Hate Aid

SAFETY by DESIGN

Wege zu sichereren sozialen Netzwerken

Inhalt

Gefördert durch

Michael Otto Foundation for Sustainability

Schöpflin Stiftung

Paradigmenwechsel: Was ist Safety by Design?

Von schnellen Lösungen zu langfristigen Ansätzen

In der Europäischen Union (EU) verlassen sich Verbraucher*innen darauf, dass Gesetze sie im Alltag vor unsicheren Produkten schützen. Verkehrsregeln schreiben vor, dass Autos mit Sicherheitsgurten, Airbags und automatischen Notbremssystemen ausgestattet sein müssen. Arzneimittelgesetze verlangen, dass neue Medikamente gründlich getestet werden, bevor sie auf den Markt kommen. Produktsicherheitsgesetze verbieten die Verwendung bestimmter Chemikalien bei der Herstellung von Spielzeug. In jeder Phase des Lebenszyklus eines Produkts – von der Gestaltung über die Herstellung bis hin zum Vertrieb – steht die Verbrauchersicherheit im Mittelpunkt. Dies wird als „Safety by Design“ (SbD) bezeichnet.

Bei digitalen Räumen wie Plattformen hinken die europäischen Gesetze dem rasanten technologischen Wandel jedoch hinterher. Obwohl diese Plattformen zu zentralen Orten unseres öffentlichen Diskurses, unserer sozialen Interaktion und unseres Nachrichtenkonsums geworden sind,[1] stellt ihr Design oft die Maximierung der Nutzungszeit über die Sicherheit der Nutzenden:[2] Algorithmen schüren Empörung, Gewalt und Desinformation;[3] Autoplay- und Infinite-Scroll-Funktionen fördern bei Kindern suchtartiges Verhalten;[4] und dopaminverstärkte Belohnungssysteme wie „Likes“ und „Reaktionen“ befeuern soziale Ängste und Selbstwertprobleme.[5]

Obwohl diese negativen Auswirkungen seit Jahren bekannt sind, sträuben sich Social-Media-Plattformen weiterhin, Änderungen am Design vorzunehmen – aus Angst, dass dies ihrem Geschäftsmodell schaden könnte. Stattdessen neigen sie dazu, belastende Befunde zu verschleiern,[6] und verlassen sich auf die willkürliche Anwendung von reaktiven Sicherheitsmaßnahmen, die weder Schäden verhindern noch deren Folgen angemessen begegnen können. Zwar tragen Maßnahmen wie das Entfernen illegaler Inhalte, die Überprüfung irreführender Behauptungen oder die Sperrung von Troll-Accounts zweifellos zur allgemeinen Sicherheit digitaler Plattformen bei, doch dieser reaktive Ansatz hat drei zentrale Nachteile:

1. **Nutzende müssen erst eine Schädigung erfahren**, bevor etwas unternommen wird. Da die meisten digitalen Plattformen mögliche Schäden nicht ausreichend antizipieren, werden ihre Nutzenden mitunter zwangsläufig zu Versuchspersonen für Designentscheidungen. Ein prominentes Beispiel sind die jüngsten Änderungen von X am Chatbot Grok, die zu Millionen Fällen von sexualisierten Deepfakes führten.[7] Das Fehlen präventiver Maßnahmen ist dabei besonders gefährlich für schutzbedürftige Nutzende wie Kinder oder Menschen mit Traumata, die bei der Konfrontation mit schädlichen Inhalten dauerhaften Schaden nehmen können.[8]

2. **Soziale Netzwerke reagieren nur** auf gemeldete Probleme. Dies überträgt die Verantwortung auf die Nutzenden und deren Bereitschaft, Vorfälle zu melden. Viele z. B. von Gewalt Betroffene reichen jedoch aus Scham, Angst vor Repressionen oder Überforderung keine Meldungen ein. Diese Zurückhaltung wird oft durch Meldewege verstärkt, die absichtlich so gestaltet sind, dass sie Nutzende verwirren oder von Meldungen abhalten.[9] Besonders besorgniserregend ist, dass viele potenziell schädliche Auswirkungen von digitalen Plattformen, wie Ängste, Sucht oder Depressionen, gar nicht erst gemeldet werden können.

3. **Viele Plattformen zögern**, konsequent zu handeln – selbst nach der Meldung von missbräuchlichen Inhalten. Eine aktuelle Erhebung von HateAid zeigt, dass 55 % der gemeldeten illegalen Inhalte auf sozialen Netzwerken online blieben.[10] Auch verwirrt die inkonsequente Anwendung von Moderationsregeln Nutzende und zeigt die Willkür der existierenden Sicherheitskonzepte.

Statt reaktiver Schadensbegrenzung braucht es daher proaktive Schadensvermeidung und -minimierung. Sicherheit muss zu einem integralen Bestandteil des Netzwerkdesigns werden. Dies ist kein revolutionärer Ansatz, sondern gängige Praxis im Bereich der Produktsicherheit.

