# The Effect of Learning Machines on Online Community Governance

*Regular Paper*

**Lei (Nico) Zheng**
Stevens Institute of Technology
Hoboken, NJ
lzheng9@stevens.edu

**Feng Mai**
Stevens Institute of Technology
Hoboken, NJ
fmai@stevens.edu

**Bei Yan**
Stevens Institute of Technology
Hoboken, NJ
byan7@stevens.edu

**Jeffrey V. Nickerson**
Stevens Institute of Technology
Hoboken, NJ
jnickers@stevens.edu

## Abstract

Automated machines that use statistical learning algorithms are increasingly used to govern online communities. Many studies conflate them with rule-based bots built on predefined conditions. This study theorizes and examines the consequence of adopting learning machines in online community's governance. Specifically, we look at how the adoption of a platform-wide, statistical learning-based anti-vandalism bot, ClueBot NG, affects the community's governance outcome: its vulnerability to vandalism (vandal attempts) and its effectiveness in fighting vandals (response time). We find that, compared with rule-based bots, the learning bot significantly reduced the number of vandalism attempts and improved the community's overall response time. In addition, the use of learning machines had second-order effects that created new challenges for community governance: after the adoption of the bot, vandals learned to attack in novel ways, and human editor's response time to revert a vandal attempt drastically increased. Our study provides new insights on how the learning machine's unique characteristics can reshape online community governance.

**Keywords:** Online Communities, Algorithmic Governance, Bots, Machine Learning, Threats, Vandalism

## Introduction

The sustainability and growth of online communities depend on their governance, which are systems for organizing the rules that regulate behavior (Forte et al. 2009). Traditionally, the governance of online communities relies on explicit rules (Butler et al. 2008), implicit norms (Faraj et al. 2011), and routines (Lindberg et al. 2016) to set boundaries, accommodate individual preferences and direct collective efforts. Recently, autonomous machines, also known as bots, are playing increasingly important roles in online community governance. For example, on Wikipedia, bots are deployed to revert vandalism, identify copyright violations, collect statistics, and organize discussion threads (Zheng et al. 2019). In other words, more online communities are now gradually transferring the "rights to rule" to machines (Müller-Birn et al. 2013).

Machines deployed in online communities differ in how they function. *Rule-based machines* are built on predefined rules and conditions. They can streamline repetitive tasks, integrate workflows, and reduce overall human efforts, but may generate high numbers of false positives and misleading feedback and require constant maintenance (Jhaver et al. 2019; Wessel et al. 2018). *Learning machines*, by contrast, are machines that use statistical learning methods to learn new rules from training data (Faraj et al. 2018). The key advantage of learning machines is that their decisions are based on generalizable patterns in existing training examples, making them more effective and robust when applied to unseen data (Abbasi et al. 2010).

However, learning machines' behaviors are also more opaque and less predictable because many statistical learning algorithms operate on high dimensional space and generate non-linear decision boundaries. The low interpretability may pose unexpected challenges to online community governance that relies on consensus-seeking and negotiation-based principles (Smith et al. 2020), as it is harder for humans to interpret and fix the errors produced by these learning machines (Jussupow et al. 2021). While some recent studies have looked at how rule-based machines are used to govern communities (He et al. 2021; Jhaver et al. 2019; Wessel et al. 2018; Zheng et al. 2019), much remains unknown about the implications of adopting learning machines. We beg the following research question: *Compared with rule-based machines, how does the adoption of learning machines affect online community governance outcomes?*

This study focuses on theorizing and examining the consequence of deploying learning machines on communities' governance outcomes. Specifically, we investigate the impact of adopting a platform-wide, statistical-learning based anti-vandalism bot, ClueBot NG, on community governance in Wikipedia. Since the main goal of Wikipedia governance is to ensure high-quality knowledge production (Kane and Ransbotham 2016), we examine the effect of applying the bot on the community's *vulnerability* to vandalism and its *effectiveness* at vandal-fighting. Our identification strategy exploits a quasi-experiment design: ClueBot NG replaced its rule-based predecessor on English Wikipedia in 2010, but not on French Wikipedia. Our analysis applies a difference-in-difference (DID) technique and compares 1,627,940 revisions of 264 pairs of the most vandalized pages from 2007 to 2018 on both the English and French Wikipedia.

