (or 'apps' as platform calls them) 'live' and talk about their 'homes' (Slack, n.d.-a). Whether these platforms denominate agents as users to anthropomorphize them or simply to point at the similarities between automated agents and humans with regard to communicative rights is unclear.

Policy-related differences are amplified by mechanisms used to enforce definitions. Only Telegram, Wikipedia, Viber and Discord include explicit mechanisms to differentiate between human and automated agents: Telegram and Wikipedia require agent accounts to include the word 'bot'; in the case of Telegram (n.d.-a), no phone number is required for creating an automated account (unlike human accounts), even while developer still requires a number; on Viber, agent should be registered on a separate platform (Viber, n.d.-b); while Discord (n.d.-c) features separate rate limits for agents. Other platforms either do not include mechanisms or treat them as non-obligatory, such as Stack Overflow that recommends adding an 'app' tag to denote agents' nature or Twitter that requests developers to reveal the non-human nature of agents, but does not provide clearly defined mechanisms yet.

The lack of information about specific mechanisms defining agents may be interpreted differently. It can be attributed to many platforms seeing agents as a subject relevant only for a small group of developers for whom the notion is self-explanatory which is also supported by the lack of definitions in documents associated with user practices. The same platforms also usually provide more limited access to their APIs that further restricts the need for introducing such mechanisms (as contrasted by Wikipedia/Telegram where such access is more open).

Agent rights and obligations

We observed some similarities in attribution of rights to agents via official platform documents (Table 3 in Online Appendix). Besides accessing user/platform data, which is almost a universal right (except for Facebook), most platforms allow agents to communicate with users. This can take the form of sending/receiving files (Telegram), posting direct messages (Twitter, Viber) or liking posts to 'to indicate acknowledgement or approval' (Facebook for Developers, n.d.). There are also platform-specific rights, such as the right to generate articles (Wikipedia), to charge subscription fees (Telegram) or to follow users (Twitter).

A separate category of rights deals with agents' ability to govern other agents, including humans. Such rights are usually communicated via explicit policy statements (e.g. that agents are eligible to hold administrative rights; Wikipedia) and more implicit references in internal developer documents (e.g. the right to delete content and ban chat members; Telegram). In some cases, such rights are informally acknowledged via informal developer discussions noting agents being misused as moderators for mass user bans (Discord).

The latter case is one of a few instances when we observed discrepancies between governance forms. While Discord policies do not explicitly grant agents governance rights, the possibility of granting these rights via programmatic mechanisms enables the acceptance of such agent uses via informal developer practices. Similarly, in the case of other platforms, we observed the tendency of practices to cover the gaps in policies by attributing more functional rights to agents, such as raising brand awareness (Twitter) or countering vandalism (Wikipedia).

In terms of agent obligations, official platform policies vary even more between platforms (Table 4 in Online Appendix). Only two obligations are relatively common: first, the need for agents to identify themselves as automated entities (Telegram and Wikipedia). Interestingly, such a requirement is not found in policies of Western social media platforms, despite them being at the centre of the debate about agent misuses. A possible explanation is their focus on *coordinated inauthentic behaviour*, whereas agents that communicate *authentically* are not obliged to self-identify. The second common obligation is the requirement to not initiate contact with human users (Telegram, WhatsApp, Viber) that can be seen as a part of a broader obligation to not engage too much in human-like actions (e.g. not to have friends and avoid categorizing humans discussed in the encyclopedia on Discord and Wikipedia, respectively).

A distinct case is made by the more normative obligations: to be helpful and responsive (Twitter), to exercise good judgement and provide good user experience (Slack) and to be harmless and useful (Wikipedia). Such requirements can be treated as part of the increasing recognition of the importance of motivating developers to acknowledge the role of normative values in system design, but their vagueness can also be used to justify sanctions against developers. Furthermore, it is unclear how such obligations are reinforced via mechanisms: while, in the case of Wikipedia, the usefulness can be verified via a trial period, in the case of Twitter no actual measurements of helpfulness are provided.

