Global Media Journal
Journal for International and Transcultural Communication
German Edition

BerlinUP Journals

Special Section: Norms, Power Relations and Injustices in Digitality
Peer-Reviewed Original Article

Content Moderation zwischen Anstandsdame, Schiedsrichter und Zensor? Zu Formen der Kommunikationskontrolle und ihren Folgeproblemen

Lukas Beckmann, Sebastian Suttner & Björn Wiegärtner

Zusammenfassung: Der Beitrag untersucht Content Moderation (CM) aus systemtheoretischer Perspektive als Form der Kommunikationskontrolle. Statt zu normativen oder juristischen Einordnungen zu gelangen, wird CM als praktisches Problem innerhalb kommunikativer Systeme verstanden. Drei historische Sozialfiguren – „Anstandsdame", „Schiedsrichter" und „Zensor" – dienen als heuristische Vergleichspunkte, um unterschiedliche Modi und Probleme der Kommunikationskontrolle zu illustrieren: von interaktiver Überwachung über organisationale Regelanwendung bis hin zur gesellschaftlichen Kontrolle des Sagbaren. Dabei zeigt sich, dass Kommunikationskontrolle stets in Zielkonflikte gerät, die die häufig beobachteten Probleme der Content Moderation zum Ausdruck bringen: zwischen Sichtbarkeit und Unsichtbarkeit, Normsetzung und -durchsetzung sowie Anschluss und Ausschluss. Die Medienlogik digitaler Kommunikation (Asynchronität, Anonymität und Reichweite) verschärft diese Konflikte. Die Untersuchung kommt zu dem Schluss, dass CM als Praxis verstanden werden muss, deren Zielkonflikte sich nicht abschließend lösen lassen, sondern deren Bearbeitung selbst zur dauerhaften gesellschaftlichen Aufgabe wird.

Schlagwörter: Content Moderation, Kommunikationskontrolle, Kommunikationstheorie, Systemtheorie, Luhmann, Hate Speech, Zensur

Abstract: This article examines content moderation (CM) from a systems theoretical perspective as a form of communication control. Rather than pursuing normative or legal classifications, CM is understood as a practical problem within communicative systems. Three historical social figures – "chaperone", "referee", and "censor" – serve as heuristic reference points to illustrate different modes and problems of communication control: from interactive monitoring to organizational rule enforcement to societal regulation of what may be said. It becomes apparent that communication control always encounters conflicts of interest that mirror the frequently observed problems of content moderation: between visibility and invisibility, norm setting and norm enforcement, and inclusion and

Global Media Journal – German Edition
Vol. 15 No. 2 2025
ISSN 2196-4807
DOI: 10.60678/gmj-de.v15i2.331

exclusion. The logic of digital communication (asynchrony, anonymity, and range) further intensifies these conflicts. The study concludes that CM must be understood as a practice whose conflicting goals cannot be conclusively resolved, but whose management itself becomes a persistent societal task.

Keywords: content moderation, communication control, communication theory, systems theory, Luhmann, hate speech, censorship

Angaben zu den Autoren:
Lukas Beckmann ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Allgemeine Soziologie an der Julius-Maximilians-Universität Würzburg. Gegenwärtige Arbeitsschwerpunkte umfassen die soziologische Systemtheorie, Akteur-Netzwerk-Theorie sowie das Verhältnis von Anthropologie zur soziologischen Theorie.
Orcid: https://orcid.org/0009-0005-7922-7754
E-Mail: lukas.beckmann@uni-wuerzburg.de

Sebastian Suttner ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Allgemeine Soziologie an der Julius-Maximilians-Universität Würzburg, wo er an seinem Promotionsvorhaben über die wissenssoziologischen Grundlagen soziologischer Krisenbeschreibungen arbeitet. Zu seinen Hauptinteressen gehören soziologische Theorie (insbesondere Systemtheorie) und Wissenssoziologie, sowie Fragen zu Geschichte der Soziologie und wissenschaftstheoretische Implikationen von soziologischer Forschung und Ideengeschichte.
Orcid: https://orcid.org/0009-0004-3613-717X
E-Mail: sebastian.suttner@uni-wuerzburg.de

Björn Wiegärtner ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Allgemeine Soziologie an der Julius-Maximilians-Universität Würzburg. Seine Forschungsschwerpunkte liegen in der Mediensoziologie, der Gesellschaftstheorie sowie der Qualitativen Sozialforschung.
Orcid: https://orcid.org/0009-0008-3271-0118
E-Mail: bjoern.wiegaertner@uni-wuerzburg.de

Author information:
Lukas Beckmann is a research associate at the Chair of Sociological Theory at Julius-Maximilians-University of Wuerzburg. His current research focuses on sociological systems theory, actor network theory, and the relationship between anthropology and sociological theory.
E-Mail: lukas.beckmann@uni-wuerzburg.de

Sebastian Suttner is a research associate at the Chair of Sociological Theory at Julius-Maximilians-University of Wuerzburg, where he is working on his doctoral thesis about the sociological foundations of knowledge in sociological descriptions of crises. His main research interests include sociological theory (especially sociological systems theory) and the sociology of knowledge, as well as questions concerning the history of sociology and the scientific implications of sociological research and the history of ideas.
E-Mail: sebastian.suttner@uni-wuerzburg.de

Björn Wiegärtner is a research associate at the Chair of Sociological Theory at Julius-Maximilians-University of Wuerzburg. His research focuses on media sociology, theory of society, and qualitative social research.
E-Mail: bjoern.wiegaertner@uni-wuerzburg.de

To cite this article: Beckmann, Lukas; Suttner, Sebastian & Wiegärtner, Björn (2025). Content Moderation zwischen Anstandsdame, Schiedsrichter und Zensor? Zu Formen der Kommunikationskontrolle und ihren Folgeproblemen. *Global Media Journal – German Edition*, *15*(2), DOI: 10.60678/gmj-de.v15i2.331

Content Moderation – vom Phänomen zum Problem

In der noch jungen Tradition der Forschungsarbeiten zum Phänomen des Community Managements und der Content Moderation (CM) werden bereits zahlreiche verschiedene Varianten der CM besprochen und kritisch eingeordnet. Die beschriebenen Reaktionsoptionen der CM reichen von der kritischen Einordnung eines Beitrags über den Verweis auf Community Guidelines, das ‚Flagging' (Melden von Beiträgen) durch andere User:innen oder sogenannte ‚Trusted Flagger', das nachträgliche Löschen eines Beitrags, die Einschränkung der Sichtbarkeit und Reichweite eines Beitrags und automatisierter, teilweise KI-unterstützer Vorab-Löschung (Upload-Filter), bis hin zur strafrechtlichen Anzeige einzelner Beiträge beziehungsweise User:innen (siehe pars pro toto Gillespie et al., 2020). Diese Praktiken lassen sich dabei allesamt unter dem Banner der CM subsummieren, die darauf abzielt, das Verhalten im Netz an kulturelle und moralische Werte zu binden. Einzelne Folgeprobleme dieser Form der Etikettedurchsetzung werden bereits ausführlich besprochen.

