Deutsche Stiftung für Recht und Informatik

# Was muss raus, was darf bleiben? –

## Einsatz künstlicher Intelligenz bei der Löschung und Sperrung rechtswidriger Äußerungen in sozialen Netzwerken

**Nils Eckert**
PLANIT // LEGAL Rechtsanwaltsgesellschaft mbH

Herbstakademie 2023


Weltweit ca. 5 Mrd. Nutzer sozialer Netzwerke,
in Deutschland ca. 72 Millionen
Äußerungen in sozialen Netzwerken
Einsatz künstlicher Intelligenz?
Weltweit ca. 100.000 Content Moderatoren,
in Deutschland ca. 5.000

# Agenda

**Teil 1:**

Welche Äußerungen

I. müssen soziale Netzwerke löschen?

II. dürfen soziale Netzwerke löschen bzw. müssen in sozialen Netzwerken einsehbar bleiben?

**Teil 2:**

Unter welchen rechtlichen Rahmenbedingungen dürfen soziale Netzwerke künstliche Intelligenz bei Löschungen von Äußerungen einsetzen?

# Teil 1

# I. Pflicht zur Löschung „rechtswidriger Inhalte“

## 1. Gesetzliche Grundlagen

- **NetzDG:** Inkrafttreten 1. Oktober 2017, novelliert im Juni 2021 und Februar 2022
- **Digital Services Act (DSA):** Ablauf Umsetzungsfrist: 17. Februar 2024, bereits ab 25. August 2023 für „sehr große Online-Plattformen“
- **Verhältnis der Gesetze:** Anwendungsvorrang der Union – ab Februar 2024 allein am DSA zu messen, ob Inhalte gelöscht werden müssen oder nicht (siehe ErwG 9 DSA)

- **Deutsche Rechtsfigur:** Störerhaftung

# I. Pflicht zur Löschung „rechtswidriger Inhalte“

## 2. Begriff der „sozialen Netzwerke“

- Legaldefinition in § 1 Abs. 1 NetzDG:

> *„Telemediendiensteanbieter, die mit Gewinnerzielungsabsicht Plattformen im Internet betreiben, die dazu bestimmt sind, dass Nutzer beliebige Inhalte mit anderen Nutzern teilen oder der Öffentlichkeit zugänglich machen.“* Journalistisch-redaktionelle Angebote und Angebote zur Individualkommunikation sind ausgeschlossen.

- DSA: keine Legaldefinition → abgestuftes Regulierungskonzept:
  - Stufe 1: Vermittlungsdienste in der EU (+)
  - Stufe 2: Hostingdienste (+)
  - Stufe 3: Online-Plattformen (+)
  - Stufe 4: Sehr große Online-Plattformen (+/-) - mehr als 45 Mio. Nutzer in der EU

# I. Pflicht zur Löschung „rechtswidriger Inhalte“

## 3. Begriff der „Löschung“ und „Sperrung“

- **Löschung:** Inhalte sind weltweit nicht mehr verfügbar

- **Sperrung:** Inhalte sind in einem bestimmten Gebiet (bestimmter Staat oder EU) nicht mehr verfügbar.

# I. Pflicht zur Löschung „rechtswidriger Inhalte“

## 4. „rechtswidrige Äußerungen“ müssen gelöscht werden

| NetzDG | DSA |
|---|---|
| „Rechtswidrige Inhalte“:<br><br>Ausschließlich auf Normen des deutschen StGB beschränkt, die in § 1 Abs. 3 NetzDG abschließend aufgezählt sind, insbesondere<br>- Beleidigung (§ 185)<br>- Üble Nachrede (§ 186)<br>- Verleumdung (§ 187)<br>- Volksverhetzende Inhalte (§ 130)<br>- Öffentliche Aufrufe von Straftaten (§ 111) | „Rechtswidrige Inhalte“:<br><br>„Sämtliche Informationen, die nicht im Einklang mit dem Unionsrecht oder dem Recht der Mitgliedsstaaten stehen.“<br><br>→ weite Definition: Alle Normen des StGB und alle sonstigen Verstöße gegen das Recht |

