MACHINE AGAINST THE RAGE
Magazin für digitale Konfliktforschung

Essentials
Nr. 8 | 2025

Nr. 8

Kollektives Faktenprüfen?

Community Notes und der Umgang mit Desinformation

Community Notes (CN) werden von US-Tech-Unternehmen als adäquates Mittel gepriesen, um schädliche und irreführende Inhalte auf digitalen Plattformen zu bändigen. Doch wie funktioniert diese crowdbasierte Form des Fact-Checking? In diesem Policy Brief erklären wir, wie jene kollektiven Anmerkungen im deutschsprachigen Kontext angewandt werden, welche Themen sie ansprechen und welche Akteure betroffen sind. Nach der Auswertung tausender CN auf X lässt sich feststellen, dass damit Desinformation kaum Einhalt geboten wird.

In aller Kürze:

- CN entwickelten sich von einem Experiment mit ausgewählten UserInnen zu einem massenhaft genutzten Tool, um irreführende Infos richtigzustellen.
- Die EU-Kommission untersucht derzeit, inwiefern CN Faktenchecks ersetzen können, um systemische Risiken wie Desinformation eingrenzen zu können.
- CN sind oft faktenbasiert, verweisen auf seriöse Quellen, werden aber nur spärlich, spät und selektiv veröffentlicht; es fehlen Anreize zur Teilnahme.
- CN sind Teil geopolitischer Konflikte über Plattformregulierung; es geht nicht nur um technische Lösungen – sondern auch um den Streit um Wahrheit.

IDZ INSTITUT FÜR DEMOKRATIE UND ZIVILGESELLSCHAFT

Datenbasis

- Grundlage der Analyse waren 35.508 deutschsprachige Community Notes (CN), die seit 2021 verfasst wurden und sich auf 22.545 Posts bezogen.
- Von den erhobenen CN wurden nur 2.228 (6%) als hilfreich bewertet und somit veröffentlicht. 1.066 (3%) wurden als nicht hilfreich eingeschätzt, 32.214 (91%) erhielten nicht genügend Stimmen und bleiben verborgen.
- Es werden immer weniger CN pro Monat verfasst (im Juni 2025: 1.029 deutschsprachige). Die Zeit zwischen Veröffentlichung des Posts und der öffentlichen Darstellung der CN beträgt im Mittel 19,6 Stunden.