Ende 2025 stattete X seinen KI-Chatbot Grok mit einer neuen Bilderstellungsfunktion aus, die es Nutzenden ermöglichte, sexualisierte Bilder von Frauen und Mädchen – einschließlich Minderjähriger und Personen des öffentlichen Lebens – zu erstellen. Innerhalb von neun Tagen generierte Grok mehr als 1,8 Millionen sexualisierte Deepfakes sowie andere schädliche und extremistische Inhalte.[11,12]

Mehrere Länder begannen mit Untersuchungen oder verhängten ein vollständiges Verbot von Grok, woraufhin X den Zugang hinter einer Paywall einschränkte.[13] **Trotz gegenteiliger Behauptungen von X erzeugt Grok weiterhin nicht-einvernehmliche sexualisierte Deepfakes in der EU, dem Vereinigten Königreich und den USA.**[14] **Das Unternehmen trägt somit weiterhin zur Förderung und Kommerzialisierung geschlechtsspezifischer Gewalt bei.**

Um dies zu ändern, hat HateAid zwei Expertengutachten in Auftrag gegeben, die eine Reihe von konkreten technischen und rechtlichen Maßnahmen zur Minimierung des Risikos von Online-Schäden und der Steigerung der allgemeinen Plattformsicherheit aufzeigen. Die Analyse fokussiert sich dabei auf allgemein bekannte, global agierende digitale Plattformen. Zudem wurde im Rahmen dieser Untersuchung eine umfassende Taxonomie erstellt, die einen strukturierten Überblick über hunderte SbD-Designelemente und -prinzipien bietet. Sie umfasst Designmaßnahmen, die bereits weltweit verbreitet sind; solche, die in der Vergangenheit genutzt wurden; und neue Ansätze, die von Expert*innen empfohlen werden. Die Taxonomie und die wissenschaftliche Expertise sollen Regulierungsbehörden und (entstehenden) sozialen Netzwerken als Leitfaden für mögliche technische Lösungen und strukturelle Präventivmaßnahmen dienen. Diese sollen die Sicherheit der Nutzenden in den Mittelpunkt des Plattformdesigns stellen.

Prof. Dr. Michael Denga ist Zivil- und Handelsrechtler mit besonderem Schwerpunkt auf der Regulierung digitaler Technologien. Seine Forschung verbindet zentrale Bereiche des deutschen Zivil- und Handelsrechts mit europäischen Perspektiven und behandelt Themen wie soziale Netzwerke, Daten und KI. Seit 2025 hat er den Lehrstuhl für Bürgerliches Recht und Handelsrecht an der BSP Business and Law School Berlin inne.

Hier geht es zum Gutachten.

Caroline Sinders forscht und engagiert sich kreativ im Bereich Machine-Learning-Design. In den letzten Jahren hat Sinders die Schnittstellen von Technologieauswirkungen auf die Gesellschaft, Interface-Design, künstliche Intelligenz, Missbrauch sowie Politik in digitalen Interaktionsräumen untersucht. Sinders gründete und leitet Convocation Research + Design (CoRD Labs), ein Forschungs- und Technologielabor für Menschenrechte.

Hier geht es zum Gutachten.

Die Prinzipien von Safety by Design

Das SbD-Konzept wird seit über zehn Jahren erforscht. Obwohl es keine allgemein anerkannte Definition gibt, besteht Konsens über die zugrundeliegenden Prinzipien.[15] Für diese Veröffentlichung stützen wir uns auf die Definition des Forschungsinstituts CoRD Labs:

> **„Safety by Design ist ein Ansatz und eine Methodik, einschließlich einer zugehörigen Taxonomie, die die Sicherheit von Einzelpersonen, Gruppen, Kollektiven und Gemeinschaften von Beginn an in den Mittelpunkt der Technologieentwicklung, des Designs und des Ideenfindungsprozesses von Software- und Hardwareentwicklung stellt.“** *(CoRD Labs, 2026)*

CoRD Labs erklärt, dass SbD über isolierte Sicherheits- und Datenschutzmaßnahmen hinausgeht, um alle Arten von Risiken anzugehen (**Sicherheit**). Dies erfordert, aufkommende Risiken während des gesamten Lebenszyklus eines sozialen Netzwerks konsequent zu bewerten und ihnen entgegenzuwirken (**Proaktivität**). Um sicherzustellen, dass bestimmte Risiken nicht übersehen werden, stellt SbD die Erfahrungen derer in den Mittelpunkt, die am stärksten von Risiken betroffen sind (**Intersektionalität**). Es stellt auch sicher, dass Nutzende verstehen (**Lesbarkeit**), wie ihre Daten, Inhalte und Netzwerkinteraktionen verarbeitet werden (**Transparenz**) – selbst, wenn sie mit Sprach- oder anderen Zugangsbarrieren konfrontiert sind (**Zugänglichkeit**). Folglich ermöglicht SbD allen Nutzenden, weitreichende Entscheidungen über ihre Sicherheit, Privatsphäre und Teilhabe online zu treffen (**Handlungsfähigkeit**).[16]

> **„Sicherheit ist nicht die Abwesenheit von Risiko, sondern das Vorhandensein von Bedingungen, die es Menschen ermöglichen, sich auszudrücken, mit anderen in Kontakt zu treten und an digitalen Räumen selbstbestimmt teilzunehmen.“**
> *(CoRD Labs, 2026)*

Diese Prinzipien stimmen mit regulatorischen Rahmenwerken in Australien und dem Vereinigten Königreich überein, wo SbD in die nationale Gesetzgebung zur Netzwerksicherheit miteingebettet ist.[17, 18] Obwohl der Digital Services Act (DSA) der EU SbD nicht explizit erwähnt, enthält er ähnliche Anforderungen, wie die Minderung systemischer Risiken durch Plattformdesign,[19] nutzerfreundliche Meldewege und die Kennzeichnung automatisierter Löschentscheidungen.[20] Da SbD-Prinzipien zunehmend an regulatorischer Bedeutung gewinnen, sollte sich der Fokus nun auf die Etablierung von Best Practices und die Verbesserung der praktischen Umsetzung von Sicherheitsverpflichtungen richten.