Different rights and obligations attributed to agents by different platforms pose challenges for universalizing structures of agent governance. While the very idea of attributing rights and obligations to AI is the subject of ongoing debate in scholarship (e.g. Perel and Elkin-Koren, 2019), our observations suggest that platforms already attribute them to agents. This illustrates how the lack of overarching regulation leads to privatization of governance rules that might facilitate the formation of legal loopholes in the relationship between platforms, developers and agents.

Scope of forbidden actions

There are various actions that are deemed off-limits for automated agents by official policy documents (Table 5 in Online Appendix). The most common is spamming, which is forbidden by all examined platforms except Wikipedia and Viber. The exact definition of spam varies, however: for Discord (n.d.-d) and Telegram (n.d.-b) spamming includes unsolicited messages/advertisements and server/channel invites, whereas for Twitter it also includes posting identical tweets (Twitter Help Center, n.d.-c). User abuse is also forbidden by many platforms (Twitter, Facebook, Viber, Slack, WhatsApp), but its definition is often vague: Twitter Automation Rules (Twitter Help Center, n.d.-b), for instance, note that 'any automated activity that encourages, promotes, or incites abuse, violence, hateful conduct, or harassment' is forbidden, but do not go into detail. Similarly, Viber condemns activities which 'defame, abuse, harass, stalk, or threaten others' (Viber, n.d.-b) without defining what agent-based stalking or defamation might actually look like.

Besides spamming and user abuse, there are more niche forbinned activities that are specific to individual platforms. One such activity is 'surprising' other users, which is forbidden by Twitter and Discord. What constitutes a surprise is only loosely defined: Twitter states that automated activity 'should honor users expectations' (Twitter Help Center, n.d.-b), whereas Discord forbids data processing 'in a way that surprises or violates Discord users' expectations' (Discord, n.d.-a). Other agent-specific forbidden activities include unauthorized agent activity (Wikipedia), self-botting that is the attribution of the agent status to a human account for using agent API access (Discord (n.d.-c) and vice versa on Stack Overflow (n.d.-b) or not sufficiently contributing to the user experience (Slack, n.d.-b).

An interesting case is the misuse of agents to manipulate public opinion that is explicitly mentioned only in policy documents of Western messengers and social media platforms. It includes prohibition of using agents for manipulation (Roth and Pickles, 2020) and fake information dissemination (Facebook; Facebook Community Standards, n.d.; WhatsApp, n.d.). While the definition of manipulation lacks specificity, it highlights the detrimental impact of agents on the public sphere in line with associated moral panic. By contrast, other platforms (in particular, Telegram) focus on more individual-level transgressions.

Through our examination of policies and developer documentation, we also identified differences between the mechanisms used to identify forbidden activities. These mechanisms usually combine manual and automated approaches. The former include user reporting (Telegram, Wikipedia, Twitter, Discord and Stack Overflow) followed by verification of the report by human moderators (Telegram, Wikipedia). Interestingly, documents do not mention punishments for false reports or compensation for unjustified punishments. Only rarely do platforms note that such errors are possible (Twitter Help Center, n.d.-a).

Even less transparency is observed in how formal forms of governance refer to automated detection of infringements. Based on developer documentation, it can be assumed that some automated approaches are based on API requests rates with higher rates being indicative of misuse (Slack). Some platforms claim to use more advanced mechanisms, for example, spam filters (Twitter, Discord) or fake content detection (Facebook) or agent monitoring (Viber), but few details are disclosed. For instance, Discord (n.d.-d) documentation notes that joining multiple servers 'might be considered spam', but does not go into detail, whereas Facebook (Facebook Community Standards, n.d.) mentions machine learning-based mechanisms, but also does not provide details. Slack notes occasional audits to detect the involvement of agents in forbidden activities, but again does not provide details (Slack, n.d.-b).