So wird forschungsseitig der Blick einerseits auf journalistisch-redaktionelle Formen der CM gerichtet und diskutiert, inwiefern sich das Aufgabenprofil von Journalist:innen erweitert, denn sie sind nicht nur mit der Recherche und Bereitstellung von Informationen, sondern zunehmend auch mit dem Kuratieren und Einordnen von Inhalten (‚Factchecking') sowie der Moderation von Kommentarspalten befasst (Paasch-Colberg & Strippel, 2021). Andererseits sind auch die Plattformbetreiber aufgrund der gesetzlichen Regelungen herausgefordert, spezifische Inhalte zu löschen, zu moderieren, in ihrer Verbreitungsreichweite zu beschränken oder eben aufgrund ihrer Profitmaximierungslogik aktiv sichtbar zu lassen (Gorwa, 2024). Dabei wird vermehrt auf automatisierte und KI-unterstützte Varianten der CM zurückgegriffen, die jeweils ihre eigenen technischen und moralischen Fragen aufwerfen: wegen deren inhärenter Schwierigkeit, problematische Inhalte überhaupt erkennen zu können (Gorwa et al., 2020), wegen der Perpetuierung von Stereotypen und Diskriminierung bei gleichzeitiger Opazität des Entscheidungsprozesses (Peterson-Salahuddin, 2024) sowie wegen der prekären und moralisch verwerflichen Arbeitsbedingungen für diejenigen, die als ‚Clickworker' zum Training automatisierter Moderationstools beitragen (Roberts, 2019). Die Frage, welche Inhalte als *problematisch* und damit moderationswürdig eingestuft werden, erweist sich (trotz der vorhandenen rechtlichen Bestimmungen) als nicht eindeutig zu beantworten (Gillespie, 2018, S. 9).[1] Unterschiedliche Social-Media-Plattformen verwenden deshalb unterschiedliche Moderationsstrategien, die sich in der Regel in Community Guidelines schriftlich manifestieren und die sich freilich im Zeitverlauf merklich ändern können (Dubois & Reepschlager, 2024).

[1] Dieser kritischen Diagnose stimmen Content Moderator:innen selbst zu. Insbesondere bei „Graubereich[en]", „Grenzbereichen" und „Grenzfällen" muss jeweils „von Fall zu Fall", mit „Gefühl", innerhalb eines „Ermessensspielraum[s]" und auf einem „feinen Grat" wandernd abgewogen und entschieden werden (Wagner, 2019, S. 129-140).

Betrachtet man diese aktuelle Forschungslage zu CM allerdings mit etwas Abstand, wird ersichtlich, dass es sich bei den hier versammelten Beiträgen um im Wesentlichen disparate Beobachtungen eines Phänomens handelt, das mit unterschiedlichen Begriffen versehen und insbesondere mit Blick auf normative Fragen und deren Anwendung untersucht wird. Dabei wird entweder die wahrgenommene Differenz zwischen Interaktionspraxis online und offline auf das Medium Internet bzw. die beteiligten Plattformen zugerechnet – und ihnen die Schuld an einer allgemeinen Diskursverrohung angelastet (Habermas, 2022). Oder aber es werden Unterschiede tendenziell verdeckt, wenn allgemein von Phänomenen des ‚Invektiven' (Ellerbrock et al., 2017), also der herabsetzenden Rede oder von Spezifika wie ‚Online Hate' (Tong, 2024) oder ‚Shitstorms' (Stegbauer, 2018) ausgegangen wird, die zu analysieren jedoch ein hohes Maß an Kontextsensibilität voraussetzen und die folgerichtig nur schwer zu generalisierbaren Aussagen führen.[2] Der sozialwissenschaftlichen Beobachtung bliebe dann nur, sich darüber zu wundern, dass „practices of moderation […] often seem to fail" (Nikunen, 2023, S. 184) – ohne konkret angeben zu können, weshalb genau dies der Fall ist.

Indem sich frühe Hoffnungen auf quasi-herrschaftsfreie und deliberative Diskurse im Web 2.0 als Illusion offenbarten, etablieren sich Formen der diskursiven Kontrolle mit dem Zweck der ‚Zivilisierung' dieser Kommunikation. Für CM stellt sich daher die Frage: Wie kann laufende Kommunikation an ethisch-moralische Maßstäbe angepasst werden? Versteht man die unterschiedlichen Formen ‚abzulehnender' Kommunikation nicht als *Phänomen*, welches zur Klassifikation invektiver Formen oder Inhalte führt, sondern als *praktisches Problem* der beteiligten Akteure, mit dem in der Situation umgegangen werden muss, lässt sich breiter ansetzen und nach den Hintergründen der Entstehung und Bearbeitung dieses Problems fragen. Es bietet sich dabei an, nach der Funktionsweise und den Problemen dieser Form der Kommunikationskontrolle zu fragen. Dieser Bezugsrahmen erhebt nicht nur Kommunikation zum Zentralbegriff von Sozialität, sondern ermöglicht es zudem, Medieneffekte sowie weiterführende Implikationen derselben zu beobachten. Statt nach normativ-ethischen Maßstäben, rechtlichen Zulässigkeiten oder unterschiedlichen Varianten des Invektiven zu fragen, lässt sich beobachtbar machen, auf welche *praktischen* Probleme Kommunikationskontrolle stößt.

In unserer Argumentation gehen wir folgendermaßen vor: Ausgehend von einer kommunikationstheoretischen Problematisierung von CM blicken wir auf drei (historische) Sozialfiguren, deren Funktion jeweils als Form der Kommunikationskontrolle beschrieben werden kann: „Anstandsdame", „Schiedsrichter" und „Zensor". An diesen Figuren wollen wir beleuchten, auf welche Probleme die Kontrolle von

[2] In der Forschungsliteratur findet sich hierfür eine Fülle an verschiedenen Begrifflichkeiten: ‚hate speech' (Matamoros-Fernández & Farkas, 2021), ‚incivility' (Su et al., 2018), ‚toxic communication/speech' (Pradel et al., 2024). Trotz der jeweils unterschiedlichen Bezeichnungen ist hier ein einheitlicher Problembezug eines deliberativen Diskurses als Ideal und der Störung dieses Diskurses durch im weitesten Sinne nicht-rationale Diskursbeiträge erkennbar. Daher wird Hate Speech auch essentialisierend als „a waste that should be removed from platforms" (Nikunen, 2023, S. 184) betrachtet. Auf diese Formen invektiver Kommunikation fokussiert der Beitrag im Folgenden.

Kommunikation stößt, wenn man diese über eine konstruktivistische Kommunikationstheorie sichtbar macht. Diese Darstellung ist dabei nicht als lineare Chronologie (von der Anstandsdame hin zur CM) gedacht. Unser Anliegen ist es vielmehr, durch die Betrachtung dieser drei Sozialfiguren eine Heuristik für die Probleme der Kommunikationskontrolle zu gewinnen, die damit das Verständnis von CM hinsichtlich ihrer abzusehenden Folgeprobleme erweitern kann. Die dabei auftretenden Zielkonflikte lassen sich analytisch in eine medien-, kommunikations- und differenzierungstheoretische Problemdimension aufgliedern.

Content Moderation als Problem der Kommunikation

Um zu verstehen, weshalb genau sich die frühe Interneteuphorie invektionsfreier Kommunikation im Netz nicht erfüllen konnte, kann auf kommunikationstheoretische Überlegungen zurückgegriffen werden. Die Kommunikationstheorie nach Luhmann formiert sich als Absetzbewegung gegenüber der klassischen Übertragungsmetapher. Diese ging noch von der Vorstellung aus, im Kommunikationsprozess würde zwischen Ego und Alter eine konstant bleibende Information übertragen. Für Luhmann bleibt hieran ungeklärt, wie sich im Akt der Übertragung dieselbe Information in zwei (oder noch mehr) Bewusstseinen verdoppeln sollte (Luhmann, 2001, S. 46). Kommunikationsprobleme lassen sich dann theorielogisch nur bei Alter sehen, insofern dieser nicht versteht, was Ego sagen will. Für Luhmann stellt sich dieses ‚Missverstehen' nicht länger als zentrales Problem dar. Die Kommunikationstheorie interessiert sich zunächst für die beobachtbaren Mitteilungen und sieht dabei, dass dieselbe Mitteilung schlichtweg „für Absender und Empfänger sehr verschiedenes bedeute[n]" (Luhmann, 2018, S. 194) kann. Dies ist kein *Problem* von Kommunikation, sondern die Bedingung der *Notwendigkeit* von Kommunikation (Luhmann, 2001, S. 47); sie wird nur dort erforderlich, wo nicht bereits Weltdeutungskonsens vorliegt.