The results reveal that overall, the adoption of the learning bot significantly reduced English Wikipedia's *vulnerability* to vandalism and improved its vandal-fighting *effectiveness* compared with rule-based bots, but increased human response time to vandalism. The adoption of the bot significantly reduced vandal attempts, human editors' workload on fighting vandalism, and the community's overall response time with respect to reverting vandal edits. Vandal attempts with a higher machine detection rate showed a bigger drop than attempts with a lower machine detection rate, suggesting that vandals also learned; they adapted by modifying their attacks. Importantly, the learning bot displaced and discouraged human vandal-fighters – their response time to vandalism increased more than threefold, potentially making Wikipedia more susceptible to subtle and novel attacks that only humans can currently identify.

Our study makes two contributions to the IS literature. First, we contribute to the online community governance literature. Extending prior studies that mainly focus on governance facilitated by humans (Forte et al. 2009; Butler et al. 2008; Joyce et al. 2013; Lindberg et al. 2016) or rule-based machines (He et al. 2021; Jhaver et al. 2019; Zheng et al. 2019), we distinguish learning machines from rule-based machines. We provide quantitative evidence of the effect of learning machines on the community's governance outcome. Second, our work also adds to the emerging field of algorithmic governance and AI use in organizations. Effects of learning machines on work and organizing have recently been studied in the context of traditional organizations whose incentives and command structures are well-defined (Faraj et al. 2018; Jussupow et al. 2021; Zhang et al. 2021). We extend this line of research by providing insights about practices and struggles to adopt learning machines in the fluid, decentralized online community context.

## Hypotheses Development

### *Machines' Effect on the Community's Vulnerability*

Compared with rule-based machines, learning machines may more effectively reduce online communities' vulnerability to online disruptive behavior for two reasons. First, learning machines may deter vandals by hampering their motivation and increasing their costs. Prior studies find that boredom, attention-seeking, and revenge motivate disruptive behavior in online communities (Shachaf and Hara 2010). However, after the adoption of learning machines, more vandal attempts are likely to be reverted promptly, therefore decreasing the psychological enjoyment and satisfaction for vandals. On the other hand, vandals may find their cost of vandalizing increased because they need to frequently switch IP addresses or register new accounts in order to surpass the bans issued by the bots. Second, learning machines may inhibit the social learning of vandals. Theory suggests that deviant behavior such as vandalizing, trolling, bullying, and harassment may be learned by observing or interacting with other deviants (Akers 2011; Lowry et al. 2016). Learning machines can inhibit this social learning process more effectively because a larger portion of vandal attempts can be quickly identified and removed, thus limiting their exposure to community members. Taken together, these theories suggest:

***H1:*** *The adoption of learning machines has a negative effect on the vulnerability of an online community, so the community receives fewer vandal attempts after it deploys learning machines.*

However, the effectiveness of learning machines on deterring vandals may correlate with the machine's detection rate over different types of disruptive behavior. Learning machines can have disproportionate detection rates on different types of vandalism – under-represented instances usually have a lower machine detection rate due to the biased distribution of the training dataset (Russell and Norvig 2002). Thus, vandals are motivated to vandalize in ways that are harder for machines to detect, thereby thwarting quick machine reversion (Abbasi et al. 2010; Barreno et al. 2010; Lappas et al. 2016).

The social learning process also plays a critical role in reinforcing the disproportionate effects of machines on different types of vandalism. Such learning involves both self-learning and peer learning (Cheng et al. 2017; Lowry et al. 2016). Through practice, vandals learn what kinds of vandalism are likely to be detected by machines. They then focus on vandalizing in alternative or novel ways to maximize their satisfaction. In addition, types of vandalism with a lower machine detection rate get a longer exposure time, which in turn helps train other vandals. In sum, as time goes on, types of vandalism that have a lower machine detection rate are likely to be used more frequently because vandals receive more immediate benefit, and because these types of vandalism remain on the site longer, provoking social learning from the other vandals.

***H2:*** *The effect of learning machines is positively associated with the machine's detection rate, leading to increased vandal attempts with lower detection rates and decreased vandal attempts with higher machine detection rates.*

### ***Machine's Effect on the Community's Effectiveness***

We also expect a significant boost in the community's effectiveness in countering vandalism after the adoption of learning machines. Compared with rule-based machines, learning machines have a higher vandalism detection rate, meaning that more vandal attempts are now reverted by machines than by humans. Moreover, machines act much faster than humans due to their scalability allows them to constantly patrol pages, monitor recent edits, and quickly revert identified vandal attempts (Geiger and Halfaker 2013; Rahwan et al. 2019). This may lead to a significant decrease in the system's average time-to-revert, thus increasing the community's overall effectiveness against vandalism.

