

Moderation von Online-Diskussionen in digitaler Bürgerbeteiligung

In deutschen Kommunen liegt die Verantwortung für die Moderation von Online-Diskussionen im Rahmen digitaler Bürgerbeteiligung bei Menschen. Der Einsatz von Künstlicher Intelligenz (KI) wird noch vermieden. Ein hybrider Einsatz ist jedoch am vielversprechendsten.

Christian P. Hoffmann, Christian Pentzold und Thomas Feiler

Auf einen Blick

- Content Moderation umfasst das Prüfen, die Organisation und die Steuerung von Online-Diskussionen und kann helfen den Austausch unter Diskussionsteilnehmern zu verbessern.
- Content Moderation wird in mehr als der Hälfte der erfassten Verfahren digitaler Bürgerbeteiligung eingesetzt.
- In Verfahren mit Content Moderation wird diese am häufigsten durch Verwaltungsangestellte durchgeführt; KI wird noch recht selten eingesetzt.
- **Empfehlung:** Bei Content Moderation in digitaler Bürgerbeteiligung die Filterung und Erkennen toxischer Beiträge u. ä. KI überlassen, während komplexere Aspekte der Bürgerbeiträge weiterhin von Menschen bearbeitet werden.

Ausgangspunkt

Wenn sich viele Menschen mit unterschiedlichen Ansichten zu lokaler Politik austauschen – etwa im Rahmen einer Online-Diskussion als Teil eines Verfahrens zur möglichen Gestaltung eines Bahnhofsvorplatzes –, stoßen unterschiedliche Ideen aufeinander, es kann zu Reibungen, Widerspruch, aber auch Streit und sogar Beleidigungen kommen. Das kann den deliberativen Austausch zwischen Bürger:innen stark beeinträchtigen. Eine mögliche Lösung dafür ist **Content Moderation**, also die Pflege und Prüfung von Diskussionen.

Dabei gibt es unterschiedliche Formen der Content Moderation (siehe Infobox *Was ist Content Moderation?*). Im Rahmen der von dem Forschungsverbund ErLE durchgeführten Befragung wurden Kommunen danach befragt, ob und wie sie Content Moderation bei ihren zuletzt durchgeführten Verfahren eingesetzt haben. Die Antworten werden im Folgenden dargestellt und vor dem Hintergrund der Forschungsliteratur zu Content Moderation und KI kritisch betrachtet, um Handlungsempfehlungen für die Praxis der digitalen Bürgerbeteiligung abzuleiten.

Moderation – nach Kommunenart

Von den insgesamt 307 Verfahren wurde in **183 Verfahren** (59.61%) in irgendeiner Form **moderiert** – im Gegensatz dazu: In 124 Verfahren (40.39%) wurde keine Moderation eingesetzt.

	Moderation	Keine Moderation	Gesamt (100 %)
Landgemeinde (bis zu 5000 Einwohner)	29 (60.42%)	19 (39.58%)	48
Kleinstadt (zwischen 5000 und 20000 Einwohner)	52 (54.74%)	43 (45.26%)	95
Mittelstadt (zwischen 20000 und 100000 Einwohner)	73 (60.33%)	48 (39.67%)	121
Großstadt (ab 100000 Einwohner)	29 (67.44%)	14 (32.56%)	43

Lesehinweis: Die Angaben zeigen, in wie vielen Verfahren einer Kommunenart Moderation zum Einsatz kam.

Obwohl sich die Häufigkeiten des Einsatzes von Moderation zwischen den Kommunenarten auf den ersten Blick unterscheiden, zeigen inferenzstatistische Analysen: **Die Kommunengröße hat keinen signifikanten Einfluss darauf, ob Moderation eingesetzt wird oder nicht.**

Was ist Content Moderation?

Content Moderation bezeichnet sowohl die Aufstellung von Regeln als auch die **praktische Tätigkeit**, Beiträge in Online-Diskussionen zu prüfen, zu organisieren und zu steuern (Ma et al., 2023; Waterschoot, 2025). Ziel dieser Arbeit ist es, die Zusammenarbeit der Nutzenden in Online-Diskussionen zu fördern und Missbrauch – wie etwa bösartige Inhalte oder die Manipulation von Informationen – zu verhindern (Ma et al., 2023). Bei der **digitalen Bürgerbeteiligung** hilft Moderation dabei, digitale Räume zu schaffen, in denen ein sinn- und respektvoller wie auch konstruktiver Austausch möglich wird (Vasilakopoulos et al., 2024; Boudebouz et al., 2025).