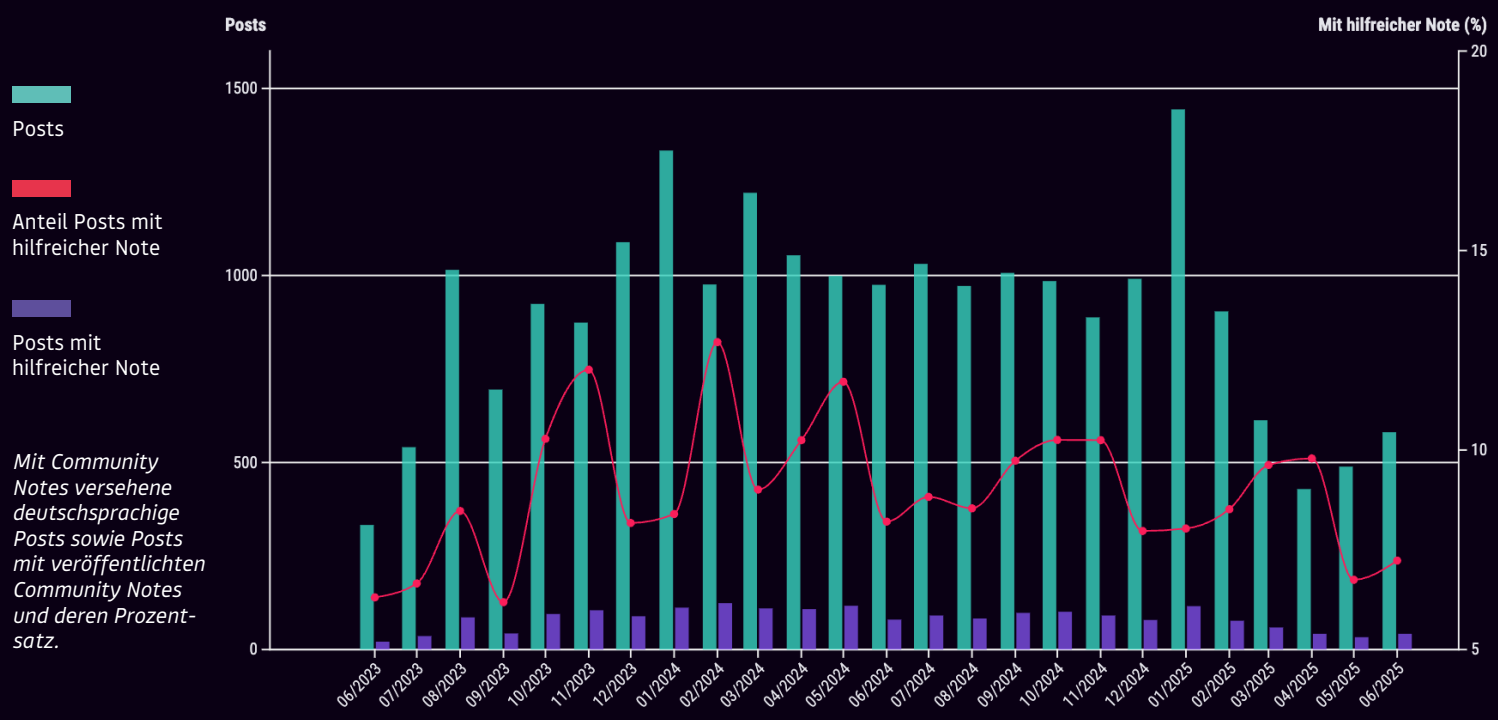
Im Januar kündigte Meta-Chef Mark Zuckerberg an, das Konzept der Community Notes, das seit nun einigen Jahren auf der Plattform X erprobt wird, für Facebook zu übernehmen. Das System, mit dem NutzerInnen über die Richtigkeit bzw. Seriosität von Inhalten urteilen sollen, wird von manchen als demokratische Alternative zum Fact-Checking gefeiert, wohingegen andere darin ein Feigenblatt für fehlende Standards der Inhaltsmoderation sehen. Das Prinzip dahinter ist, dass NutzerInnen der Plattform abstimmen können, ob Anmerkungen angebracht sind, um einen Post auf der Plattform zu berichtigen oder mit Kontext zu ergänzen. Ziel sei es, eine informiertere Debatte zu initiieren und irreführende Informationen zu markieren.

2021 initiierte die Plattform Twitter das Programm Birdwatch, wo ausgewählte NutzerInnen den Wahrheitsgehalt von Tweets kommentieren konnten. Veröffentlicht wurden jene Kommentare, die von verschiedenen politischen Gruppen als hilfreich eingeschätzt wurden. Diese identifizierte ein Algorithmus auf Grundlage des Abstimmungsverhaltens. Der Zeitpunkt der Einführung war nicht zufällig gewählt. Die Corona-Pandemie und die polarisierten Debatten hatten sich nachhaltig negativ auf das Vertrauen der BürgerInnen in Informationen und Institutionen ausgewirkt. Auch Faktenchecks waren verdächtig, da der Anspruch auf Wahrheit nicht von allen gleichsam geteilt wurde.