***H3:*** *The use of learning machines has a positive effect on the effectiveness of an online community, so the community's time to revert a vandal attempt will decrease once it deploys learning machines.*

Nevertheless, the adoption of learning machines may generate both positive and negative impacts on the effectiveness of human editors. Delegating moderation tasks to machines would alleviate the heavy workload of human editors, making them less likely to burnout, which in turn may increase their productivity. Due to the high-frequency nature of online disruptive behavior, human moderators usually need to work non-stop (Gerrard 2020). Studies have shown that human moderators usually suffer from long-lasting psychological and emotional distress due to their repetitive, prolonged exposure to disturbing content such as hate speech, harassment, and trolling (Steiger et al. 2021). Recent research has found that by delegating easy and repetitive content moderation work to machines, human moderators could become more productive and less likely to experience burnout (He et al. 2021; Jhaver et al. 2019). This suggests:

***H4a:*** *The use of learning machines has a positive effect on the effectiveness of humans, so that human's time to revert a vandal attempt decrease once the community adopts learning machines.*

On the other hand, using machines for moderation may also generate detrimental effects on the effectiveness of human moderators for three reasons. First, past research regarding human-machine work allocation has demonstrated the potential for machines to substitute for humans in routine and standard jobs (Frank et al. 2019). This is also applicable to online community governance – with more work done by learning machines, human members may find it harder to detect disruptive content themselves and therefore reduce their effort to moderate community content. Second, studies concerning machine-assisted decision-making show that, with the support of learning machines, human decision-makers become reliant on machines' decisions without even noticing them (Jussupow et al. 2021). Similar situations may also happen in online communities. The adoption of machines may create a diffusion of responsibility, so that community members develop a reliance on machine governance and become less engaged in protecting communities' knowledge contents. Over time, this will likely lead to a significant reduction in the workforce of human vandal-fighters and/or reduction in individual effort on the task, which in turn decreases the overall effectiveness of human governance. This leads to an alternative hypothesis:

***H4b:*** *The use of learning machines has a negative effect on the effectiveness of humans, so that human's time to revert a vandal attempt increase once the community adopts learning machines.*

## Method

### *Empirical Strategy*

We employ a difference-in-differences (DID) technique, a quasi-experimental research design commonly used to estimate causal effects when randomized controlled trials are infeasible. The DID design resembles a randomized controlled trial in which there are both "treated" and "untreated" groups and uses post-treatment changes at the untreated group as a counterfactual to infer what would happen at the treated group had it not adopted the treatment.

We conduct our study in English and French Wikipedia. In November 2010, English Wikipedia adopted its first and only statistical-learning based anti-vandal bot ClueBot NG, replacing its old rule-based anti-vandal bot ClueBot. The bot ensembled neural network and naïve Bayesian classifiers, enabling it to catch approximately 40% vandalism (opposed to ClueBot's 5% detection rate) while maintaining a maximal false positive rate of 0.1%. The DID strategy is appropriate in our study because English and French Wikipedia face the threats of vandalism and share the same way of governing. Most articles in the English Wikipedia have French counterparts. Finally, both sites have a long history of using bots to assist human works (Geiger and Ribes 2010), but the learning machine, ClueBot NG, has only been active on English Wikipedia. Therefore, English Wikipedia serves as the "treated" site, and French Wikipedia serves as its "untreated" counterpart in our analysis.

### ***Data***

We collect 314 articles that were identified as the most vandalized pages in English Wikipedia.[1] For each article, we remove pages if we could not find their French counterparts. This procedure resulted in 289 article pairs, or 578 articles. To enable comparison of article status before and after the adoption of ClueBot NG, we further limit our sample to articles that were created at least two years prior to the bot's launch date and had edits in the two years

[1] https://en.wikipedia.org/wiki/Wikipedia:Most_vandalized_pages

following the launch date. Our final sample contains 264 article pairs, or 528 articles, and their full revision history between January 2007 and December 2018.

Following Kittur et al. (2007), we identify vandal attempts using the following approach. First, we find the revisions that reverted vandal attempts by matching keywords "vandal" or "rvv" (short for "revert due to vandalism") left in the revision comments. Second, as vandalism in Wikipedia is corrected by restoring the article to its last version prior to the vandal attempt, we compute the character-level Jaccard similarities between the identified revert revision and the ten revisions before it. Third, we set the revisions with the minimal Jaccard similarity as the original revision (the article was restored to this revision) and the revision immediately after the original version as the identified vandal attempt (vandalized version). Hence, for each vandal attempt, we obtain three article revisions that represent the article's original revision, vandalized revision, and reverted revision. We record whether the vandal attempt was reverted by a bot or a human editor. Moreover, we calculate the *time to revert* a vandal attempt as the time difference between the vandalized article revision and the reverted revision. Meanwhile, following Chin et al. (2010), we use a gradient boost classifier to classify vandal edits into one of three categories, including *Large-Scale Editing*, *Graffiti*, and *Misinformation* based on properties of the article's original version, vandalized version, and the difference between the two versions. Overall, our model has an average accuracy rate of 75.79%.