Under the condition of non-transparency, the analysis of external developer documents associated with informal practices offers important insights into some pitfalls of formal governance forms. One of these is the omittance of some forms of agent misuse which are actually taking place. For instance, official documents usually ignore the use of agents as part of traditional cyber-security attacks (e.g. DDoS; one exception is Discord (n.d.-d)), whereas practice-related documents note them. Similarly, the abuses of governance rights (e.g. server banning raids on Discord; Reddit, 2020) are not covered by policies and mechanisms, but are acknowledged via practices.

In addition, practices highlight issues with the implementation of platform mechanisms detecting forbidden actions. A number of practice-related documents note the possibility of false positives, amplified by intentionally vague definitions of forbidden actions. Consider, for example, the following response by Twitter staff to a developer seeking clarification about the permissibility of his automated agent that posts updates from their website:

> There's no hard rule here, for many reasons, and unfortunately we are not able to provide a specific set of terms around this, because the systems are adaptive, and if there was no flexibility then it would be more straightforward for bad actors to determine ways to avoid the limits. (Twitter Developer Forum, 2020)

Sanctions for violating rules

Our analysis of official policy documents identified two approaches towards dealing with rule violations (Table 6 in Online Appendix): a universal one and a gradual one. The universal approach (Telegram, Discord, Viber, Stack Overflow and Wikipedia) treats all violations in the same manner and punishes them with the deactivation of the agent-associated account (Discord, n.d.-b; Wikipedia, n.d.-b) or limitation of its functionality (Telegram, n.d.-b). The gradual approach is utilized by major Western platforms (Facebook, Twitter) and Slack. Depending on the violation's severity, it assigns different sanctions ranging from warnings to the developer to limiting an agent's functionality (e.g. via anti-spam challenges or removal of the agent from the public access) to disabling the developer's account.

In theory, the gradual approach offers a more nuanced way of governance, in particular, as some violations can be caused by an agent malfunction (the possibility of non-intended violation, however, is noted only by Wikipedia and Stack Overflow, with the latter providing a 72-hour period for the developer to explain agent's actions). However, the policy documents are usually non-transparent about the relationship between violations and specific measures. For instance, Twitter (Twitter Help Center, n.d.-b) states that a developer account can be permanently suspended, but does not specify what infractions may not cause such a decision. Consequently, the gradual approach mainly leads to governance arbitrariness that is also noted in documents from external developer communities.

The lack of transparency on the policy level is matched by non-transparency of sanction mechanisms, in particular concerning the actor deciding on what sanctions to use. Only Telegram (n.d.-b) and Wikipedia (n.d.-b) communicate that sanctions are assigned by human moderators. Other platforms do not specify whether the decision is made manually or automatically that means that agents/developers can be subjected to automated judgement without their consent. This concern is amplified by practice-based observations noting multiple cases of non-justified sanctions. In some cases, it can be attributed to non-systematic errors (e.g. random account bans on Telegram (GitHub, 2018) or IP-based bans on Stack Overflow which might affect other users utilizing the same IPs (Stack Overflow, 2015), whereas in others (e.g. Discord or Twitter) it can be caused by systematic inconsistencies, such as sanctioning for infringements not listed in official documents.

The appeal procedures vary substantially between the platforms. In some cases (e.g. Telegram or Wikipedia), the procedure is specified in the official documents; that is, the user has to send a complaint to a specialized Telegram agent which then forwards it to human moderators (Telegram, n.d.-b). In other cases the procedure remains rather unclear and is mentioned generally, for example, as the possibility to approach the platform team (Twitter; Facebook) or the 72-hour appeal period for Stack Overflow. Such lack of clarity concerning appeal mechanisms amplifies power inequalities between developers and platforms in the context of sanction assignment.

An important component of sanction mechanisms is the question of developer accountability. For all platforms except Telegram and Viber the developer is held accountable for the agent's actions. On Twitter, for instance, the developer is declared to be 'ultimately responsible' (Twitter Help Center, n.d.-b) for all the agent's actions. Under these conditions, agents are treated as mere proxies of their creators, so their policing is approached as part of human agent governance. The rationale behind this approach is understandable in the absence of the universal structures of automated agent governance, but it also raises concerns both conceptually (i.e. do agent rights/obligations also directly apply to its developer) and practically (e.g. what if agent misuse is caused by the reasons that have nothing to do with the creator, such as a platform glitch).