# I. Pflicht zur Löschung „rechtswidriger Inhalte“

| NetzDG | DSA |
|---|---|
| Keine proaktiven Überwachungs- und Löschpflichten (Art. 8 DSA)<br>→ Löschung erst nach Meldung („notice-and-action“) | |
| Keine Durchsetzung von Persönlichkeitsrechtsverletzungen (quasinegatorischer Unterlassungsansprüche §§ 1004 Abs. 1, 823 BGB analog i.V.m. Art. 2 Abs. 1 iV.m Art. 1 Abs. 1 GG)<br><br>→ nur Störerhaftung möglich (siehe Exkurs) | Grundsätzlich Anwendung auf Persönlichkeitsrechtsverletzungen (quasinegatorische Unterlassungsanspruch)<br><br>→ aber: Haftungsprivilegierung, Art. 16 Abs. 3 DSA:<br><br>„Rechtswidrigkeit muss ohne eingehende rechtliche Prüfung feststellbar sein“ |

# I. Pflicht zur Löschung „rechtswidriger Inhalte“

| NetzDG | DSA |
|---|---|
| Pflicht zur Einrichtung eines Meldesystems (§ 3a NetzDG) | Pflicht zur Einrichtung eines Beschwerdemanagementsystems (Art. 16 DSA) |
| Pflicht zur Einrichtung eines Gegenvorstellungsverfahrens zur Überprüfung getroffener Moderationsentscheidungen (§ 3b NetzDG) | Pflicht zur Einrichtung eines internen Beschwerdemanagementsystems (Art. 20 DSA) |
| Sanktionen: Bußgeld von bis zu 5 Millionen Euro (§ 4 Abs. 2 NetzDG)<br><br>Haftung nach den Grundsätzen der Störerhaftung (siehe Exkurs) | Sanktionen: Bußgeld von bis zu 6 % des weltweiten Jahresumsatzes (Art. 52 DSA), Zwangsgelder bis 5 % Tagesumsatz, Schadensersatz (Art. 54 DSA) |

# Exkurs: Störerhaftung bei Persönlichkeitsrechtsverletzungen

- In DE: Institut der Störerhaftung analog § § 1004, 823 BGB
- Soziale Netzwerke können als Störer auf Unterlassung/Schadensersatz haften.
- Auch bei Verletzungen von Persönlichkeitsrechten (z.B. unwahren Tatsachenbehauptungen), die keine Straftatbestände erfüllen.
- Störerhaftung ist künftig an den prozeduralen Vorgaben des DSA zu messen

  → Art. 16 Abs. 3 (früher § 10 TMG): Rechtswidrigkeit muss „ohne eingehende rechtliche Prüfung erkennbar sein“.

  → bei Persönlichkeitsrechtverletzungen häufig problematisch

(P) Kann prozedurales Haftungssystem der deutschen Rspr. unter DSA aufrecht erhalten werden?