Luhmann erarbeitet deshalb einen Kommunikationsbegriff, der vom mathematischen Begriff der Information nach Shannon und Weaver (1964) inspiriert ist und diese als „eine Selektion aus einem (bekannten oder unbekannten) Repertoire von Möglichkeiten" (Luhmann, 2018, S. 195) versteht. Diese Theorievariante blendet Absenderabsichten aus und interessiert sich stattdessen für die Frage nach der Ausbildung *spezifischer Anschlüsse und deren Bedingungen*, insofern ja auch andere Anschlüsse grundsätzlich möglich erscheinen. Dass überhaupt an eine Mitteilung angeschlossen wird, setzt voraus, dass verstanden wird, dass die Information gerade nicht in der Mitteilung aufgeht: Nicht die Mitteilung ist die Information, sondern die Information liegt darin, dass es einen Grund dafür gibt, dass Ego etwas mitteilt (Luhmann, 2001, S. 45). Das Heben der Hand beispielsweise bleibt so lange bloß wahrgenommenes Verhalten, wie der informationelle Begrüßungswert nicht von der Geste selbst unterschieden wird. Mit Kommunikation bekommt man es erst dann zu tun, wenn diese Unterscheidung von (Mitteilungs-)Geste und Begrüßung(-sinformation) praktisch vollzogen wird (Luhmann, 2018, S. 198); das Heben der

Hand wird als Begrüßung (und eben etwa nicht als Sich-Kratzen, oder ‚Stopp'-Symbol) verstanden und hierauf wird seinerseits z. B. mit einer Floskel reagiert. Erst hieran zeigt sich, dass Kommunikation verstanden wurde, weil die Differenz von Mitteilung und Information in die Frage von „Annahme *oder* Ablehnung" transformiert wurde (S. 205), die sich in der Reaktion zeigt: Alter grüßt zurück (Annahme) oder ignoriert das Handheben geflissentlich (Ablehnung).

Damit wird Ego als dem Sender mit seiner Intention die zentrale Stellung im Kommunikationsprozess entzogen. Stattdessen gerät der Emergenzcharakter kommunikativer Akte in den Fokus (Luhmann, 2001, S. 47) und betrachtet auf Kommunikation im Sinne empirischer Anschlüsse, ohne in die Psyche des Individualbewusstseins blicken zu müssen, bzw. zu können. Ego ist nicht länger Träger einer (essentialistischen) Information, die man verstehen oder missverstehen könnte: „Wenn weder eine ursprüngliche Bedeutung noch eine Idee im Bewußtsein als Verankerung des ‚eigentlichen' Sinns dienen, dann kann es auch kein Mißverständnis dieser letztlich nichtexistierenden originalen Bedeutung geben." (Stäheli, 2000, S. 111) ‚Missverständnisse' sind lediglich Themen, an denen Konflikte ausgetragen werden können; das Hand-Heben kann nur ein Kratzen gewesen sein – sobald Alter es jedoch als Anlass zur Begrüßung nimmt, lässt sich dies nur noch kommunikativ zurückweisen: Man kann behaupten, es handele sich um ein Missverständnis, muss dann jedoch darauf vertrauen, dass auch das wiederum ‚richtig' verstanden wird (und nicht etwa als Unwille zu reden aufgrund persönlicher Antipathien).

Diese Kommunikationstheorie wählt deshalb als Bezugsproblem auch nicht missverständnisfreie Kommunikation, sondern die Aufrechterhaltung des Kommunikationsprozesses. Ein Kommunikationssystem, wenn erst einmal etabliert, muss sich über die ephemeren Einzelakte hinweg kontinuieren und hierfür einen Strukturwert bereitstellen (Luhmann, 2001, S. 49); es muss mit anderen Worten Anschluss*erwartungen* ausbilden und Anschluss*möglichkeiten* bereitstellen. Hier setzt dann auch das Problem der CM für die Kommunikation an: Vom Problem des Anschlusses her gedacht gibt es keine Dispräferenzen für invektive Sinngehalte; von der Bereitstellung von Anschlusswahrscheinlichkeit her gedacht, gibt es „keinen zwingenden Grund, die Konsenssuche für rationaler zu halten als die Dissenssuche" (Luhmann, 2001, S. 49). Insofern sind Phänomene invektiver Kommunikation wie Hate Speech zunächst nicht grundsätzlich unwahrscheinlicher oder wahrscheinlicher als andere, konsensorientiertere Anschlussmöglichkeiten. Dies macht die Ausbildung von CM im Grunde genommen zu einem erklärungsbedürftigen Phänomen.

Dies umso mehr, als Kommunikation so besehen als ein unwahrscheinliches Unterfangen gelten muss. Es ist nämlich nicht nur unwahrscheinlich, dass (1) Alter Ego „richtig" versteht, sondern auch dass (2) Kommunikation überhaupt geeignete Adressen findet und (3) am Ende dazu führt, dass ihr Mitteilungsgehalt antizipierbare, praktische Konsequenzen zeitigt (hierzu Luhmann, 1981, S. 26). Vor diesem Hintergrund wäre eher davon auszugehen, dass Kommunikation Ermutigung benötigte und nicht durch Kontrolle potenziell zusätzlich inhibiert wird. Die Entstehung von

Techniken der Kommunikationskontrolle wie der CM muss dann mit Blick auf die durch sie ermöglichte Lösung kommunikationsinhärenter Probleme erklärt werden. Deshalb erscheinen die medialen Settings der Kommunikation entscheidend, denn Medien arbeiten sich an den Problemen dieser Unwahrscheinlichkeiten ab. Verbreitungsmedien wie Schrift oder Telekommunikation etwa reduzieren zwar die *Unwahrscheinlichkeit des Erreichens* von geeigneten Adressen – sie wirken ihrerseits jedoch hemmend darauf, dass „der Empfänger den selektiven Inhalt der Kommunikation (die Information) als Prämisse des eigenen Verhaltens übernimmt" (Luhmann, 1981, S. 26). Im Grunde genommen herrscht eine Art negativer Korrelation zwischen diesen Unwahrscheinlichkeiten – reduziert man eine, steigt eine andere (Luhmann, 2018, S. 219). Das mediale Arrangement zeitigt insofern seine Effekte auf den Anschluss, der kommunikativ ausgebildet *wahrscheinlich* wird.

Die so rekonstruierte Kommunikationstheorie stellt das klassische Selbstverständnis der CM als ‚Moral Gatekeeper' (Boberg et al., 2018) in zweierlei Hinsicht in Frage: Wenn Kommunikationskontrolle als eine quasi-moralische Erziehung zum richtigen Handeln von Akteuren im Netz gesehen wird, erscheint dies unmöglich, da CM nicht nur einhegen müsste, was Ego sagt, sondern auch was Alter versteht. Man kann die vorgenommene Unterscheidung von Mitteilung und Information nicht vorab kontrollieren; welche Informationen aus einer Mitteilung gezogen werden oder wie daran angeschlossen wird, lässt sich nicht im Vorfeld steuern. Kommunikationssteuerung, die sich diese Aufgabe dennoch stellt, kann eigentlich nur als Verhaltenskontrolle (von Alter) imaginiert werden – und erzeugt gerade deshalb auch Widerstand in liberalen Gesellschaften. Zugleich wird kommunikationstheoretisch nun fraglich, wie viel Kontrolle ertragbar ist, bevor das System zum Erliegen kommt, denn: Wenn bereits klar ist, was und wie etwas gesagt werden soll, ist Kommunikation schlicht redundant. Dies kann nicht das Ziel von CM sein – was auch den Protagonist:innen auffällt: „several commented on how odd it is that, on a platform committed to open participation, their job is to kick some people off." (Gillespie, 2018, S. 176) Kommunikationstheoretisch geht es bei CM deshalb um die unmögliche Aufgabe, *bestimmte Formen des Anschlusses* zuzulassen und zugleich *bestimmte andere Formen* zu inhibieren und dies, obwohl man den Anschluss (der ja von Alter entschieden wird) nicht direkt kontrollieren kann. Dass diese Aufgabe als höchst konfliktuös wahrgenommen wird, lässt sich so schon kommunikationstheoretisch erklären: CM operiert in dem Spannungsfeld, möglichst großen Anschlussspielraum zuzulassen und diesen zugleich zu begrenzen.