Unsere Safety-by-Design-Taxonomie

Im Auftrag von HateAid hat CoRD Labs eine umfassende SbD-Taxonomie erstellt. Sie umfasst 214 konkrete Maßnahmen (**Interventions**), die in 30 Kernprinzipien (**Attributes**), 62 Designmodelle (**Design Patterns**) und 39 technische Werkzeuge (**Components**) unterteilt sind. Im Folgenden werden acht anschauliche Beispiele für effektive SbD-Maßnahmen vorgestellt.

Hier finden Sie die gesamte Taxonomie online.

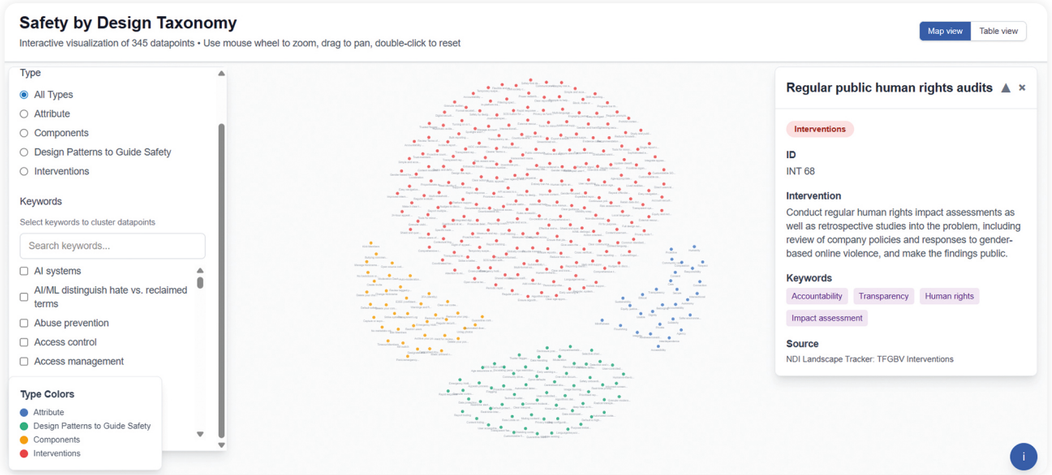

Abbildung 1 - Screenshot der Taxonomie. Die interaktive Visualisierung besteht aus 345 Datenpunkten.

Viele der aufgelisteten Maßnahmen haben sich bereits bewährt, da sie seit Jahren auf sozialen Netzwerken erprobt und getestet wurden. Zum Beispiel:

Reibungs- und Nudging-Funktionen
Instagram nutzt automatisierte Systeme, um Nutzende zu erkennen, die möglicherweise anstößige Inhalte posten möchten. Die App zeigt dann Warnungen an und informiert Nutzende darüber, dass wiederholte Verstöße gegen Community-Richtlinien zur Kontosperrung führen können. Auf ähnliche Art und Weise ermutigte Twitter Nutzende im Zuge der US-Präsidentschaftswahlen 2020 dazu, Artikel zunächst zu lesen, bevor sie diese teilen, um das Risiko der Verbreitung von Desinformationen zu minimieren. Durch den Einsatz solcher Tools könnten digitale Plattformen die Verbreitung von Belästigungen und Desinformation proaktiv eindämmen.[21]

Tools gegen digitale Belästigung
Im Jahr 2022 stellte der Google-Inkubator Jigsaw einen Open-Source-Belästigungsmanager für Twitter vor. Das Tool war explizit darauf ausgelegt, weibliche Journalistinnen zu unterstützen, die oft einem hohen Maß an geschlechtsspezifischer digitaler Gewalt ausgesetzt sind. Mithilfe des Tools konnten Betroffene „schädliche Posts leicht identifizieren und dokumentieren, Belästiger stummschalten oder blockieren und belästigende Antworten auf ihre Tweets verbergen“.[22] Aufgrund von Änderungen in der Betriebsweise von Twitter/X ist das Tool leider nicht mehr aktiv.[23] Eine neue Version dieses Tools wäre ein wertvolles Hilfsmittel für diejenigen, die mitunter am häufigsten große Wellen von Online-Belästigung erfahren.