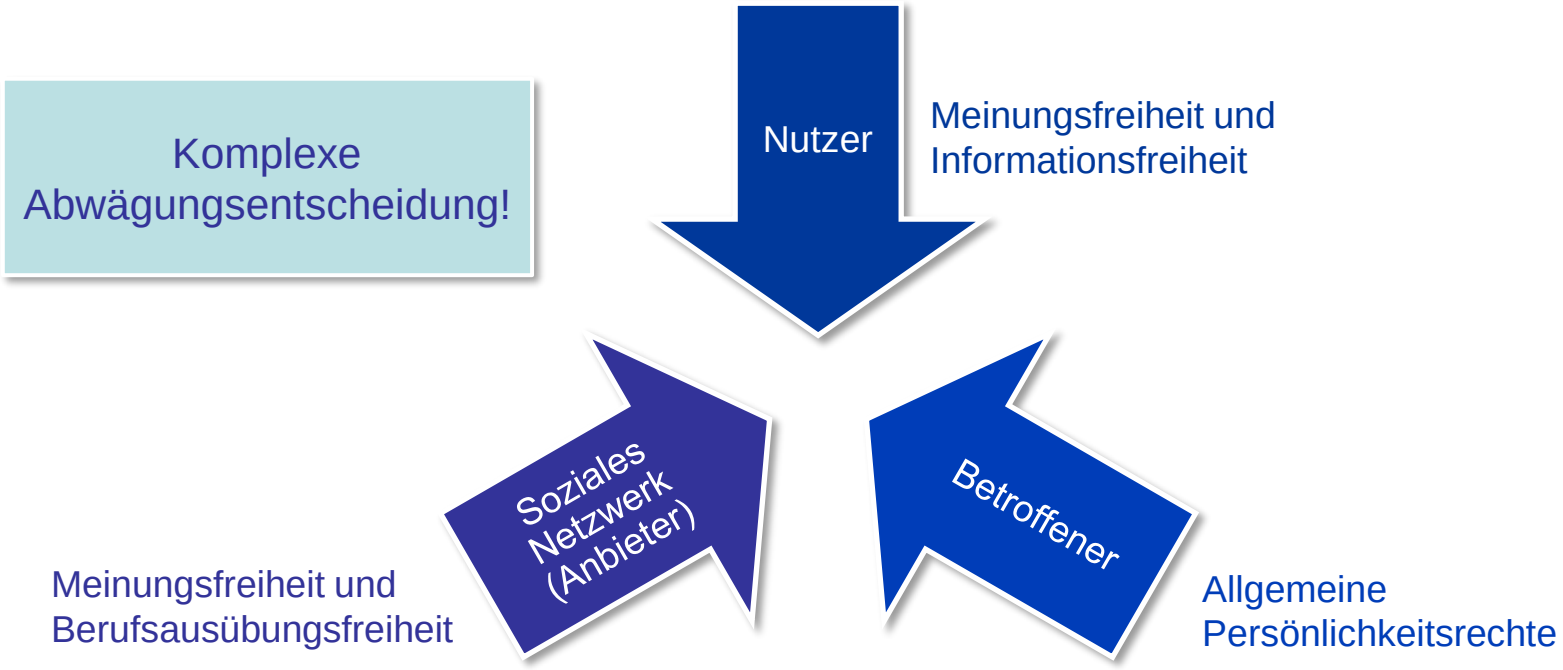
# II. Weitergehende Löschung von Äußerungen nach „Verhaltensrichtlinien“

- Praxis der Anbieter: Alle Löschungen nach „Verhaltensrichtlinien“/“Community Guidelines“

  → durch den Nutzungsvertrag einbezogen, werden vorrangig zu nationalem Recht geprüft
- Gefahr: Zu weitreichende Löschungen von Äußerungen („Overblocking“)
- Nach BGH, Urt. 29. Juli 2021 – III ZR 192/20 und III ZR 179/20 dürfen soziale Netzwerke auch Inhalte löschen, die nicht rechtswidrig sind.
- Aber: Nutzer muss von willkürlicher Ungleichbehandlung geschützt werden und Meinungsfreiheit muss bei Moderation beachtet werden (mittelbare Grundrechtswirkung)

# II. Weitergehende Löschung von Äußerungen nach „Verhaltensrichtlinien“

Die Anbieter müssen folgende Grundrecht in Ausgleich bringen (praktische Konkordanz) bei der Erstellung von Verhaltensrichtlinien und bei Löschung einzelner Äußerungen:



# II. Weitergehende Löschung von Äußerungen nach „Verhaltensrichtlinien“

Kriterien bei der Abwägung:

- Unterscheidung: Tatsachen und Werturteile
- Welche Persönlichkeitssphäre betroffen (Intimsphäre, Privatsphäre, Sozialsphäre, Öffentlichkeitsphäre)?
- Selbstöffnung des Betroffenen?
- Relevanz für den öffentlichen Diskurs
- Berücksichtigung des Umgangstons in sozialen Netzwerken

# II. Weitergehende Löschung von Äußerungen nach „Verhaltensrichtlinien“

Folgen bei rechtlich unzulässiger Löschung von zulässigen Äußerungen:

- Anspruch des Nutzers wegen vertraglicher Pflichtverletzung gemäß § § 280 Abs. 1, 249 BGB: Freischaltung/Wiederherstellung der Äußerung (Put-Back-Anspruch)

→ Spannungsverhältnis für Anbieter:

  – Inhalte entweder entfernen und dabei vertragswidriges Verhalten in Kauf zu nehmen oder
  – Äußerung belassen und Risiko eines Bußgeldes und einer Haftung in Kauf nehmen

# II. Weitergehende Löschung von Äußerungen nach „Verhaltensrichtlinien“

## Regelungen in Art. 14 DSA zur Moderationspraxis

- Im Rahmen der Moderation dürfen Inhalte gelöscht oder die Sichtbarkeit von Inhalten beschränkt, diese herabgestuft oder demonetisiert werden (Art. 3 lit. t DSA).
- Es dürfen auch nicht-rechtswidrige Inhalte gelöscht werden (z.B. Fake News)
- Art. 14 Abs. 4 DSA verpflichtet Anbieter dazu, die Rechte und berechtigten Interessen aller Beteiligter sowie die geltenden Grundrechte der Nutzer in der GRCh zu berücksichtigen, z.B. die Meinungsfreiheit → Abwägung

# II. Weitergehende Löschung von Äußerungen nach „Verhaltensrichtlinien“

## Verhältnis von bisheriger Rechtsprechung zum NetzDG und dem neuen DSA

- (P) bisher keine grundlegende mittelbare Drittwirkung auf europäischer Ebene anerkannt
- Fragestellungen an die Rechtsprechung:
  - Soll mit Art. 14 Abs. 4 DSA erstmals eine mittelbare Drittwirkung auf Unionsebene etabliert werden?
  - Soll die Reichweite der Grundrechte den Nationalstaaten obliegen oder soll der Maßstab der Union anzuwenden sein?
  - Kann die deutsche Rechtsprechung zur Moderationspraxis weiter Bestand haben oder muss die Rechtsprechung neu ausgerichtet werden?
  - Wie ist das Verhältnis zur deutschen Störerhaftung?