We construct a 12-year panel of yearly article-level data spanning two periods: the 4-year pre-treatment period before English Wikipedia adopted the protection of ClueBot NG, and the 8-year post-treatment period when the ClueBot NG was in place. The dataset includes vulnerability and effectiveness measures for 264 pairs of articles in both English and French Wikipedia sites between January 1, 2007 and December 30, 2018. An article-year observation includes the title of the article, the number of vandal attempts it received in that year, vandal attempts broken into three categories (Large-Scale Editing, Graffiti, and Misinformation), the number of vandal attempts reverted by human editors and bots, and other article characteristics.

### *Variables*

Two main dependent variables are *vulnerability* and *effectiveness*. *Vulnerability* is defined as the number of vandal attempts in an article in a year. *Effectiveness* is defined as the median value of time-to-revert that happened in an article in a year. Since these two variables are highly skewed, we transform these measures by taking their logarithm. We also control for the number of distinct editors, article anonymity, article popularity, article age, article length, and the number of talk page edits (Kittur et al. 2007). Table 1 summarizes the definition of these variables, and Table 2 shows the descriptive statistics of these variables in English and French Wikipedia before and after the adoption of ClueBot NG.

**Table 1. Variable Definition**

| Variables | Definition |
|---|---|
| $Vulnerability^*_{i,t}$ | The number of vandal attempts in article $i$ at year $t$. |
| $Effectiveness^*_{i,t}$ | The median value of time-to-revert (seconds) in article $i$ at year $t$. |
| $NumDistinctEditors^*_{i,t}$ | The number of distinct editors in article $i$ at year $t$. |
| $Anonymity$ | The percentage of anonymous editors in article $i$ at year $t$. |
| $ArticlePopularity^*_{i,t}$ | The Google trend index of the article title for article $i$ at year $t$. |
| $ArticleAge^*_{i,t}$ | The number of months article $i$ existed at the end of year $t$. |
| $ArticleLength^*_{i,t}$ | The number of characters for article $i$ at year $t$. |
| $NumTalkEdits^*_{i,t}$ | The number of revisions on the talk page associated with article $i$ at year $t$. |

Note: the variables marked with * are transformed by taking their logarithm (log(x+1)).

### *Empirical Specification*

We used a two-way fixed effect DID specification to estimate the impact of a learning machine on the dependent variables.

$$Y_{i,t} = \alpha + \beta_1 AdoptMachine_i + \beta_2 PostTreatment_t + \beta_3 AdoptMachine_i \times PostTreatment_t + Article_i + Year_t + Z_{i,t} + \epsilon_{i,t}$$