Mit der Übernahme der Plattform durch Elon Musk 2023 weitete dieser das Programm zu einer Alternative zum Fact-Checking aus. Community Notes sollten eine Ergänzung zur Moderation und eine kollaborative Form der Richtigstellung von Falschinformationen sein. Diese Praxis wurde vermehrt angenommen: Über zwei Millionen Community Notes wurden geschrieben – Tendenz fallend. Während im Januar 2025 noch 120.000 Community Notes geschrieben wurden, waren es im Juni nur noch die Hälfte. Was können Community Notes also? Und sind sie zur Eindämmung von Desinformation geeignet?

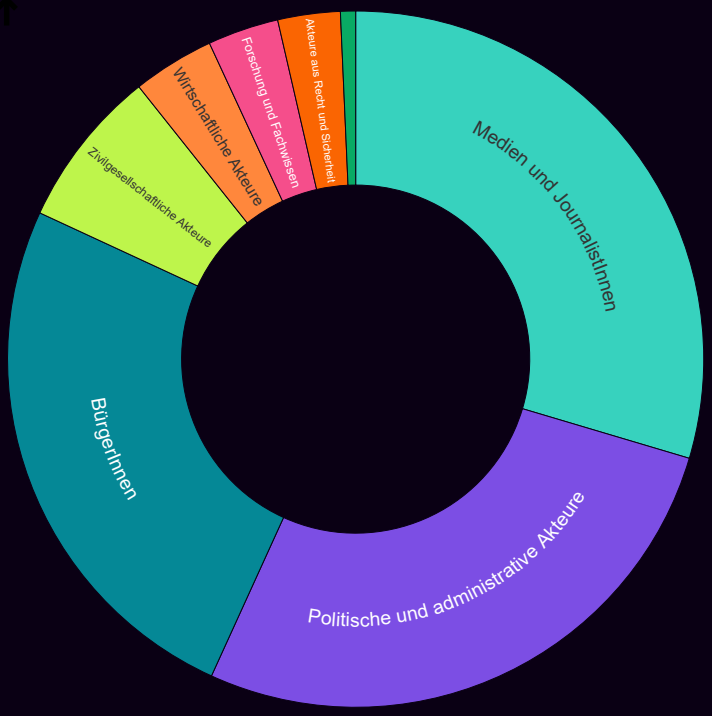
Die Logik des kollektiven Faktenchecks

Community Notes sind ein partizipatives System zur Kontextualisierung von Inhalten auf X (ehemals Twitter), das sich von professionellen Faktenchecks unterscheidet. Statt festen Regeln folgen die Beiträge individuellen Schreibstilen und subjektiven Wahrheitsmaßstäben der Nutzer-



Mit Community Notes versehene deutschsprachige Posts sowie Posts mit veröffentlichten Community Notes und deren Prozentsatz.

»Am häufigsten werden Posts wegen fehlenden Kontexts oder faktischer Fehler korrigiert, während die meisten veröffentlichten Community Notes manipulierte Medien betreffen und die wenigsten ungeprüfte Behauptungen.«



Innen. Ziel ist es, kurz, sachlich und mit Quellen belegt auf Posts zu reagieren, um Falschinformationen einzuordnen. Um selbst Notes verfassen zu dürfen, müssen NutzerInnen zunächst durch hilfreiche Bewertungen anderer Notes ihre Vertrauenswürdigkeit nachweisen – ein Prozess, der Manipulationen und koordinierte Kampagnen verhindern soll.

Bei der Beanstandung eines Posts wählen NutzerInnen zunächst aus, welcher Art der Verstoß ist, etwa fehlender Kontext, faktische Fehler, veraltete Informationen, manipulierte Medien oder Satire. Am häufigsten werden Posts wegen fehlenden Kontexts oder faktischer Fehler korrigiert, während die meisten veröffentlichten Community Notes manipulierte Medien betreffen (18,3 Prozent) und die wenigsten ungeprüfte Behauptungen (7,1 Prozent).