Conclusion

In this article, we introduced an analytical framework for studying algorithmic governance approaches used by platforms and their users in relation to automated agents. We then applied this framework to compare how different platform contexts, ranging from messengers to knowledge platforms, shape approaches to governing activities of these autonomous software products. Our analysis highlights several important points about how different governance forms – namely, policies, mechanisms and practices – are used to deal with specific aspects of agent activities within platform ecosystems.

First, existing forms of governance acknowledge the diverse roles agents play within platform ecosystems. The extent of such acknowledgement varies between platforms, in particular on the level of formal governance structures. Some platforms, in particular non-Western (Telegram) and less business-oriented (Wikipedia, Discord) ones, are more upfront in acknowledging the significance of agents for their ecosystems, whereas others (Twitter, Facebook, WhatsApp) treat them more obscurely. However, independently of these differences, all examined platforms recognize that the need to govern agents goes beyond culling the spread of disinformation and opinion manipulation that until now seems to constitute the core of research on agent governance.

One particularly interesting aspect of this recognition relates to attribution of rights to automated agents by platforms. Codified via different forms of governance, these rights range from the acceptance of agents' right to communicate with humans to more specific rights including agents' ability to moderate activity of human and non-human users. In many cases, agent rights are formulated not as rights per se, but more as accepted uses and functional capabilities. The finding, nonetheless, is important for the ongoing debate about how certain (human) rights are attributed to automated entities and demonstrates how industrial/commercial practices proceed ahead of formal legal standards.

Second, we identified a number of deficits in how platforms approach agent governance. The lack of transparency, in particular regarding forbidden actions and sanctions, is not unexpected considering insights from human agent governance, but is still concerning. In particular, insights from developer practices highlight platforms' tendency to rely on collective punishment that can result in collateral damage to humans and non-humans. The situation, in which the platforms have the exclusive right to govern automated agents and their developers raises issues similar to those related to automated content moderation and its potential for undermining individual rights (Gorwa et al., 2020; Helberger, 2020).

The emphasis on top-down and often intransparent policing that is characteristic of how many platforms approach agent governance is concerning for several reasons. Besides increasing the power imbalance between platforms and human/non-human agents, it can be slow to accommodate for the new ways of using as well as misusing agents within platform ecosystems, which are integrated faster by less formal governance structures. These observations demonstrate the importance of going beyond the exclusive focus on agent policing and also considering alternative approaches to governance, such as promoting best practices of agent use or introducing community-based dispute resolution models similar to the ones used by Wikipedia.

Third, we observed cases where there are discrepancies between different forms of governance. That is especially the case in the context of practices and policies, where analysis of external developer documents highlights the mismatch between the two on several platforms (e.g. Discord, Facebook). Sometimes, platforms enforce sanctions for actions not explicitly forbidden in the policies, whereas some apparent forms of agent misuse are not covered by official documents. This underscores the troubling lack of transparency and the arbitrariness of certain governance decisions. Partially, such mismatches can be attributed to the lack of mechanisms enforcing platform accountability: it might be hard for a developer to sue a company such as Facebook for being barred from developer access. This stresses the necessity for more accountability mechanisms, such as a platform for documenting mismatches between forms of agent governance that can be used as a foundation for a legal action or more coherent legislation dealing with automated agent governance that will prevent privatization of this subject by platforms and will level the field for different parties involved.