Sicherheitsmodi
Funktionen wie Xs „Schütze meine Posts“ und Instagrams „Privatmodus“ ermöglichen es Nutzenden, die Sichtbarkeit ihres Profils mit nur einem Klick zu begrenzen. Dies gibt ihnen mehr Kontrolle über ihre Privatsphäre und schützt sie vor unerwünschten Interaktionen mit Fremden.[24] Diese Modi können auch als Notfallknopf für Nutzende dienen, die einer plötzlichen Welle von Online-Belästigung oder Desinformation ausgesetzt sind. Um die Nützlichkeit dieser Funktion weiter zu verbessern, sollten soziale Netzwerke anpassbare Sicherheitsmodi anbieten, die differenzierte Einstellungen zulassen.[25]

Kontokennzeichnung
Soziale Netzwerke experimentieren seit langem mit Kennzeichnungen, um Desinformation zu bekämpfen und mehr Transparenz zu schaffen. So führte Twitter schon 2009 blaue Häkchen ein, um Institutionen und Personen des öffentlichen Lebens zu verifizieren. Im Jahr 2023 begann das Unternehmen, nun X, mit dem Verkauf dieser Häkchen, wodurch sein bewährtes Verifizierungssystem unwirksam wurde. Stattdessen führte X eine Kennzeichnungspflicht für Parodie-Konten ein, um satirische Inhalte zu markieren.[26] YouTube kennzeichnet staatlich finanzierte Nachrichtenkanäle seit 2018,[27] wenn auch nicht konsistent.[28] Kennzeichnungen für verifizierte oder vertrauenswürdige Quellen können Nutzenden helfen, die Glaubwürdigkeit von Inhalten zu bewerten und die Verbreitung von Desinformation zu reduzieren.

Weitere Maßnahmen in unserer Taxonomie umfassen neue Ansätze, die bestehende Mängel im Design von sozialen Netzwerken angehen:

Netzwerkübergreifende Zusammenarbeit
„Online-Gewalt verbreitet sich oft von Plattform zu Plattform und nutzt dabei deren jeweilige Schwächen aus."[29] Menschen, die Cybermobbing und Stalking betreiben, verfolgen ihre Opfer beispielsweise meist gleichzeitig auf mehreren sozialen Netzwerken. Betroffene würden daher sehr von der Möglichkeit profitieren, missbräuchliche Schlüsselwörter, Inhalte und Nutzende plattformübergreifend zu blockieren. Ebenso würden netzwerkübergreifende Doxing-Alarmsysteme dabei helfen, die Verbreitung geleakter persönlicher Informationen schnell einzudämmen, sobald sie auf einem sozialen Netzwerk erkannt wurden.

Schnellreaktionsteams
Soziale Netzwerke sollten es Nutzenden ermöglichen, Schnellreaktionsteams aus zuvor festgelegten Verbündeten zusammenzustellen, die bei der Bewältigung von Online-Belästigung helfen können. Nutzende sollten diesen Verbündeten eingeschränkten Kontozugriff gewähren können – ähnlich wie bei Gmails Delegationsfunktion –, um Belästigung zu sichten, zu dokumentieren und zu melden. Zu diesem Zweck könnten soziale Netzwerke auf bestehenden Funktionen wie Xs „Teams"-Funktion (derzeit nur für X Pro verfügbar) oder Instagrams „Rollen" (aktuell nur für geteilte Instagram-Geschäftskonten verfügbar) aufbauen.[30]

Ein-Klick-Dokumentationstool
Online-Belästigung muss ordnungsgemäß dokumentiert werden, um strafrechtlich verfolgt werden zu können. Doch viele Nutzende wissen nicht, wie sie rechtssichere Screenshots erstellen können, sind emotional belastet oder stehen unter Zeitdruck, da belästigende Inhalte schnell gelöscht oder entfernt werden können. Um sicherzustellen, dass Online-Belästigung nicht undokumentiert bleibt, könnten soziale Netzwerke eine Ein-Klick-Dokumentationsfunktion integrieren, die belästigende Posts automatisch erfasst, zusammen mit essenziellen Metadaten und jurisdiktionsspezifischen Details, wenn Nutzende Inhalte melden oder blockieren.[31]

Quarantäne für zu prüfende Inhalte
Informationen verbreiten sich rasend schnell im Internet. Dazu gehören Desinformation, persönliche Daten wie eine private Adresse oder Links zu geleaktem Material. Gemeldete Inhalte unter Quarantäne zu stellen, bis sie überprüft wurden, stellt daher einen wesentlichen Schritt dar, um die Sichtbarkeit und Verbreitung schädlicher Informationen zu verringern. Opfer großflächiger Online-Belästigung würden auch von einem Quarantäne-Dashboard profitieren, das es ihnen ermöglicht, potenziell belästigende Inhalte sicher mit Hilfe vertrauenswürdiger Verbündeter zu scannen.[32]

Die Taxonomie soll Entscheidungsträger*innen und Plattformen einen klaren und praxisnahen Überblick über sowohl etablierte als auch innovative SbD-Funktionen bieten und als Grundlage für den Aufbau sicherer Onlineplattformen dienen. Es ist jedoch wichtig zu bedenken, dass SbD nicht nur das Hinzufügen von Funktionen oder Durchsetzung der Einhaltung von Bestimmungen bedeutet. Vielmehr erfordert es die proaktive und kontinuierliche Anpassung von Sicherheitsfunktionen an eine sich ständig ändernde Online-Umgebung, wobei die Erfahrungen und Bedürfnisse der am stärksten betroffenen Nutzenden im Mittelpunkt stehen.

