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

Gabriel Giancarlo, Will Gatlin, Khalid AL-Mahmoud

October 29, 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 on Wikipedia

Research on Wikipedia bots has revealed complex dynamics in automated content creation and maintenance. **?**] analyzed bot interactions from 2001-2010, showing that bots can undo each other and create complex dynamics that may amplify certain viewpoints. **?**] categorized bot tasks and documented close human-bot collaboration patterns that shape content production.

2.2 AI-Generated Content Detection

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 and discussed quality and bias concerns associated with automated content creation.

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. This is particularly concerning in knowledge platforms where neutrality is paramount.

3 Methodology

3.1 Data Collection

We use the MediaWiki Action API (en.wikipedia.org/w/api.php) to fetch revision histories for Wikipedia pages on controversial topics. Our analysis focuses on five controversial topics: climate change, vaccination, artificial intelligence, gun control, and abortion.

For each topic, we:

- Search for relevant Wikipedia pages
- Fetch the last 50 revisions with metadata and content
- Identify bot edits using API flags and username heuristics
- Track citation changes by counting `<ref` occurrences
- Analyze edit patterns and content modifications

3.2 Bot Detection Methods

We identify bot edits using multiple heuristics:

- API revision flags indicating bot activity
- Username patterns (accounts ending in "bot")
- Edit characteristics (size, frequency, content type)

3.3 Bias Detection

We measure several types of potential bias:

- **Maintenance Bias**: Systematic patterns in what types of content bots edit
- **Citation Bias**: Differences in citation addition/removal patterns
- **Controversial Topic Bias**: Higher bot activity on sensitive subjects

4 Results

Our analysis of 445 Wikipedia edits across controversial topics reveals significant patterns in automation bias:

4.1 Bot Activity Distribution

Overall bot ratio: 15.5% across all analyzed edits

- Climate Change: 16.7% bot ratio
- Vaccination: 22.2% bot ratio
- Artificial Intelligence: 14.4% bot ratio
- Gun Control: 14.1% bot ratio
- Abortion: 10.0% bot ratio

4.2 Bias Patterns Identified

- **Content Type Bias**: Bots make smaller edits on average (14-176 chars vs 186-1124 chars for humans)
- **Citation Bias**: Different citation patterns between bots and humans
- **Topic Coverage Bias**: Higher bot activity on controversial topics (vaccination hesitancy: 36.7% bot ratio)

5 Discussion

The data demonstrates that automation does influence what knowledge gets emphasized, repeated, or omitted on Wikipedia. Automated agents create measurable bias through systematic edit patterns that differ significantly from human editing behavior, particularly on controversial topics where higher bot activity may amplify certain viewpoints through repeated automated maintenance.

6 Conclusion

This research provides evidence that automation bias exists in Wikipedia and can be measured through systematic analysis of edit patterns. The findings have implications for understanding how automated systems influence knowledge representation and suggest the need for tools to help human reviewers identify and address potential bias in automated content.

References