Finally, we observe substantive differences in how agents are governed by the platforms. Besides varying degrees of transparency towards acknowledging the roles of agents platforms adopt divergent approaches towards agent rights/obligations and mechanisms of applying sanctions for rule infringement. While the size of our sample and the limitations of document analysis as an approach for studying agent governance restrict the interpretative value of our comparative analysis, it seems that the differences between different types of platforms (e.g. messengers or social media) are less consistent in this context than differences between more (e.g. WhatsApp, Facebook) and less business-oriented (e.g. Discord, Wikipedia) platforms as well as between Western and non-Western (e.g. Telegram) platforms. In particular, business-oriented platforms, which are related to large Western corporations, tend to be less transparent in terms of its punitive mechanisms and also less inclined to grant automated agents governance rights/enforce mechanisms for clear differentiation between human and non-human agents. Future research is required to better understand the rationale behind these differences as well as how consistent they are across a broader range of platforms.

Altogether, our observations stress the importance of expanding both the research agenda and societal debates on automated agents and their governance beyond the narrow focus on preventing public opinion manipulation and disinformation. Such an expansion is essential for understanding the complex role of agents within platform ecosystems better, but also for tackling other challenges, such as (re)defining the rights of both human and non-human users and developing a generalizable normative framework for their governance. It is also crucial that this expanded agenda will take into consideration the impact of different contextual factors on agent governance, which is the matter that is increasingly acknowledged by the broader field of algorithmic governance research (Gritsenko and Wood, 2022), that prompts the need for more comparative research.

There are several shortcomings of our article besides methodological limitations of relying on document analysis for studying algorithmic governance, which we noted earlier. While ambitious in conducting a comparative analysis of different forms of governance, the scope of our study does not yet constitute a comprehensive overview of less formal practice-based approaches to governance. Further research is needed to ascertain to what extent our findings can be generalized beyond our selection of documents characterizing developer practices in relation to agent governance. To a lesser degree, it also applies to our analysis of policies and mechanisms which can also be enhanced through a broader selection of documents and more detailed examination of the possible variation in the way documents codify governance structures depending on their target audience.

As already noted, our analysis primarily sheds light on how governance structures are intended to work, and further studies are needed to assess whether specified governance approaches are actually applied as intended. Our analysis of practice-oriented documents suggests that such gaps indeed exist, so future studies utilizing more qualitative techniques (e.g. interviews) or API audits (e.g. through agent-based testing) can bring important insights here. Another important limitation (as well as a direction for future research) concerns the temporal evolution of governance forms: in our study, we looked at the state of agent governance in 2020–2021, but did not investigate how it changed over time and how external factors could affect evolution of platforms' stances towards it. Finally, while most of the platforms we examined originate from the United States, it is important to expand the geographical scope of future studies, in particular as our observations suggest the possibility of differences between Western and non-Western platforms in terms of agent governance.

Authors' note

Matthias C Kettemann is now affiliated to Department of Theory and Future of Law, University of Innsbruck, Austria.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iDs

Mykola Makhortykh https://orcid.org/0000-0001-7143-5317

Aleksandra Urman https://orcid.org/0000-0003-3332-9294

Felix Victor Münch https://orcid.org/0000-0001-8808-6790

Supplemental material

Supplemental material for this article is available online.

Notes

1. We refer to ‘agents’ instead of ‘bots’ because the latter term evokes negative connotations associated with bots being discussed primarily in the context of opinion manipulation and disinformation (see Ferrara, 2017; Howard et al., 2018; Woolley, 2020).

2. Examples vary from correcting invalid HTML (Tsvetkova et al., 2017) and conversing with customers (Brandtzaeg and Følstad, 2017) to spreading spam (Cresci et al., 2018) and disinformation (Ferrara, 2017).
3. See, for instance, Duguay et al. (2020), Gorwa (2019), Gorwa et al. (2020), Hein et al. (2016).
4. To find relevant information, we first examined platform-specific developer forums (e.g. a dedicated Viber developer community forum) when such forums existed. However, because most platforms did not have such forums, we turned to external developer discussion boards such as Stack Overflow and communities such as r/webdev on Reddit. We queried them using general terms (e.g. 'bot'/'API' + platform name) to find general discussion of developer practices and with more context-specific queries (e.g. 'ban' + 'bot' + platform name) for information on specific aspects, such as forbidden actions and sanctions. We attempted to make our search as extensive as possible, however, given the volume of information and its scattered nature, we acknowledge that our current overview of user practices can hardly be viewed as comprehensive.