Die Auswertung zeigt, dass die NutzerInnen ihre Richtigstellungen und Aussagen in der Regel mit Quellen belegen. In mehr als einem Viertel der Anmerkungen (28 Prozent) wird sogar mehr als eine Quelle verlinkt, in nur zwei Prozent der Fälle wird keine Quelle genannt. Am häufigsten verwendet werden – bei 61 Prozent der Fälle – Quellen aus dem Bereich Nachrichten und Medien, also journalistische Angebote. Danach folgen Social-Media-Quellen (23 Prozent), mit X selbst an der Spitze, und Verweise auf Behörden, Verwaltung und internationale Institutionen (20 Prozent). In absoluten Zahlen der einzelnen Domains liegt Wikipedia auf dem ersten Platz, gefolgt von X selbst und dem Account der Tagesschau.

Akteure und Themen der Community Notes

Die Analyse der Community Notes auf X zeigt, dass mehr als 60 Prozent der damit versehenen Posts von nur wenigen Accounts stammen. Besonders betroffen sind Akteure aus Medien und Journalismus (29 Prozent), Politik (27 Prozent) sowie BürgerInnen ohne Mandat (24 Prozent). Andere Bereiche wie Recht, Wirtschaft oder Kultur sind deutlich seltener vertreten. Auffällig ist, dass betroffene Accounts eine hohe Reichweite haben (im Durchschnitt: 159.550 FollowerInnen), was Sichtbarkeit und Korrekturbedarf verstärkt. Innerhalb der Mediengruppe fallen vor allem einzelne JournalistInnen, Blogger/Vlogger sowie einige wenige Leitmedien wie Tagesschau, ZDFheute oder Der Spiegel auf.

Bei politischen Akteuren stechen Markus Söder, Alice Weidel, Karl Lauterbach und andere hervor; bei veröffentlichten Notes verschiebt sich die Gewichtung leicht. Ideologisch werden besonders oft Akteure aus dem rechten bis rechtsextremen Spektrum adressiert, während ökologische, linke oder liberale Stimmen seltener auftreten – auch

Es handelt sich hier um die kondensierte Version des Themenschwerpunkts von *Machine Against the Rage*, Nr. 8 (Spätjahr 2025): Michael Schmidt, Franziska Martini, Maik Fielitz & Christian Donner, »Der Schwarm als Faktenchecker. Community Notes zwischen kollektiver Intelligenz und politischem Lagerkampf«.

Online weiterlesen – mit mehr Analysen, interaktivem Dashboard und methodischem Annex:

www.machine-vs-rage.net

MACHINE AGAINST THE RAGE

wegen des Rückzugs vieler Accounts. Thematisch betreffen Community Notes vor allem Klima und Energie (12 Prozent), gefolgt von deutscher Politik (8 Prozent), Kriegen (Ukraine, Gaza/Israel) und Wirtschafts- bzw. Sozialpolitik.

Die Plattform X stilisiert Community Notes als einen Mechanismus, um die eigenen NutzerInnen vor systemischen Risiken wie Desinformation zu schützen. Dass viele allerdings nicht annähernd in die Richtung eines Faktenchecks oder einer Kontextualisierung gehen, wird schnell deutlich, wenn Community Notes mit Blick auf ihre Zweckmäßigkeit bewertet werden. Nur 1.066 der deutschsprachigen Community Notes wurden als nicht hilfreich eingeordnet. Inwiefern die weiteren 32.214 nicht veröffentlichten Community Notes hingegen potenziell hilfreich sind oder nicht, lässt sich zunächst nicht sagen.

Ein probates Mittel gegen Desinformation?

