Automation Bias in Wikipedia: How Bots and AI Influence Knowledge Representation

Gabriel Giancarlo, Will Gatlin, Khalid AL-Mahmoud

November 5, 2025

Abstract

Wikipedia relies on a mixed ecosystem of human editors and automated tools (bots). As bots and generative-AI authored content increase, their role raises questions about neutrality, reliability, and bias: automated enforcement of citation rules or automated textual generation may introduce systematic patterns that look like "objective" fixes but in fact encode biases. This project seeks to measure how bot and AI-influenced edits are distributed across pages related to politically/socially charged topics, and to prototype tooling that surfaces bot activity and citation changes for human reviewers. Our analysis of 445 Wikipedia edits across controversial topics reveals that bots account for 15.5% of recent edits, with higher bot activity on controversial topics (up to 36.7% bot ratio). We identify systematic bias patterns including maintenance bias, citation bias, and controversial topic bias that demonstrate how automation influences what knowledge gets emphasized, repeated, or omitted on public platforms.

1 Introduction

Wikipedia has evolved into one of the most influential knowledge platforms globally, with over 6.7 million articles in English alone [?]. The platform's success relies on a complex ecosystem of human editors and automated tools (bots) working together to maintain content quality and consistency. However, as the role of automation in content creation and maintenance grows, concerns about bias and neutrality have emerged.

The problem of automation bias in knowledge platforms is increasingly relevant as AI-generated content becomes more prevalent. Recent studies have documented rising rates of AI-generated content in Wikipedia articles [?], raising questions about how automated systems may inadvertently reinforce or amplify existing biases in knowledge representation.

This research addresses three key questions about automation bias in Wikipedia:

1. **RQ1: Bot vs Human Edit Proportions** - What proportion of recent edits on controversial topics are made by accounts flagged as bots versus human editors?

2. **RQ2: Citation Pattern Differences** - How often do edits (bot or human) add or remove citations, and are bot edits more/less likely to change reference counts?

3. **RQ3: Systematic Bias Detection** - Could automated enforcement or AI-generated text unintentionally reinforce systemic bias through preferential source selection or stylistic patterns?

Our contributions include: (1) a systematic analysis of bot activity across controversial Wikipedia topics, (2) identification of measurable bias patterns in automated vs human editing behavior, and (3) development of tooling to surface automation bias for human reviewers.

2 Related Work

2.1 Bot Activity and Behavior on Wikipedia

Research on Wikipedia bots has revealed complex dynamics in automated content creation and maintenance. **?**] conducted a comprehensive analysis of bot interactions from 2001-2010, examining millions of edits across Wikipedia's history. Their study revealed that bots frequently undo each other's changes, creating complex edit wars that may amplify certain viewpoints through repeated automated enforcement. The authors found that even well-intentioned bots can create systematic patterns that influence content neutrality, providing a cautionary case for scaling automated enforcement without human oversight.

?] categorized bot tasks into distinct roles including maintenance, content creation, and quality assurance. They documented close human-bot collaboration patterns that shape content production, showing that bots often work in teams with human editors to maintain article quality. Their research identified that bots perform approximately one-third of all edits on Wikipedia, with significant variation across article categories. The study also highlighted how bot behavior patterns can influence which articles receive more maintenance attention, potentially creating systematic biases in content quality across different topics.

2.2 AI-Generated Content in Knowledge Platforms

Recent work has focused on detecting and analyzing AI-generated content in Wikipedia. **?**] documented rising rates of AI-generated content in recent English Wikipedia articles, finding that generative AI tools are increasingly being used to create and edit content. Their analysis revealed quality and bias concerns associated with automated content creation, including issues with source reliability, factual accuracy, and potential amplification of existing biases. The study raises important questions about how AI-generated content may differ from both human-authored and traditional bot-edited content in terms of bias patterns.

The proliferation of AI-generated content introduces new challenges for maintaining Wikipedia's neutrality and reliability standards. Unlike traditional bots that typically perform maintenance tasks, AI-generated content may include substantive text generation that requires different detection and analysis methods.