**Table 2. Descriptive Statistics at Article-Year Level**

| | Pre-Treatment | | | | | | Post-Treatment | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable** | **Obs** | **Mean** | **Std** | **Min** | **Max** | | **Obs** | **Mean** | **Std** | **Min** | **Max** |
| *Treatment Group (English Wikipedia)* | | | | | | | | | | | |
| Vulnerability | 1055 | 41.30 | 36.41 | 0.00 | 358.00 | | 2105 | 12.61 | 15.42 | 0.00 | 143.00 |
| Effectiveness | 1055 | 715.08 | 4373.13 | 0.00 | 81573.00 | | 2105 | 694.13 | 4508.78 | 0.00 | 68062.00 |
| NumDistinctEditors | 1055 | 388.81 | 267.67 | 1.00 | 2233.00 | | 2105 | 119.19 | 96.67 | 1.00 | 1407.00 |
| Anonymity | 1055 | 0.49 | 0.16 | 0.00 | 0.78 | | 2105 | 0.35 | 0.20 | 0.00 | 0.74 |
| ArticlePopularity | 1055 | 11.83 | 17.69 | 0.00 | 85.00 | | 2105 | 10.92 | 17.58 | 0.00 | 90.75 |
| ArticleAge | 1055 | 83.00 | 20.36 | 0.00 | 120.50 | | 2105 | 155.91 | 31.78 | 44.93 | 217.97 |
| ArticleLength | 1055 | 5428.24 | 3585.17 | 193.45 | 20189.25 | | 2105 | 8541.42 | 5673.57 | 1.00 | 33530.04 |
| NumTalkEdits | 1055 | 136.98 | 353.90 | 0.00 | 5259.00 | | 2105 | 60.54 | 468.53 | 0.00 | 13184.00 |
| *Control Group (French Wikipedia)* | | | | | | | | | | | |
| Vulnerability | 1035 | 3.98 | 7.25 | 0.00 | 86.00 | | 2080 | 2.24 | 3.96 | 0.00 | 37.00 |
| Effectiveness | 1035 | 1263.98 | 7070.51 | 0.00 | 86312.00 | | 2080 | 922.84 | 5568.20 | 0.00 | 85207.00 |
| NumDistinctEditors | 1035 | 56.45 | 52.07 | 1.00 | 391.00 | | 2080 | 32.60 | 32.92 | 1.00 | 313.00 |
| Anonymity | 1035 | 0.36 | 0.17 | 0.00 | 0.83 | | 2080 | 0.34 | 0.18 | 0.00 | 1.00 |
| ArticlePopularity | 1035 | 4.91 | 12.81 | 0.00 | 85.00 | | 2080 | 3.90 | 10.75 | 0.00 | 85.50 |
| ArticleAge | 1035 | 59.15 | 23.10 | 0.00 | 107.77 | | 2080 | 130.65 | 34.71 | 22.50 | 205.20 |
| ArticleLength | 1035 | 3192.59 | 3728.17 | 13.10 | 30180.30 | | 2080 | 5577.97 | 5732.44 | 76.00 | 36629.08 |
| NumTalkEdits | 1035 | 8.86 | 28.64 | 0.00 | 469.00 | | 2080 | 4.14 | 28.94 | 0.00 | 731.00 |

A critical assumption of DID approach is the parallel trend assumption, which requires the differences in the variable of interest are the same for both the treatment and control groups prior to the treatment (Gertler et al. 2016). Figure 1 shows the quarterly trend between the English and French Wikipedia for two dependent variables before and after the launch of the bot.

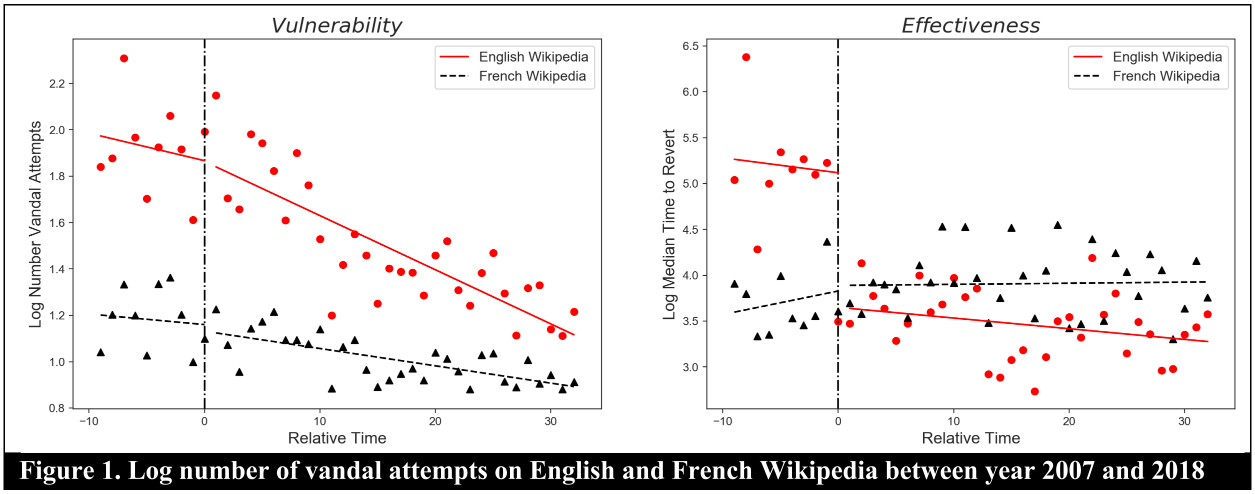

**Figure 1. Log number of vandal attempts on English and French Wikipedia between year 2007 and 2018**

Note: The dotted (solid) lines represent the fitted trend of the treatment group (English Wiki) using first-order polynomials. Each dot is the dependent variable at the article level averaged within a three-month window.