> **„Echtes Safety by Design bedeutet, vorherzusehen, wie Systeme instrumentalisiert werden können [...]. Es beinhaltet auch den Aufbau von Schutzmechanismen, bevor sich die ersten Nutzenden überhaupt anmelden, und die Echtzeit-Reaktion auf Schäden, um diese leichter und in größerem Umfang zu erkennen und zu mildern."** *(CoRD Labs, 2026)*

Umsetzung von Safety-by-Design-Maßnahmen

Hindernisse bei der Umsetzung

Warum setzen die meisten sozialen Netzwerke ihre Nutzenden stattdessen Risiken aus, obwohl etliche Sicherheitsfunktionen existieren? Basierend auf den Aussagen der konsultierten Expert*innen können drei zentrale Gründe für die mangelnde Bereitschaft der Plattformen, SbD-Maßnahmen umzusetzen, hervorgehoben werden:

Geschäftsmodelle sozialer Netzwerke

Soziale Netzwerke generieren Einnahmen durch das Sammeln und Monetarisieren der Daten ihrer Nutzenden. Um ihre Gewinne zu maximieren, streben sie ständig danach, die Anzahl der Nutzenden und deren Interaktion auf dem sozialen Netzwerk zu erhöhen. Je mehr Nutzende und je länger sie mit einem Netzwerk interagieren – durch (Dis)Likes, Kommentare und Posts –, desto mehr Daten können gesammelt und desto besser können Werbeanzeigen personalisiert werden. Daher fördern soziale Netzwerke virale Inhalte, die oft kontrovers, unwahr oder sogar schädlich, aber äußerst ansprechend sind. Sie nutzen auch suchtfördernde Designelemente, die Nutzende an ihre Bildschirme fesseln.[33]

SbD stellt dieses Gewinnmodell direkt infrage, indem es das Wohlbefinden und die Sicherheit der Nutzenden über ihr Interaktionslevel und ihre Bildschirmzeit stellt. Neben der Begrenzung der Reichweite polarisierender und schädlicher Inhalte fördert SbD auch Privatsphäre und Transparenz und befähigt Nutzende, das Tracking ihres Online-Verhaltens einzuschränken. Es sind nicht nur die finanziellen Kosten der Einführung neuer Sicherheitsfunktionen, sondern der potenzielle Verlust von Nutzendendaten, der mit den kommerziellen Interessen großer sozialer Netzwerke in Konflikt steht.

Machtasymmetrien und Vorurteile

Ein weiteres maßgebliches Hindernis sind Machtasymmetrien innerhalb der internen Designprozesse von digitalen Plattformen. Designteams treffen Entscheidungen oft auf der Grundlage ihrer Annahmen über „durchschnittliche Nutzende". Dies verstärkt bestehende Vorurteile und vernachlässigt typischerweise marginalisierte, weniger monetarisierbare Gruppen, die dem höchsten Schadensrisiko ausgesetzt sind.[34] Im Gegensatz zur Mehrheit der Nutzenden sind die Verantwortlichen hinter digitalen Plattformen eine relativ homogene Gruppe meist männlicher, technikaffiner und hochbezahlter Fachleute, die bestimmte Werte und Interessen teilen. Anders als bei Ärzt*innen oder Anwält*innen gibt es weder einen Eid noch einen Qualifikationsnachweis oder einen Branchenkodex, der die Designer*innen sozialer Netzwerke verpflichtet. Dementsprechend fehlen den Verantwortlichen klare Anreize, die Bedürfnisse von Gruppen, die nicht den eigenen entsprechen, in der Gestaltung ihrer sozialen Netzwerke zu berücksichtigen.

Als Beispiel dient TikToks jüngste Entscheidung, das Vertrauens- und Sicherheitsteam durch fehlerhafte automatisierte Moderationssysteme zu ersetzen.[35] Obwohl diese Designänderung das Schadensrisiko für Nutzende, insbesondere für diejenigen, die

marginalisierten Gemeinschaften angehören, massiv erhöht, wurde keine umfassende Risikoanalyse für verschiedene Zielgruppen durchgeführt. Durch die Fokussierung von Designentscheidungen auf die Erfahrungen der am stärksten Betroffenen droht SbD, diese Machtverhältnisse umzukehren.

Fehlende regulatorische Anreize

In der EU unterliegen die meisten Arten von Verbraucherprodukten strengen Sicherheitsstandards, um Schäden zu verhindern. Wenn Herstellende diesen nicht nachkommen, können sie von den durch ihre Nachlässigkeit Geschädigten zur Rechenschaft gezogen werden. Im Gegensatz dazu genossen digitale Plattformen lange Zeit weitreichende Haftungsprivilegien, um die freie Verbreitung nutzergenerierter Inhalte zu gewährleisten. Gleichzeitig versuchte die EU durch geringe Sicherheitsanforderungen mehr Online-Dienstleistende in die Union zu locken. Nach zwei Jahrzehnten minimaler Regulierung verabschiedete die EU den DSA, der deutlich strengere Sicherheitsstandards setzt. Dieser legt auch fest, dass digitale Plattformen nur so lange von der Haftung für nutzergenerierte Inhalte befreit sind, bis diese gemeldet werden.[36] Obwohl der DSA auch einzelne SdD-Maßnahmen beinhaltet, fehlt ihm jedoch ein umfassendes SbD-Mandat, das die einzelnen Maßnahmen sinnvoll miteinander verknüpft.