2.3 Automation Bias in Knowledge Systems

The broader literature on automation bias suggests that automated systems can introduce systematic patterns that appear objective but encode human biases [**?**]. In knowledge platforms where neutrality is paramount, automation bias can manifest through several mechanisms: (1) preferential enforcement of certain editorial policies, (2) systematic differences in which content areas receive automated attention, and (3) algorithmic amplification of certain viewpoints through repeated automated edits.

Research in human-computer interaction has shown that automation bias can lead users to overrely on automated systems and accept their outputs uncritically [**?**]. In the context of Wikipedia, this suggests that high bot activity on certain topics may create an appearance of consensus or quality that masks underlying bias patterns.

2.4 Citation Patterns and Source Quality

Studies of citation patterns in Wikipedia have revealed that bots and humans differ in their approach to citation management. Bots typically focus on technical maintenance tasks such as fixing broken links, standardizing citation formats, and removing duplicate references. However, these maintenance activities can inadvertently influence which sources are retained or removed, potentially creating systematic biases in source representation.

The quality and diversity of sources cited in Wikipedia articles is critical for maintaining neutrality. Research has shown that controversial topics often have more polarized citation patterns, with different editors preferentially citing sources that support their viewpoints [**?**]. The role of bots in managing these citations raises questions about whether automated maintenance may systematically favor certain types of sources or citation patterns.

2.5 Methodological Approaches to Bias Detection

Previous research on bias detection in Wikipedia has employed various methodologies including network analysis of editor interactions, content analysis of edit patterns, and statistical comparison of bot vs human editing behavior. However, few studies have specifically focused on automation bias as a distinct phenomenon that can be measured through systematic analysis of edit characteristics.

Our work builds on these methodological foundations while introducing new approaches to measuring automation bias through comparative analysis of edit patterns, citation changes, and content modifications across controversial topics.

3 Methodology

3.1 Research Design

Our study employs a quantitative analysis of Wikipedia edit patterns across controversial topics to measure automation bias. We selected five controversial topics (climate change, vaccination, artificial intelligence, gun control, and abortion) based on their high edit activity, public controversy, and relevance to current policy debates. For each topic, we analyzed the most recent 30-50 revisions from multiple related Wikipedia pages to capture a representative sample of recent editing activity.

The research design follows a comparative approach, analyzing differences between bot and human editing patterns across multiple dimensions: edit frequency, content size changes, citation modifications, and topic coverage. This multi-dimensional analysis allows us to identify systematic bias patterns that may not be apparent when examining any single metric in isolation.

3.2 Data Collection

We use the MediaWiki Action API (`https://en.wikipedia.org/w/api.php`) to programmatically fetch revision histories for Wikipedia pages on controversial topics. The data collection process is implemented in our main analysis script, available at github.com/GabeGiancarlo/EchoChamber/blob/main/src/we

The data collection workflow consists of the following steps:

1. **Page Search**: For each controversial topic, we search Wikipedia using the MediaWiki API's search endpoint to identify relevant pages. The search returns up to 5 pages per topic, prioritizing high-traffic articles that are likely to have recent editing activity.

2. **Revision Fetching**: For each identified page, we fetch the last 30-50 revisions using the API's revision history endpoint. Each revision includes metadata such as revision ID, timestamp, editor username, edit flags, content size, and the full article content.

3. **Data Storage**: Revision data is collected in real-time during analysis and stored in structured JSON format for further analysis. The complete data collection code includes error handling, rate limiting to respect Wikipedia's API usage policies, and data validation to ensure completeness.

The complete implementation of our data collection system is available in the repository at github.com/GabeGiancarlo/EchoChamber/tree/main/src, with detailed documentation in the code comments and README files.

3.3 Bot Detection Methods

We identify bot edits using a multi-layered heuristic approach that combines API flags, username patterns, and edit characteristics. The bot detection algorithm is implemented in the `is_bot_revision()` function, available at github.com/GabeGiancarlo/EchoChamber/blob/main/src/week7.py#L9
L112.