Techniken: Restriktiv vs. Konstruktiv

Unabhängig davon, ob Menschen oder Maschinen moderieren, gibt es zwei Hauptstrategien (Yu et al., 2024):

- **Restriktive Techniken:** Dies ist das klassische „Eingreifen“, bei dem Beiträge gelöscht, verborgen oder Nutzende gesperrt werden (Yu et al., 2024). **Content Moderation durch KI** wird häufig für diese Form der schnellen Filterung von Beiträgen genutzt (Gillespie, 2020).
- **Konstruktive Techniken:** Diese gehen über das Löschen hinaus und versuchen, die Diskussion aktiv zu verbessern, indem etwa wertvolle Beiträge hervorgehoben oder Meinungen zusammengefasst werden (Yu et al., 2024). In der **digitalen Bürgerbeteiligung** wird KI bereits eingesetzt, um Argumente zu extrahieren oder die Konsensbildung in großen Gruppen zu unterstützen (Vasilakopoulos et al., 2024).

Vorteile von KI	Nachteile von KI
Hohe Skalierbarkeit: Bewältigung von zahlreichen Beiträgen in Echtzeit (Yu et al., 2024).	Mangelndes Kontextverständnis: Schwierigkeiten bei Sarkasmus oder kulturellen Feinheiten (Gillespie, 2020).
Entscheidungskonsistenz: Sorgt für eine gleichbleibende Anwendung von Regeln über große Datenmengen hinweg (Yu et al., 2024).	Systematische Fehler (Bias): Gefahr falscher Moderationsentscheidungen aufgrund der (ggf. vermuteten) Identität des Beitragsverfassers (Waterschoot, 2025; Oh & Downey, 2025).

Wer moderiert?

Blickt man nun darauf, wer die Moderation durchführt, ergibt sich folgendes Bild:

1. In **137 Verfahren (74.86%)** sind **Verwaltungsmitarbeiter** involviert;
2. in **67 Verfahren (36.61%)** sind **externe Dienstleister** involviert;
3. in nur **8 Verfahren (4.37%)** ist **Künstliche Intelligenz** involviert.

(Hinweis: Da in einem Verfahren mehrere Akteure im Einsatz sein können, übersteigt die Summe der Werte die Gesamtzahl von 183 Verfahren bzw. 100 %.)

In **155 (84.7%)** der moderierten Verfahren wird **Moderation in ausschließlich einer Form** durchgeführt – davon: 110 (60.11%) durch Mitarbeiter der Verwaltung, 41 (22.4%) durch externe Dienstleister und 4 (2.19%) durch Künstliche Intelligenz.

Bei den 27 Verfahren (14.75%), bei denen zwei Formen kombiniert werden, treten folgende Kombinationen auf:

- **24 Verfahren: Mitarbeiter der Verwaltung + externe Dienstleister**
- **2 Verfahren: Externe Dienstleister + Künstliche Intelligenz**
- **1 Verfahren: Mitarbeiter der Verwaltung + Künstliche Intelligenz**

In nur einem Verfahren wurde Moderation durch alle drei Formen gleichzeitig durchgeführt.

Handlungsempfehlungen

Einerseits wird Content Moderation in mehr als der Hälfte der Verfahren genutzt, andererseits jedoch nicht in allen – obwohl sie ein wirksames Instrument zur Verbesserung der Diskussionsqualität darstellt. Weiterhin wird Content Moderation (noch) überwiegend von Menschen durchgeführt, was aufgrund der Komplexität eines diskursiven Austausches notwendig, aber bei einer hohen Anzahl von Beiträgen sehr aufwändig sein kann.