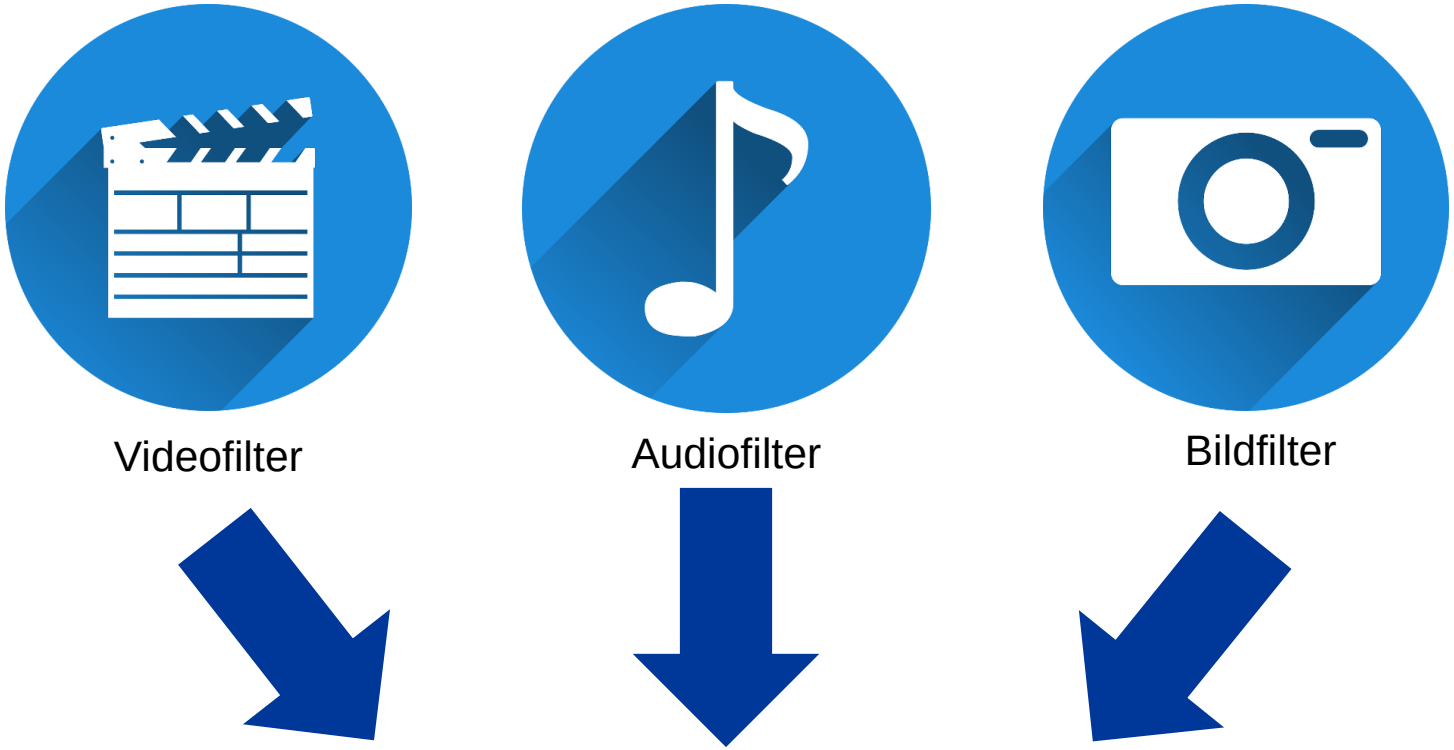
# Fazit von Teil 1

- Deutsche Rechtsprechung hat unter Geltung des NetzDG Regelungen zur Löschpraxis in sozialen Netzwerken entwickelt
- NetzDG wird durch den DSA unanwendbar (Anwendungsvorrang des Unionsrechts)

→ Aufgabe an Rechtsprechung: Kann die Löschpraxis auf die Anwendung des DSA übertragen werden und inwieweit müssen die Grundrechte beachtet werden?

# Teil 2

# I. Derzeitiger Einsatz von KI bei Löschungen in sozialen Netzwerken




Intelligente Filtersysteme: Bei Upload (Uploadfilter) und bei Meldung durch Nutzer im Urheberrecht

# II. Unterschiedliche Systeme künstlicher Intelligenz

(P) Kann nicht eingesetzt werden: Rein textbasierte Abgleichsysteme (ohne Text- und Kontextverständnis)
Ausnahme: Auffinden wortgleicher Äußerungen

→ Komplexe Abwägungsentscheidung unter Berücksichtigung des Äußerungskontextes erforderlich

Allein anwendbar: Künstliche Intelligenz mit automatischer Textqualifizierung, die Einordnung in Äußerungskontext vornimmt und umfassendes Textverständnis entwickelt

→ Derzeit wird KI vor allem bei Vorfilterung/Vorsortierung eingesetzt

# III. Rechtliche Voraussetzungen zum Einsatz von KI zur automatisierten Filterung von Äußerungen

## 1. Verstoß gegen Art. 22 Abs. 1 DSGVO?

→ Nein, Art. 22 Abs. 3 DSGVO: Durch Beschwerdeverfahren (Art. 20 DSA) kann Moderationsentscheidung angefochten werden.

## 2. „Automatisierte Mittel“ in Art. 16 Abs. 6 S. 2 DSA und Art. 20 Abs. 6 DSA

- Art. 16 Abs. 6 S. 2 DSA: Im Meldeverfahren grds. zulässig.
- Art. 20 Abs. 6 DSA: Im internen Beschwerdemanagementsystem: Nicht zulässig
- Löschungen auf Grundlage von Verhaltensrichtlinien: keine Regelungen