Das so formulierte, soziologische Bezugsproblem der Kommunikationskontrolle lässt sich nun als Vergleichsmoment für historische Rollen heranziehen, die ebenso an der Funktion einer solchen Kommunikationskontrolle angesiedelt sind. Systemtheoretisch werden hierbei drei Formen der Kontrolle sichtbar: (1) das *Unterdrücken* von Äußerungen, (2) die mitlaufende Kommunikation über richtiges Verhalten (*conduct*) und dessen Sanktion und (3) das Verhindern *der Verbreitung* von Kommunikation. Diese Formen der Kommunikationskontrolle lassen sich an drei rollenförmig ausdifferenzierten Sozialfiguren (Moebius & Schroer, 2018) darstellen: an

der „Anstandsdame“, welche durch interaktionale Teilhabe als taktlos verstandene Äußerungen einschränkt; dem „Schiedsrichter“, der in quasi-organisierten Situationen über regelkonformes Verhalten kommuniziert; und dem „Zensor“, welcher unliebsame Mitteilungen aus dem Sagbarkeitsbereich der Gesellschaft löscht. An diesen historischen Experimentalformen der Kommunikationskontrolle werden dabei nicht nur die Effekte des wechselnden Kommunikationsmediums (von Face-to-Face-Situationen zur Kommunikation im Netz; siehe etwa Barth & Wagner, 2024) sichtbar, sondern auch Probleme, die sich als Folge der gewählten Systemreferenzen wie Interaktion, Organisation und Gesellschaft (dazu Luhmann, 1975) ergeben.

Kommunikationskontrolle in Interaktionen: die Anstandsdame

Eine Sozialfigur, die auf das Problem der Unterdrückung von Äußerungen abzielt, ist die Anstandsdame. Im Übergang von stratifikatorischer zu funktionaler Differenzierung findet sich in der Oberschichtenkommunikation ein Interaktionsideal, welches sich an „Selbstbeherrschung“ in der eigenen Selbstdarstellung orientiert (Luhmann, 1993, S. 91). Es gilt allgemein ein „umsichtige[s] Maßhalten in sozialen Beziehungen“ (S. 103) und insbesondere in der Interaktion mit Damen, „Vernunft, Maß und Passion“ (S. 99) zu verbinden. Im 18. Jahrhundert etabliert sich zur mitlaufenden Kontrolle dieser immer komplexer werdenden oberschichtlichen Etikette die Sonderrolle der Anstandsdame (Wouters, 2004, S. 58).

In Interaktionen, die besonderes Gefährdungspotential hinsichtlich einer möglichen Übertretung der Interaktionsetikette bieten, wird eine dritte Instanz installiert, welche – ohne selbst allzu aktive Interaktionsteilnehmerin zu sein – die Etikette überwacht. Dies gilt insbesondere für junge, unverheiratete Damen sowie verlobte Paare bis zur Hochzeit im Kontext öffentlichen Auftretens wie dem Ausgehen, Reisen oder Situationen des Hof-Machens (Wouters, 2004, S. 58 f.). Das Spezifikum dieser unwahrscheinlichen Rollen-Institutionalisierung zeigt sich dabei insbesondere darin, dass dieses Arrangement auf Face-to-Face-Interaktionen begrenzt ist, die körperlich ko-präsente und sich wechselseitig wahrnehmende Teilnehmende voraussetzen (Goffman, 1982). Eingreifen muss die Anstandsdame dabei lediglich im Extremfall. In der Kommunikation selbst kann sie sonst weitestgehend als ‚abwesend‘ (nicht-)behandelt werden (Kieserling, 1999, S. 64 f.). Die Anstandsdame kann trotz *körperlicher Anwesenheit* aufgrund ihrer *interaktionellen Abwesenheit* Ziel von Spott werden – die noch heute gängige Semantik des ‚Anstandswauwaus‘ drückt dies aus. Erfüllt die Anstandsdame ihre Überwachungsfunktion, bleibt sie paradoxerweise interaktional (als Teilnehmerin wie als Thema) unsichtbar.

Man kann an dieser Technik der Kommunikationskontrolle einige Parallelen zu Formen der CM beobachten, die sich auf interaktive Momente der Kommunikation online beziehen: Vordergründig besteht die Herausforderung der CM ebenfalls darin, Kommunikation zunächst einmal auf im weiteren Sinne moralische Übertretungen *mitlaufend* hin zu beobachten. Die Parallele zeigt sich in ‚Live‘-Settings. Dabei wird

CM zu einer interaktiven Kommunikationsüberwachung, die in Echtzeit kontrollieren und gegebenenfalls eingreifen muss, um Entgleisungen zu verhindern oder zu sanktionieren – was auch simultan sichtbar wird (Thach et al., 2022, S. 4041). Auf vielen Plattformen wurde hierfür die Rolle der ‚Mods' etabliert. An diesen „unpaid volunteers, without relevant purposeful training in their important moderator obligations" (2022, S. 4038) zeigt sich auch, dass die mangelnde Durchsetzungsfähigkeit von Sanktionen das zu Unterbindende geradezu provoziert. Die Anwesenheit einer ansonsten unbeteiligten Kontrollinstanz mit oftmals beschränkten Sanktionsmöglichkeiten lädt insofern zum Testen von Grenzen und Provokationen ein. So ist es – analog zur Anstandsdame – gerade *die in der Interaktion auffallende Anwesenheit der Moderation*, welche dazu tendiert, invektive Anschlusskommunikation heraufzubeschwören. Während die Anstandsdame in der Ständegesellschaft jedoch trotz ihrer interaktiven Prekarität noch immer als legitime (also: an die Ständeordnung gekoppelte) Beobachterrolle institutionalisiert war, gerät dieses Moment bei der CM unter Druck. In der Folge lässt sich die stets präsente Dauerüberwachung durch Dritte interaktiv kaum mehr vor Spott und Hohn bewahren.

Zugleich multipliziert sich für die CM die Komplexität dieses Arrangements in zeitlicher und sozialer Hinsicht, da nun statt den für die Anstandsdame üblichen Charakteristika mündlicher Anwesenheitssettings zusätzlich Schriftmomente die Kommunikation tragen. An- und Abwesenheiten können interaktiv nicht mehr in jedem Fall eindeutig ausgemacht werden und es kommt durch digitale Verbreitungsmedien zu einer „kommunikative[n] Form *anwesender Abwesenheit*" (Barth, 2023, S. 5, Hervorhebung im Original; siehe auch Barth & Wagner, 2024), da Teilnehmendenzahlen permanent schwanken und das Tempo eines Chats sich von den Komplexitätsgrenzen einer körpervermittelten Interaktion unterscheidet. Nikunen beobachtet gerade in der daraus resultierenden kommunikativen Beschleunigung ein Spezifikum von Online-Hate Speech (Nikunen, 2023, S. 179). Nun genügt die bloße Anwesenheit eines Dritten nicht länger, um Verstöße gegen die Etikette prophylaktisch zu verhindern. Zu distanziert, zu vage und reaktionsträge gestaltet sich die personifizierte Technik der Kommunikationskontrolle, als dass Überschreitungen unterblieben. Zu den ohnehin mit der Rolle der Anstandsdame einhergehenden Folgeproblemen der Kommunikationskontrolle gesellen sich im Setting der sozialen Medien damit auch Probleme einer sich einstellenden Asymmetrie zwischen wenigen Kontrollinstanzen gegenüber vielen Kommunikationsofferten. Zudem entfallen die interaktiven Mittel ‚sanfter' Sanktion durch Blicke oder Gesten und können in dieser Hinsicht kaum substituiert werden.