Aus diesen Ergebnissen sowie dem Forschungsstand zum Gegenstand können folgende Handlungsempfehlungen abgeleitet werden:

Kommunalverwaltung: Expert:innen für Kontext und Qualität

Allgemein gelten **menschliche Akteure** bei der Moderation als unverzichtbar, wenn es darum geht, kommunikative Nuancen wie Sarkasmus, Ironie oder kulturelle Hintergründe zu verstehen (Gillespie, 2020; Waterschoot, 2025). Besonders bei der Identifizierung von konstruktiven Inhalten – also Beiträgen, die eine Diskussion voranbringen – sind menschliche Moderator:innen der KI überlegen (Waterschoot, 2025). Allerdings ist die rein von menschlichen Akteuren getragene Moderation arbeits-, zeit- und ressourcenaufwendig (Ma et al., 2023).

KI: Vorteil bei Skalierung und Mustererkennung

Die **Moderation durch KI** zeigt ihre Stärken überall dort, wo große Datenmengen in Echtzeit verarbeitet werden müssen, was für Menschen herausfordernd wäre (Gillespie, 2020; Yu et al., 2024). KI-Systeme sind besonders effizient darin, explizite Beleidigungen oder bekannte unerwünschte Inhalte durch Musterabgleich zu blockieren (Gillespie, 2020; Oh & Downey, 2025). Ein Schwachpunkt bleibt jedoch die Einordnung: KI-Tools neigen dazu, unhöfliche Sprache als „toxisch“ abzustrafen, während sie subtilen, „höflich“ formulierten Hass oft übersehen (Oh & Downey, 2025). Zudem besteht das Risiko, dass Algorithmen durch Vorurteile (Bias) systematisch die Stimmen von Minderheiten unterdrücken (Waterschoot, 2025; Yu et al., 2024).

Einsatz in der digitalen Bürgerbeteiligung

In der **digitalen Bürgerbeteiligung** ist der Diskurs oft politisch aufgeladen und komplex, weshalb für die abschließende Bewertung und die Konsensbildung meist menschliche Beteiligte (oft in Form von Moderatoren oder Freiwilligen) benötigt werden, um Legitimität und Vertrauen zu sichern (Mariani et al., 2025; Vasilakopoulos et al., 2024). Die KI wird hier strategisch für das „Screening“ eingesetzt, um ungewollte

Inhalte vorzusortieren oder um aus vielen Beiträgen automatisch Themencluster und Zusammenfassungen zu erstellen (Borchers et al., 2024; Mariani et al., 2025).

Die optimale Kombination: Das Hybrid-Modell

Die effektivste Strategie ist folglich eine **hybride Content Moderation**, bei der die KI die einfacheren Routineaufgaben übernimmt, indem sie zum Beispiel offensichtlich toxische Inhalte filtert und Massendaten sortiert, während der Mensch für die Feinarbeit zuständig bleibt (Gillespie, 2020; Waterschoot, 2025). In diesem Modell markiert die KI verdächtige Beiträge, aber die endgültige Entscheidung über eine Löschung oder die Kennzeichnung eines besonders wertvollen Kommentars treffen Personen (Waterschoot, 2025). In der digitalen Bürgerbeteiligung durch KI verspricht diese Kombination, dass Verwaltungen auch bei sehr hohen Teilnehmendenzahlen reaktionsfähig bleiben, ohne den demokratischen Kern einer individuellen Prüfung zu verlieren (Mariani et al., 2025).

Verantwortliche für digitale Bürgerbeteiligungsverfahren sollten strategisch entscheiden, welche restriktiven und konstruktiven Moderationstechniken (siehe Infobox: *Was ist Content Moderation?*) zu welchem Zeitpunkt sinnvoll sind und wer die Durchführung übernimmt. Folgende Leitfragen helfen bei dieser Entscheidung:

- Wie lassen sich zeitliche, finanzielle oder personelle Ressourcen schonen?
- Wie steigern wir die Qualität des Diskurses und des gesamten Verfahrens?
- Welche Entscheidungen kann KI zuverlässig treffen – und wo bleibt menschliche Moderation unverzichtbar?