# III. Rechtliche Voraussetzungen zum Einsatz von KI zur automatisierten Filterung von Äußerungen

## 3. Künstliche Intelligenz = „automatisierte Mittel“?

- Art und Ausmaß nicht geregelt
- Automatisierte Mittel weit zu verstehen: Umfasst Abgleichsysteme ohne Text- und Kontextverständnis als auch künstliche Intelligenz mit Textqualifizierung

## 4. Technische Anforderungen

- Bei „rechtswidrigen Äußerungen“ - Art. 16 Abs. 6 S. 1 DSA: „*sorgfältig, frei von Willkür und objektiv*“
- Bei Löschungen aufgrund von Verhaltensrichtlinien – Art 14 Abs. 4 DSA: „*sorgfältig, objektiv und verhältnismäßig*“ unter Berücksichtigung der Rechte und berechtigten Interessen sowie der Grundrechte, insbesondere freie Meinungsäußerung.

# IV. Umsetzung der Anforderungen an KI

- Differenzierung nach konkret eingesetztem „automatisierten System“/KI
- Hohe Anforderungen bei automatisierter Löschung von Äußerungen:
  - Komplizierte Abwägungsentscheidung im Einzelfall
  - Meinungsfreiheit schlechthin konstituierend für freiheitlich-demokratische Grundordnung
  - Zentrale Rolle von sozialen Netzwerken für öffentliche Meinungsbildung
  - Fehlende Kontextsensibilität, Erkennung von Ironie und Satire nicht fehlerfrei
  - Gegenüber menschlicher Kontrolle wohl noch erhöhte Fehleranfälligkeit: Gefahr Overblocking

→ Pflicht für Anbieter: Nachweis, dass eingesetzte KI Anforderungen bei der Löschung von Äußerungen erfüllt

# V. Nachweis erfolgt über Transparenzanforderungen

- Ohne Nachweis muss eine menschliche Nachkontrolle vor Löschung erfolgen.
- Transparenzpflichten (Art. 15, 24 und 42 DSA): Aussagekräftige Angaben zur Moderation, einschließlich der Nutzung automatisierter Tools
    - Bei sehr großen Online-Plattformen: Zusätzlich jährliche Risikobewertung, auch algorithmischer Systeme
- So viele Informationen zu der Funktionsweise von KI mitteilen, dass gerichtliche Überprüfung möglich ist.
- Derzeit wird dies von noch keinem sozialen Netzwerk ausreichend differenziert dargelegt.

→ Aufgabe an Rechtsprechung: Konkretisierung der konkreten Nachweispflichten/Transparenzpflichten

# VI. Ideen zum Nachweis (Konkretisierung durch Rspr.)

- Verbindlichen Qualitätsmaßstab für die Entscheidung festlegen: Durchführung von Testreihen mit juristisch qualifizierte menschliche Content-Moderatoren
- KI darf nicht wesentlich vom Qualitätsmaßstab abweichen
- Bei Äußerungen gibt es viele Grenzfälle, in denen sich kein klares Bild ergibt, gewisse Unsicherheit muss akzeptiert werden
- Bei klar rechtswidrigen Inhalten: Höhere Anforderungen an die KI-Systeme
- Gleiche Maßstäbe an die Begründung der Entscheidung durch KI (Art. 17 DSA).

# Fazit von Teil 2

- Löschungen von Äußerungen durch KI: Pflicht für Anbieter zum Nachweis eines sorgfältiges, willkürfreies und objektives Verfahren, das Grundrechte und Nutzerinteressen ausreichend berücksichtigt.
- Nachweis kann im Rahmen der Erfüllung von Transparenzpflichten erfolgen (z.B. in Transparenzberichten)
- Konkrete Anforderungen an Nachweis müssen durch Rspr. ausgestaltet werden
- Derzeit: Kein ausreichender Nachweis durch die Anbieter

→ Bis auf Weiteres: Pflicht zur menschlichen Nachkontrolle bei Einsatz künstlicher Intelligenz mit Textqualifizierung bei Äußerungen in sozialen Netzwerken

# Vielen Dank

Nils Eckert

PLANIT // LEGAL
Rechtsanwaltsgesellschaft mbH
Jungfernstieg 1
20095 Hamburg