References

Assenmacher D, Clever L, Frischlich L, et al. (2020) Demystifying social bots: on the intelligence of automated social media actors. *Social Media+society* 6(3): 1–14.

Bloch-Wehba H (2020) Automation in moderation. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3521619 (accessed 8 January 2021).

Brandtzaeg P and Følstad A (2017) Why people use chatbots. In: Kompatsiaris I (ed.) *Internet Science*. Cham: Springer, pp. 377–392.

Cresci S, Petrocchi M, Spognardi A, et al. (2018) From reaction to proaction: unexplored ways to the detection of evolving spambots. In: *Companion Proceedings of the Web Conference 2018*, Lyon, 23–27 April, pp. 1469–1470. New York: ACM.

Discord (n.d.-a) Discord developer policy. Available at: http://tiny.cc/6155tz (accessed 8 January 2021).

Discord (n.d.-b) Discord terms of service. Available at: https://discord.com/terms (accessed 8 January 2021).

Discord (n.d.-c) OAuth 2. Available at: http://tiny.cc/2155tz (accessed 8 January 2021).

Discord (n.d.-d) Tips against spam and hacking. Available at: http://tiny.cc/4155tz (accessed 8 January 2021).

Duguay S, Burgess J and Suzor N (2020) Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence* 26(2): 237–252.

Facebook Community Standards (n.d.) Falsenews. Available at: http://tiny.cc/7155tz (accessed 8 January 2021).

Facebook for Developers (n.d.) Bots for workplace. Available at: http://tiny.cc/8155tz (accessed 8 January 2021).

Ferrara E (2017) Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* 22(8). Available at: https://firstmonday.org/article/view/8005/6516 (accessed 8 January 2021).

Fox A (2019) Automated political speech: regulating social media bots in the political sphere. *First Amendment Law Review* 18: 114–166.

Gallwitz F and Kreil M (2021) The rise and fall of 'social bot' research. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3814191 (accessed 10 May 2021).

Geiger R (2014) Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society* 17(3): 342–356.

Geiger R (2018) The lives of bots. arXiv. Available at: https://arxiv.org/ftp/arxiv/papers/1810/1810.09590.pdf (accessed 10 May 2021).

Gillespie T (2014) The relevance of algorithms. In: Gillespie T, Boczkowski PJ and Foot KA (eds) *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge, MA: MIT Press, pp. 167–194.

GitHub (2018) Phone number have been banned shortly after entering authentication code. Available at: https://github.com/tdlib/td/issues/312 (accessed 8 January 2021).

Gorwa R (2019) What is platform governance? *Information, Communication & Society* 22(6): 854–871.

Gorwa R, Binns R and Katzenbach C (2020) Algorithmic content moderation: technical and political challenges in the automation of platform governance. *Big Data & Society* 7(1): 1–15.

Gorwa R and Guilbeault D (2020) Unpacking the social media bot: a typology to guide research and policy. *Policy & Internet* 12(2): 225–248.

Gritsenko D and Wood M (2022) Algorithmic governance: a modes of governance approach. *Regulation & Governance* 16: 45–62.

Gunkel D (2018) *Robot Rights*. Cambridge, MA: MIT Press.

Hein A, Schreieck M, Wiesche M, et al. (2016) Multiple-case analysis on governance mechanisms of multi-sided platforms. In: *Proceedings of Multikonferenz Wirtschaftsinformatik*, Ilmenau, Germany, 9–11 March, pp. 9–11. Berlin: GITO-Verlag.

Helberger N (2020) The political power of platforms: how current attempts to regulate misinformation amplify opinion power. *Digital Journalism* 8(6): 842–854.

Howard P, Woolley S and Calo R (2018) Algorithms, bots, and political communication in the US 2016 election. *Journal of Information Technology & Politics* 15(2): 81–93.

Hukal P, Berente N, Germonprez M, et al. (2019) Bots coordinating work in open source software projects. *Computer* 52(9): 52–60.