## RESULTS

**Table 3. Impact of Learning Machine on Vandal Attempts.**

| | **(1)** | **(2)** | **(3)** |
|---|---|---|---|
| **VARIABLES** | **NumVandalEdits** | **NumBotRevert** | **NumHumanRevert** |
| Adopt | -0.297*** | 0.593*** | -0.859*** |
| | (0.049) | (0.027) | (0.031) |
| LnNumVandalEdits | | 0.797*** | 0.690*** |
| | | (0.012) | (0.011) |
| Constant | -2.947*** | -0.342* | 0.023 |
| | (0.647) | (0.196) | (0.167) |
| Article Controls | Yes | Yes | Yes |
| Time Fixed Effect | Yes | Yes | Yes |
| Article Fixed Effect | Yes | Yes | Yes |
| Observations | 6,275 | 6,275 | 6,275 |
| R-squared | 0.681 | 0.838 | 0.845 |
| Article Groups | 528 | 528 | 528 |

Note: Robust standard errors clustered at article level are reported. *** p<0.01, ** p<0.05, * p<0.1

Table 3 presents the results on how the adoption of learning machines affects the behavior of vandals. In model 1, we find that the adoption of a learning machine reduced 29.7% vandal attempts per year in the treatment articles as compared to articles in the control group. In model 2 and model 3, we find that the number of vandal edits reverted by bots increased by 59.3% per year for articles in our treatment group as compared to the control group, and the number of vandal edits reverted by humans decreased by 85.9%. These results support H1. Table 4 examines the impact of learning machines on the type of vandal attempts. While there is a significant decrease in the number of vandal attempts across all vandal categories, the magnitude of the reduction correlates with the machine's detection rate of each vandal type. Specifically, we find that there is a 40.3% decrease in *Large-Scale Editing*, which is relatively easy for machines to detect. For more sophisticated types of vandal edits, the use of the learning machine only led to a 7.4% decrease in *Graffiti* and a 10.5% decrease in *Misinformation*. These ultimately resulted in an increase in more sophisticated types of vandal edits. We find the adoption of learning machines leads to a 12.2% increase in the proportion of *Graffiti* and a 7.2% increase in the proportion of *Misinformation*. These findings are consistent with H2.

**Table 4. Impact of Learning Machine on Vandal Attempts across Different Vandal Categories.**

| | **Number of Vandal Attempts** | | | **Proportion of Vandal Attempts** | | |
|---|---|---|---|---|---|---|
| **VARIABLES** | **Large-scale Editing** | **Graffiti** | **Misinformation** | **Large-scale Editing** | **Graffiti** | **Misinformation** |
| Adopt | -0.403*** | -0.074*** | -0.105*** | 0.007 | 0.122*** | 0.072*** |
| | (0.030) | (0.022) | (0.029) | (0.014) | (0.018) | (0.013) |
| LnNumVandalEdits | 0.601*** | 0.768*** | 0.572*** | 0.086*** | 0.142*** | 0.065*** |
| | (0.012) | (0.010) | (0.012) | (0.006) | (0.008) | (0.006) |
| Constant | -0.434** | -0.035 | -0.366*** | 0.062 | 0.357*** | 0.060 |
| | (0.176) | (0.143) | (0.123) | (0.089) | (0.127) | (0.075) |
| Article Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Time Fixed Effect | Yes | Yes | Yes | Yes | Yes | Yes |
| Article Fixed Effect | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 6,275 | 6,275 | 6,275 | 6,275 | 6,275 | 6,275 |
| R-squared | 0.752 | 0.867 | 0.707 | 0.127 | 0.162 | 0.062 |
| Article Groups | 528 | 528 | 528 | 528 | 528 | 528 |

Note: Robust standard errors clustered at article level are reported. *** p<0.01, ** p<0.05, * p<0.1

Table 5 shows the effects of learning machines on the community's effectiveness to revert a vandal attempt. The adoption of machines has significantly improved the community's overall effectiveness by reducing the response time to revert a vandal attempt by 84.9%. This finding is consistent with H3 and can be attributed to that a larger number of vandal edits being reverted by bots. In model 2, we do not find a significant effect of learning machines on the bot's response time. In model 3, we find the use of learning machines significantly *increases* the response time of human moderators by a magnitude of 339%. This finding supports Hypothesis 4b: learning machines have a negative impact on the effectiveness of human vandal-fighting.