Zur Datenbasis und Methode

Datenbasis. Vollerhebung unter deutschen Gemeinden (N = 10.751) zum Angebot digitaler Bürgerbeteiligung; Erhebungszeitraum: 11. März 2025 bis 10. November 2025; realisierte Stichprobe $n = 2.389$ (Rücklauf 22,2 %). Der Fragebogen bestand aus zwei Teilen: Teil 1 befasste sich mit der Kommune und der Situation der digitalen Bürgerbeteiligung; Teil 2 wurde nur an die Kommunen weitergeleitet, die angaben, dass sie Erfahrung mit digitaler Bürgerbeteiligung haben, und befasste sich mit den letzten beiden durchgeführten Verfahren. Die Befunde dieses Policy Briefs basieren auf Daten des zweiten Fragebogenteils.

Im Rahmen der deutschlandweiten Befragung konnten Kommunen Angaben zu ihren zuletzt durchgeführten Verfahren digitaler Bürgerbeteiligung machen. Somit haben wir Informationen über 307 Verfahren erhalten; davon stammen 125 Verfahren (40,72 %) aus Kommunen, die Fragen zu nur einem Verfahren beantwortet haben, während sich 182 Verfahren (59,28 %) auf die 91 Kommunen verteilen, die Antworten zu jeweils zwei Verfahren gegeben haben.

Detaillierte Daten siehe:

Kneuer, Marianne, Hoffmann, Christian, Marschall, Stefan, Pentzold, Christian, Wimmer, Maria, Donix, Johannes, Feiler, Thomas, Pestow, Radomir, Stock, Bastian & Wilker Nadja. 2026. *Facilitating and hindering factors of digital participation at the local level*. 2026. Data Report. OSF. DOI: [10.17605/OSF.F.IO/FQNV7](https://doi.org/10.17605/OSF.F.IO/FQNV7)

Christian P. Hoffmann, Christian Pentzold & Thomas Feiler. Moderation von Online-Diskussionen in digitaler Bürgerbeteiligung

. Policy Brief, Juli 2026. Leipzig

Kontakt: kontakt@lokale-partizipation.de - Projekt *Erfolgsfaktoren lokaler E-Partizipation*

<https://lokale-partizipation.de/> - gefördert durch die Stiftung Mercator.

Literatur

- Borchers, M., Soroko, D., Tavanapour, N. & Bittner, E. (2024). Hear Me Out: Supporting Citizens to Create Comprehensible Contributions on Urban Participation Platforms. *Proceedings Of The ACM On Human-Computer Interaction*, 8(CSCW2), 1–26. <https://doi.org/10.1145/3686971>
- Boudebouz, A., Carvalho, J. A., & Tavares, A. F. (2025). e-Participation for community empowerment: A systematic literature review. *Transforming Government: People, Process and Policy*. <https://doi.org/10.1108/TG-01-2025-0008>
- Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2), 1–5. <https://doi.org/10.1177/2053951720943234>
- Ma, R., You, Y., Gui, X., & Kou, Y. (2023). How do users experience moderation?: A systematic literature review. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), Article 278. <https://doi.org/10.1145/3610069>
- Mariani, I., Mortati, M., Rizzo, F., & Deserti, A. (2025). *Design thinking as a strategic approach to e-participation: From current barriers to opportunities*. Springer Nature. <https://doi.org/10.1007/978-3-031-72160-1>
- Oh, D., & Downey, J. (2025). Does algorithmic content moderation promote democratic discourse? Radical democratic critique of toxic language AI. *Information, Communication & Society*, 28(7), 1157–1176. <https://doi.org/10.1080/1369118X.2024.2346531>
- Vasilakopoulos, Z., Tavantzis, T., Promikyridis, R., & Tambouris, E. (2024). The use of artificial intelligence in eParticipation: Mapping current research. *Future Internet*, 16(6), 198. <https://doi.org/10.3390/fi16060198>
- Waterschoot, C. (2025). Governing the “third half of the internet”: The dynamics of human and AI-assisted content moderation. In J. Van Dijck, K. van Es, A. Helmond, & F. van der Vlist (Eds.), *Governing the digital society: Platforms, artificial intelligence, and public values*. Taylor & Francis Group. https://doi.org/10.5117/9789048562718_ch03
- Yu, Z., Otto, L., Assenmacher, D., & Wagner, C. (2024). A systematic review of the effects of AI-assisted moderation on individuals and groups. *Human-Machine Communication*, 9, 167–191. <https://doi.org/10.30658/hmc.9.10>