Dies hat es digitalen Plattformen ermöglicht, Verpflichtungen durch oberflächliche und partielle Rechtseinhaltung zu unterwandern. Nutzende werden aufgrund der Gestaltung der Meldewege beispielsweise davon abgebracht, schädliche Inhalte zu melden.[37] Ohnehin bleibt ein erheblicher Teil der gemeldeten Inhalte trotzdem online.[38] Weder die Transparenzberichte der sozialen Netzwerke noch die Transparenzdatenbank der Europäischen Kommission geben ausreichend Einblick in die inneren Arbeitsweisen von Empfehlungssystemen oder die Qualität der Inhaltsmoderation. Forschende, die Datenzugang beantragen, um die internen Prozesse sozialer Netzwerke auszuwerten, erhalten häufig Informationen, die zu unspezifisch oder unstrukturiert für wissenschaftliche Zwecke sind.[39] Externe Audits, die systemische Risiken bewerten, wie öffentliche Gesundheit oder Sicherheit demokratischer Wahlen, bleiben oberflächlich, unvollständig und ohne valide Beweise.[40] Ebenso haben die meisten großen Netzwerkbetreibenden bislang kaum Interesse daran gezeigt, die von ihren Diensten ausgehenden systemischen Risiken angemessen zu analysieren und auszubessern.

Obgleich einige europäische und nationale Aufsichtsbehörden Verfahren eingeleitet haben, um Rechtsverstöße zu ahnden, bestehen weiterhin erhebliche Durchsetzungsdefizite. Einerseits fehlen den Regulierungsbehörden schlicht das Personal, die Expertise und die finanziellen Mittel, um die große Anzahl von Beschwerden über Netzwerkfehlverhalten zeitnah zu bearbeiten. Dieses Problem wird durch sich überschneidende Mandate und regulatorische Rahmenwerke verschärft, die langwierige Koordinationsprozesse zwischen verschiedenen Gerichtsbarkeiten und Aufsichtsbehörden erfordern. Andererseits sehen sich Vollzugsbehörden zunehmend politischem Gegenwind ausgesetzt, da Plattformbetreibende und die US-Regierung die EU immer offensiver unter Druck setzen, in der Absicht, eine Rücknahme der EU-Netzwerkregulierung zu erwirken. Die Durchsetzung des DSA ist somit zu einem Spielball in geopolitischen Verhandlungen geworden, wobei die Frage der Rechtseinhaltung als Werkzeug in umfassenderen transatlantischen Streitigkeiten genutzt wird. Die Zukunft der digitalen Souveränität Europas hängt folglich vom Willen und der Fähigkeit der Regulierungsbehörden ab, ihr Regelwerk durchzusetzen. Sie müssen sich dabei gegen von der US-Regierung gestützte Plattformen behaupten, die sich nicht scheuen, die Gesetze der Länder zu brechen, in denen sie Geschäfte machen.[41]

Empfehlungen für politische Entscheidungsträger*innen

Europas Ansatz zur Netzwerkregulierung beruht auf seinem einzigartigen Einsatz für demokratische Werte, Rechtsstaatlichkeit und grundlegende Menschenrechte. Um diese Prinzipien zu verteidigen und die digitale Souveränität zu wahren, muss die EU Nutzendenrechte durchsetzen – sowohl gegen die finanziellen Interessen der Plattformbetreibenden als auch gegen transatlantischen politischen Druck. In seinem Rechtsgutachten betont Prof. Denga, dass dies keine bloße politische Entscheidung, sondern eine rechtliche Verpflichtung ist, vergleichbar mit der Pflicht der EU-Kommission, das Wettbewerbsrecht durchzusetzen. Er argumentiert, dass der Gesetzgeber den DSA vollständig umsetzen muss, da „das Effektivitätsinteresse europäischer Digitaler Souveränität zu einer erheblichen Ermessensreduktion der beauftragten Behörden, potentiell bis hin zur Aufgreifpflicht, [führt].“ [42]

Obwohl der DSA eine robuste rechtliche Grundlage bietet, hängt sein Erfolg davon ab, über bloße Rechtseinhaltung hinauszugehen. Ein ganzheitlicher SbD-Ansatz ist entscheidend, um systemische Risiken anzugehen, die der DSA allein nicht vollständig mildern kann. Um Regulierungsbehörden in diesem Übergang zu unterstützen, haben wir ein dreistufiges Programm entwickelt, das SbD operationalisiert und sicherstellt, dass Europas digitale Infrastruktur mit seinen Kernwerten übereinstimmt und einen globalen Standard für die verantwortungsvolle Governance digitaler Plattformen setzt.