Sofern Community Notes nützlich im Umgang mit fehlerhaften oder irreführenden Informationen sein sollten, müssten sie demnach entweder versuchen, Behauptungen zu widerlegen oder einen wichtigen Kontext hinzuzufügen. Allerdings fällt in unserer Untersuchung auf, dass eine Mehrheit der Community Notes gar nicht dem Anspruch standhält, Aussagen richtigzustellen, sondern am häufigsten andere Community Notes selbst kommentieren und auf deren eigentliche Funktion hinweisen. Diese Community Notes werden oft mit dem Hinweis NNN (Note Not Needed) versehen und sollen eine vorherige Community Note als überflüssig oder illegitim kennzeichnen. Fakten- und kontextbasierte Community Notes kombiniert machen zwar fast die Hälfte aller Post aus (44,4 Prozent). Allerdings steht jenen potenziell hilfreichen Community Notes eine Mehrheit (55,5 Prozent) gegenüber, die nicht nur gegen die eigenen Standards von Community Notes verstoßen; sie haben auch prinzipiell nichts mit der Informationsprüfung zu tun.

Dass Community Notes also systemische Risiken wie Desinformation auf digitalen Plattformen im großen Stil minimieren würden, darf im Falle von X also bezweifelt werden: Kontext zu misleading Posts hinzuzufügen, ist weit weniger präsent als man es vermuten würde. Community Notes – in der Form wie sie auf X umgesetzt werden – tragen daher kaum zur nachhaltigen Minimierung von systemischen Risiken bei, wie sie der DSA beschreibt.

»Dass Community Notes systemische Risiken im großen Stil minimieren würden, darf im Falle von X bezweifelt werden.«

Über das Magazin

Machine Against the Rage (MATR) ist ein wissenschaftliches Online-Magazin für digitale Konfliktforschung. Mit einem Schwerpunkt auf digitalen Hass und Desinformation analysiert es auch, wie politische Auseinandersetzungen im Netz über diese polarisierenden Begriffe geführt werden. Es nutzt dazu Methoden wie Data Science, Netzwerkanalyse und sozialwissenschaftliche Forschung. MATR wird halbjährlich veröffentlicht und ist am IDZ in Jena angesiedelt.

Über das IDZ

Das Institut für Demokratie und Zivilgesellschaft ist eine außeruniversitäre Forschungseinrichtung mit Sitz in Jena, die demokratische Kultur und Zivilgesellschaft in Thüringen, Ostdeutschland und darüber hinaus stärkt. Es untersucht Ursachen, Formen und Folgen von Ideologien der Ungleichwertigkeit sowie Fragen von Diskriminierung und Empowerment. Einen Schwerpunkt legt das IDZ auf die Erforschung digitaler Konflikte und Polarisierung.

Über toneshift

toneshift ist ein bundesweites Netzwerk gegen Hass im Netz und Desinformation. Der Verbund aus sechs etablierten Organisationen hat sich zusammengeschlossen, um Expertise zu Themen wie Hassrede, Gewalt und Desinformation im digitalen Raum zu bündeln und Ansprechpartner für Medien, MultiplikatorInnen und Betroffene zu sein. Teil des Netzwerks ist das IDZ, das in Form von MATR wissenschaftliches Reflexionswissen für die Praxis bereitstellt.

CRI

Gefördert vom Bundesministerium für Bildung, Familie, Senioren, Frauen und Jugend im Rahmen des Bundesprogramms Demokratie leben!

tone shift Netzwerk gegen Hass im Netz und Desinformation

Für inhaltliche Aussagen und Meinungsäußerungen tragen die Publizierenden dieser Veröffentlichung die Verantwortung.

Machine Against the Rage: Magazin für digitale Konfliktforschung
Redaktion: Holger Marcks & Maik Fielitz

Institut für Demokratie und Zivilgesellschaft
Tatzendpromenade 2a
07445 Jena

IDZ INSTITUT FÜR DEMOKRATIE UND ZIVILGESELLSCHAFT

E-Mail: redaktion@machine-vs-rage.net
www.machine-vs-rage.net | www.idz-jena.de

Das am Institut für Demokratie und Zivilgesellschaft (IDZ) angesiedelte Projekt befindet sich in Trägerschaft der Civic Research and Innovation gGmbH (CRI)