Our bot detection method uses three complementary approaches:

- **API Flags**: We check for explicit bot flags in the revision metadata provided by the MediaWiki API. The API includes flags and tags that indicate when an edit was made by an account with bot privileges.
- **Username Heuristics**: We identify bot accounts through username patterns, checking for common indicators such as "bot", "automated", "script", or "maintenance" in the editor username. This heuristic captures bots that may not have explicit API flags but are identifiable through naming conventions.
- **Edit Characteristics**: We analyze edit characteristics such as edit size, frequency, and content type. While not definitive, systematic patterns in these characteristics can help identify automated edits that may have been missed by other methods.

This multi-method approach provides robustness against variations in how bots are flagged or named across different Wikipedia pages. The implementation includes validation and testing to ensure accuracy, with detailed logging of bot detection decisions for manual verification.

3.4 Bias Detection and Measurement

We measure several types of potential automation bias through systematic analysis of edit patterns. The bias detection algorithms are implemented in the `detect_bias_patterns()` function at github.com/GabeGiancarlo/EchoChamber/blob/main/src/week7.py#L189-L256.

3.4.1 Maintenance Bias

Maintenance bias refers to systematic patterns in what types of content bots edit, particularly the tendency for bots to make smaller, more frequent maintenance edits rather than substantive content changes. We measure this by comparing the average edit size (in characters) between bot and human edits. The analysis code tracks size changes for each revision and calculates mean differences between bot and human editing patterns.

3.4.2 Citation Bias

Citation bias measures differences in how bots and humans modify citations in Wikipedia articles. We track citation changes by counting `<ref` occurrences in revision content and computing differences between successive revisions. The analysis compares citation addition/removal patterns between bot and human edits to identify systematic differences that may indicate bias.

The citation analysis is implemented in the `count_citations()` and `analyze_content_changes()` functions, which process revision content to extract citation patterns and track changes over time.

3.4.3 Controversial Topic Bias

We measure controversial topic bias by comparing bot activity ratios across different topics, with particular attention to topics that are known to be controversial or politically charged. Higher bot activity on controversial topics may indicate systematic bias in how automated systems maintain or moderate content on sensitive subjects.

The analysis identifies controversial topics through keyword matching (e.g., ”controversy”, ”debate”, ”denial”) and compares bot activity ratios to detect potential bias patterns.

3.4.4 Content Amplification Bias

Content amplification bias measures whether bots systematically contribute to large content additions or removals that may influence article emphasis. We track content size changes that exceed 50% of the previous revision size and analyze whether bots or humans are responsible for these significant content modifications.

3.5 Content Analysis Pipeline

The complete content analysis pipeline processes revision data through multiple stages:

1. **Content Extraction**: Revision content is extracted from the API response, handling different content formats and API versions.

2. **Change Detection**: For each revision, we compare content size, citation count, and other metrics to the previous revision to identify changes.

3. **Pattern Analysis**: We aggregate changes by editor type (bot vs human) to identify systematic patterns.

4. **Bias Calculation**: Multiple bias metrics are calculated and combined to identify potential automation bias patterns.

The complete analysis pipeline is documented in the repository at github.com/GabeGiancarlo/EchoChamber/tr with the main analysis logic in `week7.py` and supporting API server code in `week7_api.py`.

3.6 Web Interface and API

To facilitate interactive analysis and result visualization, we developed a web-based interface that provides real-time access to our analysis results. The Flask-based API server is available at github.com/GabeGiancarlo/EchoChamber/blob/main/src/week7_api.py and provides JSON endpoints for querying analysis results.

The web interface allows users to:

- Query specific topics for automation bias analysis
- View real-time results with bot activity ratios and bias indicators
- Access detailed per-page analysis with links to Wikipedia articles
- Explore citation patterns and content change metrics

The complete web interface code and documentation are available in the repository, including HTML templates and JavaScript for interactive visualization.

3.7 Data Validation and Quality Assurance

To ensure the reliability of our results, we implemented several validation measures:

- **API Response Validation**: All API responses are validated for completeness and correctness before processing.
- **Data Consistency Checks**: We verify that revision sequences are consistent and that content changes are properly tracked.
- **Manual Verification**: Sample results were manually verified by comparing our bot detection results against Wikipedia's revision history pages.
- **Result Verification Script**: A dedicated verification script (`verify_real_data.py`) confirms that analyzed pages exist and that results match expected patterns.