Schritt 1: Entschlossene Durchsetzung bestehender Regeln

Die Rücknahme des TikTok Lite Rewards-Programms in Europa, das süchtig machendes Verhalten durch Gutscheine und finanzielle Entschädigungen fördert, zeigt, wie durch konsequente und schnelle Rechtsdurchsetzung großflächige Schäden verhindert werden können.[43] Der dringendste Schritt besteht daher in der umfassenden Durchsetzung des DSA, insbesondere seiner designbezogenen Elemente. So muss darauf geachtet werden, dass soziale Netzwerke nutzendenfreundliche Meldewege schaffen und illegale Inhalte konsequent und prompt entfernen. Es muss auch sichergestellt werden, dass Forschende nützliche Daten erhalten, aussagekräftige Transparenzberichte erstellt werden und soziale Netzwerke von unabhängigen Gutachter*innen geprüft werden. Durchsetzungsbehörden benötigen dringend mehr finanzielle und personelle Ressourcen, um die Rechtseinhaltung zu überwachen. Sie müssen auch vor politischer Einmischung geschützt werden. Zur Gewährleistung dessen wird eine unabhängige Durchsetzungsbehörde auf europäischer Ebene empfohlen. Dies könnte eine neue europäische Agentur oder sogar ein völlig neues EU-Gremium sein, das – wie die Europäische Zentralbank – seine Aufsicht ohne direkte politische Weisungen ausübt.[44]

Die Unterstützung durch die US-Regierung könnte bestimmte Plattformbetreibende dazu ermutigen, europäische Regeln offen zu missachten. In diesen Fällen ist es wichtig zu bedenken, dass die Durchsetzung von Netzwerkpflichten „nicht alleine durch die aufsichtsrechtliche und privatrechtliche Haftung der Anbieter von Vermittlungsdiensten, sondern auch durch die persönliche Haftung ihres Personals“[45] erfolgt. Obwohl sich DSA-Verpflichtungen an soziale Netzwerke und nicht an deren Mitarbeitende richten, weist Prof. Denga darauf hin, dass sich aus dem DSA dennoch eine persönliche Haftbarkeit für Rechtsverstöße ableiten lässt. Sie basiert auf den objektiven Qualitätsstandards der Vermittlungsdienste, die durch den DSA festgelegt werden, dem dringenden öffentlichen Interesse, das im Prinzip der europäischen digitalen Souveränität verankert ist, und dem haftungsrechtlichen Nutzen-Lasten-Paradigma.[46] Folglich könnte das Leitungspersonal digitaler Plattformen im Falle schwerwiegender Schäden und systemischen Versagens persönlich haftbar gemacht werden, insbesondere wenn grobe Fahrlässigkeit oder vorsätzliche Nichtbeachtung nachgewiesen werden kann.[47]

Schritt 2: Einführung umsetzbarer Sicherheitsstandards

Die Europäische Kommission,[48] das Europäische Parlament,[49] und der Europäische Rat[50] sind alle zu dem Schluss gekommen, dass die aktuellen Sicherheitsstandards unzureichend sind, um Nutzende online angemessen zu schützen. Dementsprechend ist neue und aktualisierte Gesetzgebung erforderlich, um unter anderem die Erstellung und Verbreitung von nicht-einvernehmlichen Deepfakes, süchtig machenden Designelementen wie Autoplay oder Endlos-Scrolling sowie Praktiken wie die Verwendung von Social-Bots und Shadow-Banning, die den öffentlichen Diskurs verzerren, zu verbieten.[51] Neben dem Verbot manipulierender oder schädlicher Praktiken müssen Entscheidungsträger*innen auch proaktiv spezifische Sicherheitsstandards und -maßnahmen vorschreiben. Dazu gehört die Einstufung von Netzwerk-Algorithmen und inhaltsgenerierenden KI-Tools wie Grok als Hochrisiko-KI-Systeme im Sinne des AI-Acts. Nur so kann eine Rechtseinhaltung mit Transparenzverpflichtungen wie Datenqualitätsprüfungen oder Bias-Monitoring sichergestellt werden.[52] Weitere Beispiele finden sich unter den 214 Maßnahmen, die in der SbD-Taxonomie aufgeführt sind und als Bezugspunkte dienen können.[53]

Um oberflächliche Rechtseinhaltung zu verhindern, muss neue Gesetzgebung individuelle Sicherheitsmaßnahmen mit einem übergeordneten SbD-Mandat untermauern. „Analog zu den ‚Fit & Proper'-Anforderungen im Kapitalmarktrecht sollten für Führungskräfte von Social-Media-Plattformen sowie deren lokale Repräsentanten spezifische Qualifikationsanforderungen etabliert werden, um sicherzustellen, dass diese den Herausforderungen digitaler Meinungsräume gerecht werden."[54] Darüber hinaus sollte das Mandat verfahrensrechtliche Garantien in den Designprozessen sozialer Netzwerke einführen, wie regelmäßige Konsultationen mit Interessensvertreter*innen, um deren Sicherheit, Transparenz, Lesbarkeit, Zugänglichkeit, Intersektionalität und Fokus auf Nutzendenfreundlichkeit zu gewährleisten. Angesichts der jüngsten Zunahme von SLAPP-Klagen durch Plattformbetreibende, Online-Gewalt und sogar Repressionsakten ausländischer Regierungen sind stärkere Schutzmaßnahmen für einzelne Forschende und zivilgesellschaftliche Organisationen gegen solche Angriffe erforderlich.