**Table 5. Impact of Learning Machine on Response Time.**

| | **(1)** | **(2)** | **(3)** |
|---|---|---|---|
| **VARIABLES** | **Response Time** | **ResponseTime Bot** | **ResponseTime Human** |
| Adopt | -0.849*** | -0.033 | 3.391*** |
| | (0.156) | (0.059) | (0.223) |
| LnNumVandalEdits | 0.945*** | | |
| | (0.067) | | |

| | | | |
|---|---|---|---|
| LnNumBotRevert | | 0.763*** | |
| | | (0.031) | |
| LnNumHumanRevert | | | 2.469*** |
| | | | (0.091) |
| Constant | 3.270*** | -0.486 | 2.589* |
| | (1.207) | (0.421) | (1.369) |
| Article Controls | Yes | Yes | Yes |
| Time Fixed Effect | Yes | Yes | Yes |
| Article Fixed Effect | Yes | Yes | Yes |
| Observations | 6,275 | 6,275 | 6,275 |
| R-squared | 0.188 | 0.318 | 0.354 |
| Article Groups | 528 | 528 | 528 |
| Note: Robust standard errors clustered at article level are reported. *** p<0.01, ** p<0.05, * p<0.1 | | | |

## DISCUSSION AND CONCLUSION

Learning machines are increasingly used in the governance of online platforms to streamline workflow, encourage positive engagement, and reduce human effort in platform maintenance. Unlike traditional autonomous machines that use predefined rules, learning machines are capable of discovering, interpreting, modifying, and enforcing new rules that are opaque to humans. In this paper, we examined if and how such learning machines affect governance outcomes of online communities. We conducted a DID analysis of 264 pairs of most vandalized articles on English and French Wikipedia to quantitatively examine the effect of a learning machine on platform governance. We found that, compared to communities governed by rule-based bots, the learning bot significantly reduced the community's vulnerability to vandalism and improved its vandal-fighting effectiveness. The use of such machines also had second-order effects -- vandals learned to attack in novel ways, and it took longer for the community's human editors to identify and revert vandal attempts.

Our study has a number of implications. First, the creation and maintenance of a machine that learns is not just a software engineering challenge. These machines have impacts on online communities that go beyond statistical improvements to the tasks they perform. The machines are part of a system that depends on humans, who in turn adjust the machines' behavior in a continuous cycle of learning. This cycle is not always virtuous: outsiders may also learn and seek to disrupt the system. Thus, community governance includes remaining vigilant and proactively searching for external threats that disguise themselves as benign. Second, the use of learning machines can create human deskilling, because machines act more quickly, and the work that remains may be less rewarding. Communities face the challenge of communicating to their members about the machines' capability and limitations, with the goal of motivating humans to continue contributing and learning in the presence of machines that are also learning.

## Acknowledgement

This material is based upon work supported by the National Science Foundation under grants 2128906, 1909803, and 1717473.

## References

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., and Nunamaker, J. F. 2010. "Detecting Fake Websites: The Contribution of Statistical Learning Theory," *MIS Quarterly*(34:3), pp. 435–461.

Akers, R. L. 2011. *Social Learning and Social Structure: A General Theory of Crime and Deviance*.

Barreno, M., Nelson, B., Joseph, A. D., and Tygar, J. D. 2010. "The Security of Machine Learning," *Machine Learning* (81:2), pp. 121–148.

Butler, B., Joyce, E., and Pike, J. 2008. "Don'T Look Now, but We'Ve Created a Bureaucracy: The Nature and Roles of Policies and Rules in Wikipedia," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 1101–1110.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., and Leskovec, J. 2017. "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions," *Proceedings of the Conference on Computer-Supported Cooperative Work,* CSCW '17, pp. 1217–1230.

Chin, S.-C., Street, W. N., Srinivasan, P., and Eichmann, D. 2010. "Detecting Wikipedia Vandalism with Active Learning and Statistical Language Models," in *Proceedings of the 4th Workshop on Information Credibility*, WICOW '10, New York, pp. 3–10.

Faraj, S., Jarvenpaa, S. L., and Majchrzak, A. 2011. "Knowledge Collaboration in Online Communities," *Organization Science* (22:5), pp. 1224–1239.

Faraj, S., Pachidi, S., and Sayegh, K. 2018. "Working and Organizing in the Age of the Learning Algorithm," *Information and Organization* (28:1), pp. 62–70.

Forte, A., Larco, V., and Bruckman, A. 2009a. "Decentralization in Wikipedia Governance," *Journal of Management Information Systems* (26:1), pp. 49–72.