Katzenbach C and Ulbricht L (2019) Algorithmic governance. *Internet Policy Review* 8(4): 1–18.

Kettemann M (2020) *The Normative Order of the Internet: A Theory of Rule and Regulation Online*. Oxford: Oxford University Press.

Lamo M and Calo R (2019) Regulating bot speech. *UCLA Law Review* 66: 988–1028.

Langford M (2020) Taming the digital leviathan: automated decision-making and international human rights. *AJIL Unbound* 114: 141–146.

Lessig L (1999) The law of the horse: what cyberlaw might teach. *Harvard Law Review* 113(2): 501–549.

Lior A (2020) AI Entities as AI agents: artificial intelligence liability and the AI respondeat superior analogy. *Mitchell Hamline Law Review* 46(5): 1–58.

Livingstone R (2016) Population automation: an interview with Wikipedia Bot pioneer ramman. *First Monday*. https://firstmonday.org/ojs/index.php/fm/article/download/6027/5189 (accessed 10 May 2021).

Lokot T and Diakopoulos N (2016) News bots: automating news and information dissemination on Twitter. *Digital Journalism* 4(6): 682–699.

Maréchal N (2016) When bots tweet: toward a normative framework for bots on social networking sites. *International Journal of Communication* 10: 5022–5031.

Martineau P (2019) What is a bot? *Wired*. https://www.wired.com/story/the-know-it-alls-what-is-a-bot/ (accessed 24 August 2021).

Pedrazzi S and Oehmer F (2020) Communication rights for social bots?: options for the governance of automated computer-generated online identities. *Journal of Information Policy* 10: 549–581.

Perel M and Elkin-Koren N (2019) Separation of functions for AI: restraining speech regulation by online platforms. *Lewis & Clark Law Review* 24(3): 857–898.

Rauchfleisch A and Kaiser J (2020) The false positive problem of automatic bot detection in social science research. *PLoS ONE* 15(10): 1–20.

Reddit (2020) Discord bot banned like half of our discord, what do we do? Available at: http://tiny.cc/7055tz (accessed 8 January 2021).

Roshchinskaya N (2020) How to create a Telegram chatbot for your business with SendPulse. Available at: http://tiny.cc/g055tz (accessed 8 January 2021).

Roth Y and Pickles N (2020) Bot or not? The facts about platform manipulation on Twitter. Available at: http://tiny.cc/h055tz (accessed 8 January 2021).

Seering J, Wang T, Yoon J, et al. (2019) Moderator engagement and community development in the age of algorithms. *New Media & Society* 21(7): 1417–1443.

Slack (n.d.-a) API Documentation Overview. Available at: https://api.slack.com/start/overview (accessed 10 May 2021).

Slack (n.d.-b) Slack App Developer Policy. Available at: https://api.slack.com/developer-policy (accessed 10 May 2021).

Stack Overflow (2015) Is my public IP banned from Stack Overflow? https://meta.stackexchange.com/questions/260168/is-my-public-ip-banned-from-stack-overflow (accessed 24 August 2021).

Stack Overflow (n.d.-a) Stack Exchange, Inc. API Terms of Use. Available at: https://stackoverflow.com/legal/api-terms-of-use (accessed 10 May 2021).

Stack Overflow (n.d.-b) Stack Exchange Network Acceptable Use Policy. Available at: https://stackoverflow.com/legal/acceptable-use-policy (accessed 10 May 2021).

Stoker G (1998) Governance as theory: five propositions. *International Social Science Journal* 50(155): 17–28.

Telegram (n.d.-a) Bots: an introduction for developers. Available at: https://core.telegram.org/bots (accessed 8 January 2021).

Telegram (n.d.-b) Spam FAQ. Available at: https://telegram.org/faq_spam (accessed 8 January 2021).

Telegram (n.d.-c) Telegram privacy policy. Available at: https://telegram.org/privacy (accessed 8 January 2021).

Timberg C and Dwoskin E (2018) Twitter is sweeping out fake accounts like never before, putting user growth at risk. *The Washington Post*. https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/ (accessed 24 August 2021).