Kontrolle in organisierten Kommunikationssettings: der Schiedsrichter

Stärker organisierte und regelbasierte Interaktionssettings können auf andere Sanktionsmöglichkeiten zurückgreifen und entwickeln dementsprechend andere Rollen, um die Einhaltung einer Kommunikationsetikette zu gewährleisten und Konflikte zu bearbeiten, oder diese – durch die kommunikative Steuerung von

Themen – gar nicht erst aufkommen zu lassen. Als eine solche Rolle erscheint etwa der Schiedsrichter als neuzeitlicher unparteiischer Dritter, der mit zunehmender Professionalisierung von Spielsituationen installiert wird (Huggins, 2023). Zielpunkt dieser Rolle ist weniger die vorgreifende Verhinderung von Zwischenfällen, wie dies noch in der ungezwungenen Interaktion rund um die Anstandsdame der Fall ist; stattdessen geht es hier darum, in abgegrenzten Situationen für kommunikative Moderation in der Form von Regeldurchsetzung und Bearbeitung von Konflikten zu sorgen.

Der Rekurs auf durchzusetzende Regeln erfolgt dabei nicht an einem Kommunikationsideal (etwa: der Oberschicht) orientiert, sondern wird unter den Bedingungen von organisationaler Mitgliedschaft formuliert: Im Fokus stehen durch Mitgliedschaft an der Spielsituation rollenmäßig abgesonderte Spielende, die zeitlich begrenzt auf (regelkonforme) Teilhabe verpflichtet sind und gegebenenfalls exkludiert werden können. Die Entscheidungen des Schiedsrichters haben eine bindende Funktion für Mitglieder, lassen sich jedoch von Nicht-Mitgliedern wie dem Publikum mehr oder minder ignorieren. Die Rolle des Schiedsrichters exponiert sich in hohem Maße, da Entscheidungen in einer mehrdeutigen und komplexen Gesamtlage unter Zeitdruck offen kommuniziert werden müssen und in der Folge ein polarisierendes und konfliktinduzierendes Potential entfalten (Weigelin, 2022). Damit können diese Entscheidungen wiederum selbst zum Bezugspunkt invektiver Kommunikationen werden. Seine Sanktionsfähigkeit beschränkt sich jedoch auf Mitglieder des organisierten Arrangements, das heißt auf Spielende oder gegebenenfalls Trainer:innen, die motivational an die Spielsituation gebunden sind. Gegenüber dem Publikum und seinen möglichen invektiven Initiativen bleibt der Schiedsrichter in seiner Rolle hingegen machtlos.

Überträgt man die Eigenlogik und kommunikativen Folgeprobleme dieser Form der Kommunikationskontrolle auf die CM, wird ersichtlich, dass auch hier oft stark organisierte Interaktionskontexte vorliegen, insofern die Mitglieder der Plattformen durch ihre Teilhabe grundsätzlich auf das Einhalten von Verhaltensmaßgaben (gewissermaßen ‚Spielregeln') in der Form von Community-Richtlinien verpflichtet werden. Diese sind als Entscheidungen über richtiges Verhalten kodifiziert, die wiederum zur Legitimation von Moderationsentscheidungen verwendet werden können (Gillespie, 2018, S. 45, 71). Es bedarf zudem in vielen dieser Settings nicht nur der strengen Beobachtung der Kommunikation, sondern auch der entschiedenen Markierung von Konflikten, die – teilweise erst mit Zeitverzug aufgrund der Schriftlichkeit – kommunikativ als solche behandelt und bearbeitet werden müssen. Diese Aufgabe übernehmen zum Beispiel ‚Faktenchecker', welche konfliktbehaftete Inhalte regulierend einordnen und damit für kommunikative Orientierung sorgen, indem sie vom Standpunkt der kodifizierten Verhaltensregeln besehen, ablehnenswerte Inhalte markieren, einsortieren oder rahmen, ohne diese Inhalte zu entfernen. Damit exponiert sich diese Form der CM auf ähnliche Weise wie der entscheidende Schiedsrichter als unparteiischer Dritter, macht sich selbst angreifbar und wird (zumindest außerhalb seines organisatorischen Kompetenzbereichs) zum

möglichen Referenzpunkt von Hass-Kommunikation (Obermaier, 2023). Schließlich lassen sich die Entscheidungen der CM von außen jederzeit als kontingent, willkürlich, falsch oder parteiisch beobachten und insofern wiederum moralisch diskreditieren (siehe auch Wagner, 2019, S. 139 f.). Häufig entstehen hierzu abseits des ursprünglichen Kontextes der Entscheidung eigene kommunikative Arrangements (als Facebook- oder Telegram-Gruppen, ebenso wie als eigenes Netzwerk wie Truth Social), die sich damit der Mitgliedschafts-bezogenen Sanktionierbarkeit durch die CM entziehen.

Auch hier tritt zu den kommunikationstheoretischen Folgeproblemen der analogen Sozialfigur im Kontext mediatisierter Kommunikation ein weiteres Problem: Die Sanktionierbarkeit durch CM beschränkt sich auf Plattformmitglieder, die über Usernamen oder Accounts sichtbar gemacht werden und meist nur auf dieser Ebene der Sanktion bis hin zur Sperrung des Accounts unterliegen. Während Spielsituationen ebenso wie Organisationen Mitgliedschaft als „Prämisse für Eintritts- und Austrittsentscheidungen“ (Luhmann, 1995, S. 39) definieren und darüber Motivation für den organisierten Handlungszusammenhang voraussetzen können (S. 42), vermag die im Digitalen vollzogene Abstraktion von der eigenen Identität (z. B. durch ein anonymes Social-Media-Profil) die Sanktionsfähigkeit der Organisation zu desavouieren. Im Netz können darüber hinaus selbst noch rollenspezifische Sanktionen umgangen werden, indem neue Accounts bei der gleichen Plattform geschaffen werden. Dies verstärkt sich noch, wenn gruppenintern der Ausschluss von einer Plattform als Auszeichnung verstanden wird. Damit ist der CM als Schiedsrichter nicht nur – wie sein analoges Pendant – machtlos gegenüber Nicht-Mitgliedern, sondern seine Sanktionsfähigkeit gegenüber Mitgliedern ist aufgrund des medialen Settings auch erheblich eingeschränkt.

Die Entscheidung über kommunikative Existenz: der Zensor

Zuletzt etabliert sich mit zunehmender Verbreitung gedruckter Medien eine weitere, eigens hierauf ausgerichtete Rolle der kommunikativen Kontrolle. Zensur wird institutionalisiert, um die Kontrolle von Kommunikation als Kontrolle des Sagbaren zu ermöglichen (Roßbach, 2024, S. 15). Wo zensiert wird, ist am Material entweder gar kein Eingreifen sichtbar, womit der Zensor hinter der Zensur verschwindet, da kein Senden mehr stattfindet; oder es bleibt die Markierung einer Leerstelle in Form geschwärzter Textabschnitte als Hinweis auf die Entscheidung zur Zensur und die Existenz des Zensors, aber nicht mehr auf den Inhalt. Anders als Schiedsrichter und Anstandsdame kann der Zensor selbst anonym bleiben, da die Entscheidung nicht auf ihn als Person verweist. Diese meist stark rechtlich-organisatorisch und politisch institutionalisierte Variante der Kommunikationskontrolle lässt sich nur in dafür geeigneten Medienkontexten ausüben und bedarf eines hohen Maßes an technischer Eingebundenheit in Verbreitungs- und Entscheidungsprozesse. Echtzeitkommunikation hingegen lässt sich nicht zensieren. Zugleich besetzt der Zensor eine Grenzstelle zwischen politisch-rechtlicher Reglementierung und

organisationsförmig-massenmedialer Verbreitung. Er lässt sich klassischerweise in eigens eingerichteten Zensurämtern autokratischer Staaten beobachten und ist mit der Prüfung literarischer oder massenmedialer Schriften befasst. Für die Frage nach den Folgeproblemen von Kommunikationskontrolle interessiert dabei weniger der zensierte – also empirisch unauffindbare Inhalt, sondern die Rolle des Zensors selbst. Schließlich ist es soziologisch betrachtet nur das Wissen um die Existenz einer Zensurinstanz, welche den Charakter von (veröffentlichter) Kommunikation verändert und damit auf die Systemreferenz Gesellschaft verweist – wenn man Gesellschaft als den Horizont aller möglicher Kommunikation begreift (Luhmann, 1975).[3]