The validation scripts and procedures are documented in the repository, ensuring reproducibility and reliability of our analysis.

4 Results

Our analysis of 445 Wikipedia edits across five controversial topics reveals significant patterns in automation bias. The complete results dataset is available in JSON format at github.com/GabeGiancarlo/EchoChamb

4.1 Bot Activity Distribution

Overall, bots account for 15.5% of recent edits across all analyzed topics. However, this ratio varies significantly by topic:

- **Climate Change**: 16.7% bot ratio (15 bot edits out of 90 total edits across 3 pages)
- **Vaccination**: 22.2% bot ratio (20 bot edits out of 90 total edits across 3 pages)
- **Artificial Intelligence**: 14.4% bot ratio (13 bot edits out of 90 total edits across 3 pages)
- **Gun Control**: 14.1% bot ratio (12 bot edits out of 85 total edits across 3 pages)
- **Abortion**: 10.0% bot ratio (9 bot edits out of 90 total edits across 3 pages)

The variation in bot activity across topics suggests that automation is not uniformly applied, with some controversial topics receiving significantly more automated maintenance than others. The vaccination topic shows the highest overall bot ratio, which may reflect increased maintenance activity related to ongoing public health debates.

4.2 Page-Level Analysis

At the page level, we observed substantial variation in bot activity. For example, within the vaccination topic, the "Vaccine hesitancy" page showed a 36.7% bot ratio (11 bot edits out of 30 revisions), while the "Anti-vaccine activism" page showed only 6.7% bot ratio (2 bot edits out of 30 revisions). This variation within the same topic category suggests that bot activity may be influenced by specific page characteristics such as edit frequency, controversy level, or maintenance needs.

4.3 Bias Patterns Identified

Our analysis identified several distinct bias patterns:

4.3.1 Content Type Bias (Maintenance Bias)

Bots consistently make smaller edits on average compared to human editors. Across all topics, bot edits averaged 14-176 characters in size, while human edits averaged 186-1124 characters. This pattern indicates that bots primarily perform maintenance tasks such as fixing links, updating templates, or making minor formatting changes, rather than adding substantive content.

This maintenance bias creates a systematic difference in what types of content changes are made by automated systems versus humans, potentially influencing which aspects of articles receive more attention or maintenance.

4.3.2 Citation Bias

We observed different citation patterns between bots and humans. On pages with sufficient citation activity, bots showed different average citation deltas compared to human editors. For example, on the "Artificial general intelligence" page, bots showed an average citation delta of -1.0 (removing citations), while humans showed an average delta of 0.0 (maintaining citation counts).

This difference suggests that bots may be systematically removing or modifying citations in ways that differ from human editing patterns, potentially influencing which sources are retained or emphasized in articles.

4.3.3 Controversial Topic Bias

We found higher bot activity on pages explicitly related to controversial subtopics. The "Vaccine hesitancy" page showed 36.7% bot activity, significantly higher than the overall topic average. Similarly, the "Climate change denial" page showed 20% bot activity, higher than the main "Climate change" page (6.7%).

This pattern suggests that controversial or disputed topics may receive more automated maintenance attention, potentially as a response to higher edit activity or content disputes. However, this increased automation could also amplify certain viewpoints through repeated automated enforcement.

4.3.4 Content Amplification Bias

In some cases, bots were responsible for significant content amplifications (large content additions). For example, on the "Artificial general intelligence" page, bots were responsible for 100% of content amplifications (1 out of 1 amplification events), where amplification is defined as content size increases exceeding 50% of the previous revision.

While this represents a small number of events, it suggests that bots can contribute to substantial content changes, not just maintenance tasks, which could influence article emphasis and coverage.

4.4 Statistical Significance

While our sample size (445 edits across 15 pages) provides initial insights into automation bias patterns, larger-scale analysis would be needed to establish statistical significance for all observed differences. However, the consistent patterns observed across multiple topics and pages suggest that automation bias is a measurable phenomenon that warrants further investigation.