Tsvetkova M, García-Gavilanes R, Floridi L, et al. (2017) Even good bots fight: the case of Wikipedia. *PLoS ONE* 12(2): 1–13.

Twitter Developer Forum (2020) Can you give me the specifics of the Twitter automation rules? Available at: https://twittercommunity.com/t/can-you-give-me-the-specifics-of-the-twitter-automation-rules/142034 (accessed 8 January 2021).

Twitter Help Center (n.d.-a) About suspended accounts. Available at: http://tiny.cc/o055tz (accessed 8 January 2021).

Twitter Help Center (n.d.-b) Automation rules. Available at: http://tiny.cc/s055tz (accessed 8 January 2021).

Twitter Help Center (n.d.-c) Platform manipulation and spam policy. Available at: http://tiny.cc/t055tz (accessed 8 January 2021).

Uyheng J and Carley K (2020) Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines. *Journal of Computational Social Science* 3(2): 445–468.

Viber (n.d.-a) Viber Developer Distribution Agreement or Viber API Terms of Service. Available at: https://developers.viber.com/docs/general/api-terms-of-service/ (accessed 10 May 2021).

Viber (n.d.-b) Viber Python Bot API. Available at: https://developers.viber.com/docs/api/python-bot-api/ (accessed 10 May 2021).

Walsh J (2020) Social media and moral panics: assessing the effects of technological change on societal reaction. *International Journal of Cultural Studies* 23(6): 840–859.

WhatsApp (n.d.) Terms of service. Available at: http://tiny.cc/y055tz (accessed 8 January 2021).

Wikipedia (n.d.-a) Help:Creatingabot. Available at: http://tiny.cc/1155tz (accessed 8 January 2021).

Wikipedia (n.d.-b) Wikipedia:Bot policy. Available at: http://tiny.cc/w055tz (accessed 8 January 2021).

Woolley S (2020) *Bots and Computational Propaganda: Automation for Communication and Control*. Cambridge: Cambridge University Press.

Zheng L, Albano CM, Vora NM, et al. (2019) The roles bots play in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 3: 1–20.

Author biographies

Mykola Makhortykh is a Postdoctoral Researcher at the Institute of Communication and Media Studies at the University of Bern. His research on, among others, news recommender systems, search engines and digital memory studies has appeared in the *European Journal of Communication*, *Internet Policy Review* and *New Media & Society*.

Aleksandra Urman is a Postdoctoral Researcher at the Social Computing Group, University of Zurich. Her PhD dissertation defended in May 2020 examines polarization on social media from a comparative perspective. Her research interests include online political communication, algorithmic biases and computational research methods.

Felix Victor Münch is an Early Career Researcher in computational social science with a PhD from the Digital Media Research Centre at QUT (Brisbane, Australia). Currently, he works as a Postdoc Researcher at the Leibniz Institute for Media Research | Hans-Bredow-Institut in Hamburg. His main fields of interest are currently network science methods, social media analytics and theories regarding the public sphere.

Amélie Heldt is a Legal Scholar and Doctoral Researcher at Leibniz Institute for Media Research in Hamburg where she focuses on the horizontal effect of freedom of expression in the digital sphere, platform regulation and social media governance.

Stephan Dreyer is Senior Researcher for Media Law and Media Governance and head of the research programme ‘Transformation of Public Communication’ at the Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg. His research focuses on the regulatory aspects of communication in a datafied society; currently, he is working on legal questions that arise with regard to algorithm-driven personalisation and information flows, automated decision-making systems in journalism and communicative (social) bots.

Matthias C Kettemann heads a research programme on online rules at the Leibniz Institute for Media Research | Hans-Bredow-Institut and coordinated research groups at the Humboldt Institute for Internet and Society, Berlin, and the Sustainable Computing Lab of the Vienna University of Economics and Business. His latest book, *The Normative Order of the Internet. A Theory of Rule and Regulation Online*, is available at Oxford University Press.