Das zentrale Problem der Kommunikationskontrolle zeigt sich hier entsprechend in der Relation von Kommunikation zum Zensor: Eine Mitteilung, die vernehmbar ist, hat es durch die Zensur geschafft und kann damit als sagbar verstanden werden. Sie ist damit *gesellschaftliche* Kommunikation. Die Mitteilung kann vor dem Hintergrund ihrer möglichen Zensur nun jedoch bedeutungsmäßig verdoppelt werden: Weiß man um den Zensor, wird schon die Beobachtbarkeit von Kommunikation selbst zur Information: Das Mitgeteilte ist auch *offiziell* sagbar. Dies legt andererseits die Suche nach verdeckten Bedeutungen im Text nahe: Man kann Inhalte codieren oder nach versteckten Codes suchen, die von der Zensur nicht erfasst werden, weil sie nur idiosynkratisch oder subtextuell verstehbar sind. Alternativ kann der Zensor dann durch Wechsel des Kommunikationskontextes, wie durch den Rückzug ins Private, umgangen werden (Feuchert & Ehrhardt, 2024, S. 336).

Insofern Kommunikationskontrolle durch Zensur auf die Grenze zwischen Kommunikation und Nicht-Kommunikation verweist, lassen sich diese Erkenntnisse auf die Situation der CM übertragen. Auch bei CM besteht die Möglichkeit, auf Codierungen auszuweichen.[4] Das online Kommunizierte verändert seinen Charakter bereits durch das Wissen um vorgängige Filter oder nachträgliche Löschungen: Indem jede Kommunikation nun vermittels der Differenz von Löschung/Akzeptanz betrachtet werden kann, rückt die Frage nach den Grenzen des Sagbaren in den Vordergrund. Dabei lässt sich dann die gewählte Form zensierender CM selbst noch moralisieren, sobald Sagbarkeit zum Bezugsproblem moralischer Kommunikation erhoben wird. Erst hier, und nicht bei ‚Mods' oder ‚Flagging', bekommt man es deshalb mit politisch-rechtlichen Problemen zu tun (wo sich auch historisch die Grenzrolle des Zensors ausdifferenziert).[5]

---

[3] An diesem Punkt entzünden sich juristische und demokratietheoretische Debatten, zum Beispiel um den sogenannten ‚Chilling-Effekt' (Reinbacher, 2022, S. 807 ff.).

[4] Dieses Phänomen der bewusst eingesetzten kreativen Sprachvariation wird unter dem Stichwort „Algospeak" besprochen (Steen et al., 2023). In der Forschungsliteratur zu Hate Speech wird zudem darauf hingewiesen, dass Hass-Botschaften gezielt in Form von Bildern und Memes kommuniziert werden – wiederum auch um CM zu umgehen (Oehmer-Pedrazzi & Pedrazzi, 2024). Munn (2020) kommt daher zu folgendem Schluss: „the inventiveness of users and the ambiguity of language mean that toxic communication remains complex and difficult to address" (S. 2).

[5] Differenzierungstheoretisch betrachtet, hat man es unter den spezifischen Bedingungen der Moderne in diesen Fällen mit Fragen der Eigenlogik und dem dazugehörigen Kontext von Kommunikation zu tun. So kann Hasskommunikation sich als politisch erfolgreich erweisen, jedoch rechtlich problematisch sein; sie kann rechtlich unproblematisch sein und doch finanzielle Probleme erzeugen

Durch die Mediatisierung der Kommunikation lässt sich nunmehr die unsichtbar gewordene Zensurinstanz algorithmisch automatisieren: als Upload-Filter (vorher) oder KI-gestützte Bewertung und Löschung von Inhalten (nachher). Das (nachträgliche) Löschen von Beiträgen wird durch die verantwortlichen CMs gemeinhin als Ultima Ratio betrachtet – nicht nur, weil die Grenze zwischen dem Gerade-noch-Sagbaren und dem Gerade-nicht-mehr-Sagbaren nicht eindeutig zu bestimmen ist, sondern insbesondere auch weil die Löschpraxis als solche sich anschließend skandalisieren lässt und so zu einem als erhöht empfundenen Legitimationsdruck führt. Stattdessen scheinen Plattformen mit verschiedenen Formen der Zensur zu experimentieren – die kommunikativen Folgeprobleme hiervon sind jedoch noch weitestgehend unabschätzbar: Eine Alternative zur vollständigen Entfernung im Sinne einer Zensur, die wiederum nur unter den medialen Bedingungen von Online-Kommentarspalten funktioniert, ist etwa die Einschränkung der Reichweite und Sichtbarkeit eines Postings, was seinerseits moralische Fragwürdigkeiten aufweist (Wagner, 2019, S. 139 f.). Anstelle des vollständigen Löschens eines Beitrags wird bei dieser Form der CM lediglich moduliert, welchen Userprofilen dieser angezeigt wird (sogenanntes ‚Shadowbanning'). Auch auf Social Media-Plattformen kann eine Form der „visibility moderation" (Zeng & Kaye, 2022) zum Einsatz kommen; hier werden in der Regel bestimmte Postings in ihrer durch Ranking-Algorithmen gesteuerten Zirkulation gedrosselt – ein invektives Posting schlägt folglich keine hohen (Erregungs-)Wellen mehr (siehe auch Döveling & Seyfert, 2023, S. 32).[6]

Fazit: Das dreifache Problem der Content Moderation

Ausgangspunkt unserer Argumentation war eine *kommunikationstheoretische Problematisierung* des Selbstverständnisses von Kommunikationskontrolle der CM als moralische Instanz, welche im Netz richtiges Verhalten bedingt. Die Perspektive der systemtheoretischen Kommunikationstheorie ermöglichte es dabei, auf die Probleme, die mit Techniken der Kommunikationskontrolle einhergehen, soziologisch scharf zu stellen und zu problematisieren, inwiefern bestimmte Anschlüsse wahrscheinlicher gemacht werden können als andere. Hieraus ergab sich die Frage nach dem Vergleich zu anderen historischen Figuren der Kommunikationskontrolle, die sich – ebenso wie CM – mit den Folgeprobleme ihrer eigenen Kontrollversuche konfrontiert sehen. Diese kommunikationstheoretische Problematisierung wird dabei einerseits durch eine differenzierungstheoretische Perspektive

(Barth et al., 2023). Daran interessiert für unseren Fall wiederum, dass die Frage der ‚Sagbarkeit' ihre eigenen Anschlussprobleme in den jeweiligen Funktionssystemen der Gesellschaft erzeugt. Zensur kann also, sobald sie erkannt ist, immer unter Bedingungen politischer Klugheit, rechtlicher Legitimität, massenmedialen Informationswerts oder wirtschaftlicher Rentabilität und so weiter beobachtet und kritisiert oder eingefordert werden.

[6] Überlässt man die Entscheidung der Reichweiteneinschränkung eines Postings allerdings einer automatisierten/KI-gestützten CM, so kann es auf Social-Media-Plattformen, für die die ungehinderte Zirkulation (gerade) von diskussionswürdigen Beiträgen zum Geschäftsmodell gehört, freilich auch „zu Situationen kommen, in denen problematische Beiträge sowohl belohnt als auch bestraft werden" (Döveling & Seyfert, 2023, S. 32).

ergänzt. Andererseits zeigt das medientheoretische Argument auf, dass sich die Unabschätzbarkeit möglicher Folgeprobleme drastisch potenziert.