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., Wang, D., Youn, H., and Rahwan, I. 2019. "Toward Understanding the Impact of Artificial Intelligence on Labor," *Proceedings of the National Academy of Sciences of the United States of America* (116:14), pp. 6531–6539.

Geiger, R. S., and Halfaker, A. 2013. "When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes?," in *Proceedings of the 9th International Symposium on Open Collaboration*, WikiSym '13, pp. 6:1–6:6.

Geiger, R. S., and Ribes, D. 2010. "The Work of Sustaining Order in Wikipedia: The Banning of a Vandal," in

*Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pp. 117–126.

Gerrard, Y. 2020. "Behind the Screen: Content Moderation in the Shadows of Social Media," *New Media & Society* (22:3), pp. 579–582.

Halfaker, A., Geiger, R. S., Morgan, J. T., and Riedl, J. 2013. "The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline," *The American Behavioral Scientist* (57:5), pp. 664–688.

He, Q., Hong, Y., and Raghu, T. S. 2021. "The Effects of Machine-Powered Platform Governance: An Empirical Study of Content Moderation," *Available at SSRN*, January 17.

Jhaver, S., Birman, I., Gilbert, E., and Bruckman, A. 2019. "Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator," *ACM Trans. Comput.-Hum. Interact.* (26:5), pp. 1–35.

Joyce, E., Pike, J. C., and Butler, B. S. 2013. "Rules and Roles vs. Consensus: Self-Governed Deliberative Mass Collaboration Bureaucracies," *The American Behavioral Scientist* (57:5), pp. 576–594.

Jussupow, E., Spohrer, K., Heinzl, A., and Gawlitza, J. 2021. "Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence," *Information Systems Research: ISR* (isre.2020.0980).

Kane, G. C., and Ransbotham, S. 2016. "Research Note---Content and Collaboration: An Affiliation Network Approach to Information Quality in Online Peer Production Communities," *Information Systems Research* (27:2), pp. 424–439.

Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. 2007. "He Says, She Says: Conflict and Coordination in Wikipedia," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pp. 453–462.

Lappas, T., Sabnis, G., and Valkanas, G. 2016. "The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry," *Information Systems Research* (27:4), pp. 940–961.

Lindberg, A., Berente, N., Gaskin, J., and Lyytinen, K. 2016. "Coordinating Interdependencies in Online Communities: A Study of an Open Source Software Project," *Information Systems Research* (27:4), pp. 751–772.

Lowry, P. B., Zhang, J., Wang, C., and Siponen, M. 2016. "Why Do Adults Engage in Cyberbullying on Social Media? An Integration of Online Disinhibition and Deindividuation Effects with the Social Structure and Social Learning Model," *Information Systems Research* (27:4), pp. 962–986.

Müller-Birn, C., Dobusch, L., and Herbsleb, J. D. 2013. "Work-to-Rule: The Emergence of Algorithmic Governance in Wikipedia," in *Proceedings of the 6th International Conference on Communities and Technologies*, C&T '13, pp. 80–89.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A.' sandy', Roberts, M. E., Shariff, A., Tenenbaum, J. B., and Wellman, M. 2019. "Machine Behaviour," *Nature* (568:7753), pp. 477–486.

Russell, S., and Norvig, P. 2002. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, USA.

Shachaf, P., and Hara, N. 2010. "Beyond Vandalism: Wikipedia Trolls," *Journal of Information Science and Engineering* (36:3), pp. 357–370.

Smith, C. E., Yu, B., Srivastava, A., Halfaker, A., Terveen, L., and Zhu, H. 2020. "Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems," in *Proceedings of the 2020 CHI Conference*, CHI '20, April 21, pp. 1–14.

Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J., and Lease, M. 2021. *The Psychological Well-Being of Content Moderators*.

Wessel, M., de Souza, B. M., Steinmacher, I., Wiese, I. S., Polato, I., Chaves, A. P., and Gerosa, M. A. 2018. "The Power of Bots: Characterizing and Understanding Bots in OSS Projects," *Proc. ACM Hum. -Comput. Interact.* (2:CSCW), pp. 182:1–182:19.

Zhang, Z., Lindberg, A., Lyytinen, K., and Yoo, Y. 2021. "The Unknowability of Autonomous Tools and the Liminal Experience of Their Use," *Information Systems Research, forthcoming*,

Zheng, L. N., Albano, C. M., Vora, N. M., Mai, F., and Nickerson, J. V. 2019. "The Roles Bots Play in Wikipedia," *Proc. ACM Hum. -Comput. Interact.* (3:CSCW), pp. 1--20.