5 Discussion

5.1 Implications of Automation Bias

Our findings demonstrate that automation does influence what knowledge gets emphasized, repeated, or omitted on Wikipedia. The systematic differences between bot and human editing patterns reveal several important implications:

5.1.1 Maintenance Amplification

The consistent pattern of bots making smaller, more frequent edits creates a measurable bias toward certain types of content changes over time. While individual bot edits may be neutral and well-intentioned, the cumulative effect of automated maintenance can influence article emphasis and coverage. Articles with higher bot activity may receive more frequent maintenance attention, potentially creating systematic differences in content quality or freshness across topics.

5.1.2 Citation Patterns and Source Representation

The observed differences in citation patterns between bots and humans raise questions about how automated maintenance may influence source representation in Wikipedia articles. If bots systematically remove or modify citations in ways that differ from human patterns, this could affect which sources are retained or emphasized, potentially influencing article neutrality and source diversity.

5.1.3 Controversial Topic Bias

The higher bot activity on controversial topics suggests that automation may be responding to increased edit activity or content disputes. However, this increased automation could also amplify certain viewpoints through repeated automated enforcement of certain policies or standards. The 36.7% bot ratio on ”Vaccine hesitancy” pages is particularly notable, as it suggests that controversial health topics may receive disproportionate automated attention.

5.2 Limitations

Our study has several limitations that should be considered when interpreting the results:

- **Sample Size**: We analyzed 445 edits across 15 pages, which provides initial insights but may not be representative of all Wikipedia content. Larger-scale analysis would strengthen the generalizability of our findings.

- **Bot Detection Accuracy**: Our heuristic-based bot detection method may miss some bots or misclassify some human edits. While we use multiple detection methods to improve accuracy, manual verification of all results was not feasible.
- **Time Window**: We analyzed only recent revisions (last 30-50 per page), which captures current patterns but may not reflect historical trends or long-term bias patterns.
- **Topic Selection**: Our focus on controversial topics may limit generalizability to less controversial content areas where bot activity patterns may differ.

5.3 Future Work

Future research could expand on our findings in several directions:

- **Large-Scale Analysis**: Analyzing a much larger sample of Wikipedia edits across more topics and time periods would provide stronger statistical evidence for automation bias patterns.
- **AI-Generated Content Detection**: Extending the analysis to specifically detect and analyze AI-generated content, as distinct from traditional bot edits, would provide insights into how generative AI tools influence Wikipedia content.
- **Longitudinal Analysis**: Tracking bias patterns over time would reveal how automation bias evolves and whether certain patterns are temporary or persistent.
- **Source Domain Analysis**: Analyzing which source domains are added or removed by bots versus humans could provide insights into citation bias and source representation.
- **Intervention Studies**: Developing and testing tools to help human reviewers identify and address automation bias could provide practical solutions for mitigating bias in Wikipedia content.

6 Conclusion

This research provides evidence that automation bias exists in Wikipedia and can be measured through systematic analysis of edit patterns. Our analysis of 445 edits across controversial topics reveals that bots account for 15.5% of recent edits, with significant variation across topics and pages. We identified several measurable bias patterns including maintenance bias, citation bias, and controversial topic bias that demonstrate how automation influences knowledge representation.

The findings have important implications for understanding how automated systems influence knowledge platforms. As Wikipedia and similar platforms increasingly rely on automation for content maintenance and quality assurance, it is critical to understand how these automated systems may introduce systematic biases that affect what knowledge gets emphasized, repeated, or omitted.

Our work also demonstrates the feasibility of developing tooling to surface automation bias for human reviewers. The web-based interface and API we developed provide a foundation for tools that could help Wikipedia editors and administrators identify and address potential automation bias in articles.

As AI-generated content becomes more prevalent in knowledge platforms, the need for robust bias detection and mitigation strategies will only increase. Our research contributes to this effort by providing methods and tools for measuring and understanding automation bias in Wikipedia and similar collaborative knowledge platforms.

References