Das Spannungsfeld, in welchem CM sich bewegt, ist dabei wesentlich durch das *Medium der Kommunikation* selbst bedingt. Sie reagiert nicht nur auf das ständig mitlaufende Potential von Kommunikation, als Abweichung markiert und als Konflikt ausgetragen werden zu können, sondern auch auf die Erfordernisse der Medienumgebung sozialer Medien selbst. CM kann das Medium schließlich nicht umgehen, in dem die zu bearbeitende Kommunikation abläuft. Vielmehr zeigt sich am Medium, welche Grenzen und Schwierigkeiten diese Form der Kommunikationskontrolle zu überwinden und beständig zu lösen hat. Zu einem guten Teil besteht die in der Einleitung angesprochene schwierige Lage der CM daher in einem *Medienproblem*. Das Medium der Online-Kommunikation beinhaltet Momente mündlicher (synchroner) sowie schriftlicher (asynchroner) Settings *zur selben Zeit* (Barth, 2023, S. 6). Unterschiedliche Strategien der Kommunikationskontrolle werden daher erforderlich und beobachtbar, deren Funktionieren alles andere als wahrscheinlich ist.

Andererseits ergeben sich jenseits der Medieneffekte Zielkonflikte, wie sie den drei vorgestellten Sozialfiguren ebenfalls inhärent sind und die sich auf Anforderungen der jeweiligen *Systemreferenz* beziehen lassen. Unser Ziel war es dabei nicht, den Sozialfiguren Systemkompetenzen zu attestieren, sondern vielmehr anhand jener Typik spezifische Folgeprobleme von Kommunikationskontrolle sichtbar zu machen: Die Anstandsdame unterliegt den Dynamiken der Interaktion als abwesende Anwesende und kann genau in dieser Rolle verächtlich gemacht werden. Als 'Anstandswauwau' wird ihrer widersprüchlichen Rolle ironisierend Rechnung getragen. An ihr zeigt sich, dass die Kontrolle von *Kommunikation durch Anwesenheit* dort scheitert, wo die Kommunikation Anwesende in den Indifferenzbereich der Abwesenheit abschiebt. Dieses Problem hatten wir auch im Fall der Echtzeit-Moderation von Chats ausmachen können. Der Versuch, bestimmte Äußerungen durch die Darstellung von interaktiver Autorität zu unterdrücken, läuft beständig Gefahr, als solche beobachtet und verächtlich gemacht zu werden und Abweichung erst heraufzubeschwören.

Der Zielkonflikt des Schiedsrichters ergibt sich dagegen aus der Notwendigkeit, zwar sanktionsfähige, aber damit umso exponiertere Entscheidungen zu treffen, um Konflikte in organisierten Kontexten zu bearbeiten. Während seine Entscheidungen für Mitglieder der Spielsituation bindend sind, bleiben Konflikte gegenüber dem Publikum ungelöst – der Schiedsrichter kann Publikumskommunikation nicht unterbinden. Eben jene Exponiertheit bei der Moderation von Beiträgen lässt sich auch online beobachten und macht den entscheidenden Moderator zur möglichen Zielscheibe invektiver Kommunikationen. Der Versuch, *Kommunikation durch Entscheidung zu kontrollieren*, läuft also darauf hinaus, unter den Bedingungen von Kommunikation agieren zu müssen und unterliegt damit denselben Dynamiken, die eine Kontrolle erst nötig gemacht hatten. Im Falle der CM vermag Mitgliedschaft

nicht länger zu disziplinieren, sondern muss darauf setzen, dass Mitglieder mit einem Eigeninteresse an kommunikativer Disziplin beitreten.

Zuletzt lässt sich im Fall der Zensur beobachten, dass diese Form der Kommunikationskontrolle zu einer Codierung der Kommunikation führen kann oder auf andere Formate ausweicht. Auch in diesem Fall offenbart sich daher ein Zielkonflikt bei der Kontrolle von Kommunikation durch *Unterbinden von Kommunikation*, der im Fall der CM analog beobachtet werden kann: Hier bilden sich eigene Codes aus, welche CM – insbesondere in ihrer automatisierten Form – bis in die Unmöglichkeit erschweren. Diese dritte Form der Kommunikationskontrolle führt damit letztlich wieder zurück zum kommunikationstheoretischen Ausgangsargument: Entsteht die Information nicht beim Sender, sondern wird sie vom Empfänger generiert, kann Kommunikation codiert werden und dadurch Zensur umgehen. Zudem können Kommunikationskanäle nicht ubiquitär kontrolliert werden, was gerade unter den technischen Voraussetzungen des Mediums ‚Internet' zensorischen Versuchen zuwiderläuft.

Unter den Bedingungen von digitaler Medialität (asynchron, anonym, reichweitenstark), invektivem Hintergrundpotential und Steuerungsaversion jeder Kommunikation sowie dem Verschmelzen unterschiedlicher Systemreferenzen braut sich für CM ein ‚perfect storm' zusammen. Bei diesen vielschichtigen Anforderungen an die CM erscheint es unmöglich, eine Strategie zu wählen, welche ihre Folgen (‚shitstorm' versus ‚candystorm') abschätzbar werden lässt.

Im eingangs angesprochenen Dilemma, das Hate Speech entweder als Medieneffekt oder aber als kontextunspezifisches Phänomen einer (zunehmenden) ‚Inzivilität' betrachtet, lässt sich nun beobachten, dass beide Perspektiven letztlich auf eine Verkürzung der Situation der CM als Form der Kommunikationskontrolle hinauslaufen. Keineswegs erzeugen die Plattformen einfach Invektivität, die durch die Einführung von Moderationsformen dauerhaft oder gänzlich verhindert werden könnte, um so zu einer Art Paradies konsensualer oder nicht-invektiver Kommunikation zurückzugelangen. Zugleich kann das Medium der Kommunikation nicht ausgeblendet werden, indem man stattdessen auf konstante Formen des Invektiven fokussiert, die in ihrer phänomenalen Qualität gewissermaßen von außen durch Individuen in die Kommunikation getragen würden.

Als kommunikationsinhärentes Problem, das folglich auch nur *in der Kommunikation* adressiert werden kann, scheint Invektivität durch drei wesentliche Strategien behandelbar: (1) das Unterdrücken von Äußerungen, (2) die Kontrolle von Kommunikation durch Kommunikation und (3) das Verhindern der Verbreitung von Kommunikation. Wir konnten zeigen, dass alle drei Zugänge die oben thematisierten Folgeprobleme mit sich bringen. Für Zivilisiertheit der Kommunikation bedeutet das, sie sollte weder als Naturzustand ex ante noch als prinzipiell unmögliches und daher irrelevantes Phänomen verstanden werden. Wo die Frage nach der Zivilisiertheit von Kommunikation als praktisches Problem erkannt und institutionalisiert

wurde, entwickelten sich entsprechende Mechanismen wie Etikette, Takt oder Erziehung. Auch diese können die Eigendynamik von Kommunikation nicht hintergehen. Sie generieren stattdessen Folgeprobleme, die auch für jegliche Formen der CM von entscheidender Bedeutung sind, und beispielsweise im Rahmen Digitaler Ethik verhandelt werden (siehe etwa Howard, 2024). Für die soziologische Forschung eröffnet sich in der Beobachtung der jeweiligen Folgeprobleme solcher Zivilisierungsversuche ein sowohl empirisches als auch theoretisches Forschungspotential.

Bibliografie

Barth, N. (2023, 31. Mai). *Mehr Vulgarität? Gesetz der steigenden Negationsrate*. KWI-Blog. https://doi.org/10.37189/kwi-blog/20230531-0830

Barth, N. & Wagner, E. (2024). Indiskrete Indiskretionen: Klatschkommunikation über anwesend Abwesende. In R. Gaderer & V. Grömmke (Hrsg.), *Hass teilen: Tribunale und Affekte virtueller Streitwelten* (S. 225–248). Transcript.

Barth, N., Wagner, E., Raab, P., & Wiegärtner, B. (2023). Contextures of hate: Towards a systems theory of hate communication on social media platforms. *The Communication Review*, *26*(3), 209–252. https://doi.org/10.1080/10714421.2023.2208513

Boberg, S., Schatto-Eckrodt, T., Frischlich, L., & Quandt, T. (2018). The Moral Gatekeeper? Moderation and Deletion of User-Generated Content in a Leading News Forum. *Media and Communication*, *6*(4), 58–69. https://doi.org/10.17645/mac.v6i4.1493

Döveling, K. & Seyfert, R. (2023). Digitale Affektkulturen. Soziale Medien als affektive Intensitätsmedien. In G. L. Schiewer, J. Szczepaniak & J. Pociask (Hrsg.), *Emotionen – Medien – Diskurse. Interdisziplinäre Zugänge zur Emotionsforschung* (S. 23–36). Harrassowitz.

Dubois, E., & Reepschlager, A. (2024). How harassment and hate speech policies have changed over time: Comparing Facebook, Twitter and Reddit (2005–2020). *Policy & Internet*, *16*, 523–542. https://doi.org/10.1002/poi3.387

Ellerbrock, D., Koch, L., Müller-Mall, S., Münkler, M., Scharloth, J., Schrage, D. & Schwerhoff, G. (2017). Invektivität – Perspektiven eines neuen Forschungsprogramms in den Kultur- und Sozialwissenschaften. *Kulturwissenschaftliche Zeitschrift*, *2*(1), 2–24. https://doi.org/10.25969/mediarep/3593

Feuchert, S. & Ehrhardt, J. (2024). Das 20. Jahrhundert und die (fast) totale Zensur in der Moderne. In N. Roßbach (Hrsg.), *Zensur: Handbuch für Wissenschaft und Studium* (S. 313–350). Nomos.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., Roberts, S. T., Sinnreich, A., & Myers West, S. (2020). Expanding the debate about content moderation: scholarly research agendas for the coming policy debates. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1512

Goffman, E. (1982). Die Interaktionsordnung. In H. Knoblauch (Hrsg.), *Interaktion und Geschlecht* (S. 50–104). Campus.

Gorwa, R. (2024). *The Politics of Platform Regulation: How Governments Shape Online Content Moderation*. Oxford University Press.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Habermas, J. (2022). *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik*. Suhrkamp.

Howard, J. (2024). The Ethics of Social Media: Why Content Moderation is a Moral Duty. *Journal of Practical Ethics*, *11*(2), 33–52. https://doi.org/10.3998/jpe.6195

Huggins, M. (2023). Umpires, Referees, Judges and Stewards: Changing Modes of Judgment and Arbitration in English Sport c.1600–c.1900. *The International Journal of the History of Sport*, *40*(8), 661–679. https://doi.org/10.1080/09523367.2023.2242282

Kieserling, A. (1999). *Kommunikation unter Anwesenden. Studien über Interaktionssysteme*. Suhrkamp.

Luhmann, N. (1975). Interaktion, Organisation, Gesellschaft. Anwendungen der Systemtheorie. In N. Luhmann, *Soziologische Aufklärung 2. Aufsätze zur Theorie der Gesellschaft*. (S. 9–21). Westdeutscher Verlag.

Luhmann, N. (1981). Die Unwahrscheinlichkeit der Kommunikation. In N. Luhmann, *Soziologische Aufklärung 3. Soziales System, Gesellschaft, Organisation* (3. Aufl., S. 25–34). Westdeutscher Verlag.

Luhmann, N. (1993). *Gesellschaftsstruktur und Semantik. Studien zur Wissenssoziologie der modernen Gesellschaft*. (Bd. 1). Suhrkamp.

Luhmann, N. (1995). *Funktionen und Folgen formaler Organisation: Mit einem Epilog 1994*. Duncker & Humblot.

Luhmann, N. (2001). Was ist Kommunikation? In N. Luhmann, *Short Cuts* (2. Aufl., S. 41–63). Zweitausendeins.

Luhmann, N. (2018). *Soziale Systeme: Grundriß einer allgemeinen Theorie* (17. Aufl.). Suhrkamp.

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, Hate Speech, and Social Media: A Systematic Review and Critique. *Television & New Media*, *22*(2), 205–224. https://doi.org/10.1177/1527476420982230

Moebius, S. & Schroer, S. (2018). *Diven, Hacker, Spekulanten. Sozialfiguren der Gegenwart*. Suhrkamp.

Munn, L. (2020). Angry by design: toxic communication and technical architectures. *Humanities and Social Sciences Communication*, *7*(53). https://doi.org/10.1057/s41599-020-00550-7

Nikunen, K. (2023). Affective Temporalities of Digital Hate Cultures. In M. Lünenborg & B. Röttger-Rössler (Hrsg.), *Affective Formations of Publics. Places, Networks, and Media* (S. 173–190). Routledge.

Obermaier, M. (2023). Occupational Hazards: Individual and Professional Factors of Why Journalists Become Victims of Online Hate Speech. *Journalism Studies*, *24*(7), 838–856. https://doi.org/10.1080/1461670X.2023.2173955

Oehmer-Pedrazzi, F. & Pedrazzi, S. (2024). "An image hurts more than 1000 words?": Sources, channels, and characteristics of digital hate images. *Communications*, *49*(3), 421–443. https://doi.org/10.1515/commun-2023-0117

Paasch-Colberg, S., & Strippel, C. (2021). "The Boundaries are Blurry…": How Comment Moderators in Germany See and Respond to Hate Comments. *Journalism Studies*, *23*(2), 224–244. https://dx.doi.org/10.1080/1461670X.2021.2017793

Peterson-Salahuddin, C. (2024). Repairing the harm: Toward an algorithmic reparations approach to hate speech content moderation. *Big Data & Society*, *11*(2). https://doi.org/10.1177/20539517241245333

Pradel, F., Zilinsky, J., Kosmidis, S., & Theocharis, Y. (2024). Toxic Speech and Limited Demand for Content Moderation on Social Media. *American Political Science Review*, *118*(4), 1895–1912. https://doi.org/10.1017/S000305542300134X

Reinbacher, T. (2022). „Das wird man doch wohl noch sagen dürfen!" Politische Meinungsäußerungen im Internet als strafbare Beleidigung. *Zeitschrift für das Juristische Studium*, *15*(6), 802–810. https://www.zjs-online.com/dat/artikel/2022_6_1686.pdf

Roberts, S. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Roßbach, N. (Hrsg.). (2024). *Zensur: Handbuch für Wissenschaft und Studium*. Nomos. https://doi.org/10.5771/9783748930037

Shannon, C. E., & Weaver, W. (1964). *The Mathematical Theory of Communication*. University of Illinois Press.

Stäheli, U. (2000). *Sinnzusammenbrüche: Eine dekonstruktive Lektüre von Niklas Luhmanns Systemtheorie*. Velbrück Wissenschaft.

Steen, E., Yurechko, K., & Klug, D. (2023). You Can (Not) Say What You Want: Using Algospeak to Contest and Evade Algorithmic Content Moderation on TikTok. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231194586

Stegbauer, C. (2018). *Shitstorms: Der Zusammenprall digitaler Kulturen*. Springer VS.

Su, L. Y.-F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., & Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. *New Media & Society*, *20*(10), 3678–3699. https://doi.org/10.1177/1461444818757205

Thach, H., Mayworm, S., Delmonaco, D., & Haimson, O. (2022). (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society*, *26*(7), 4034–4055. https://doi.org/10.1177/14614448221109804

Tong, S. T. (2024). Foundations, Definitions, and Directions in Online Hate Research. In J. B. Walther & R. E. Rice (Hrsg.), *Social Processes of Online Hate* (S. 37–72). Routledge.

Wagner, E. (2019). *Intimisierte Öffentlichkeiten: Pöbeleien, Shitstorms und Emotionen auf Facebook*. Transcript.

Weigelin, M. (2022). Entscheidungen und ihre Bewertungen – Zur Mikrosoziologie des Schiedsrichter-Pfiffs. *Österreichische Zeitschrift für Soziologie*, *47*(3), 225–246. https://doi.org/10.1007/s11614-022-00500-4

Wouters, C. (Hrsg.). (2004). *Sex and manners: Female emancipation in the West, 1890-2000*. SAGE.

Zeng, J., & Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, *14*(1), 79–95. https://doi.org/10.1002/poi3.287