Schritt 3: Dezentralisierung der Netzwerk-Infrastruktur

In den letzten 20 Jahren haben Unternehmen ihre digitalen Plattformen so gestaltet, dass sie so aufmerksamkeitserregend und anziehend wie möglich sind. Dies ermöglichte es einigen kleinen Tech-Start-ups, sich zu mächtigen multinationalen Konzernen zu entwickeln. Es führte auch zur Ausbreitung digitaler Gewalt und Online-Desinformation. Statt dafür Verantwortung zu übernehmen und entsprechende Maßnahmen zu ergreifen, wälzen Plattformbetreibende die Schuld ab und verlagern die Last auf Strafverfolgungsbehörden, mit denen sie nur begrenzt zusammenarbeiten. Obwohl sie in der Lage wären das Design ihrer Plattformen anzupassen, um mehr Sicherheit zu gewährleisten, sind die meisten nicht bereit, ihre Gewinne zu gefährden.

Infolgedessen ziehen sich immer mehr Menschen aus dem öffentlichen Diskurs zurück, während das Profitinteresse einiger weniger Unternehmen die gesellschaftliche Spaltung vorantreibt. Dies hat schwerwiegende Folgen für unsere Demokratie und Sicherheit. Entscheidungsträger*innen sollten nach Wegen suchen, um alternative Netzwerk-Modelle zu fördern. Bestehende dezentralisierte, Open-Source-Netzwerke wie das Fediverse oder Eurosky bieten Alternativen zu etablierten Plattformen und operieren ohne überwachungsbasierte Werbung oder süchtig machende Algorithmen.[55] Diese nicht-kommerziellen Netzwerke haben jedoch Schwierigkeiten, privates Kapital anzuziehen. Daher sind mehr öffentliche Fördergelder für europäische Netzwerk-Initiativen und die Einführung rechtlich verankerter Portabilitätsrechte, die die Migration von Nutzenden zwischen Netzwerken erleichtern, erforderlich. Darüber hinaus ist das bestehende Haftungsregime für Hosting-Anbieter*innen nicht darauf ausgelegt, mit dezentralisierten Netzwerken umzugehen, und stellt rechtliche Herausforderungen und Unsicherheiten dar. Daher müsste die EU gezielt Sonderregeln einführen, um ihr Wachstum zu ermöglichen. Dies ist bislang in einem Rechtsrahmen, der vor allem die mächtigsten und missbräuchlichsten Akteure kontrollieren soll, kaum realisierbar. Ausnahmen für nicht-kommerzielle Plattformen und mehr regulatorische Sandboxes sind daher notwendig.

Durch die Förderung der Entwicklung und Verbreitung europäischer Netzwerk-Modelle, die die in der Taxonomie genannten SbD-Prinzipien umsetzen, kann die EU systemischen Online-Schäden effektiv begegnen und gleichzeitig ihre digitale Souveränität verteidigen. Wie in der jüngsten Resolution des Europäischen Parlaments festgestellt wurde, ist die Erhaltung der digitalen Souveränität Europas nicht nur eine technische Frage, sondern ein demokratisches Gebot.[56] In einer Ära, in der digitale Räume politische Realitäten prägen, ist demokratische Kontrolle über soziale Netzwerke entscheidend, um die europäische Autonomie und grundlegende Werte aufrechtzuerhalten. Es ist an der Zeit, dass die EU ihre Prinzipien in die Tat umsetzt und eine digitale Zukunft aufbaut, die ihr Engagement für Sicherheit, Rechenschaftspflicht und Rechtsstaatlichkeit widerspiegelt.

Quellenverzeichnis

1 Eurobarometer (2025): Social Media Survey 2025 (FL014EP), S. 12, https://europa.eu/eurobarometer/surveys/detail/3592 (letzter Zugriff: 31.01.2026).

2 Bhargava, Vikram R. & Manuel Velasquez (2021): Ethics of the Attention Economy: The Problem of Social Media Addiction, Business Ethics Quarterly, 31 (2021), 321–59. DOI: 10.1017/beq.2020.32.

3 Munn, Luke (2020): Angry by design: toxic communication and technical architectures, Humanities and Social Sciences Communications 7(1): 1–11. DOI: 10.1057/s41599-020-00550-7.

4 Mujica, Alejandro L., Crowell, Charles R.; Villano Michael A. & Uddin, Khutb M. (2022): Addiction by design: Some dimensions and challenges of excessive social media use, Medical Research Archives, 10(2), 1–29. DOI: 10.18103/mra.v10i2.2677.

5 Lee, Hae Yeon; Jamieson, Jeremy P.; Reis, Harry T.; Beevers, Christopher G.; Josephs, Robert A.; Mullarkey, Michael C.; O'Brien, Joseph & Yeager, David S. (2020): Getting fewer "Likes" than others on social media elicits emotional distress among victimized adolescents, Child Development, 91(6), 2141–2159. DOI: 10.1111/cdev.13422.

6 Horwitz, Jeff (2025): Meta buried ‚causal' evidence of social media harm, US court filings allege, Reuters, https://www.reuters.com/sustainability/boards-policy-regulation/meta-buried-causal-evidence-social-media-harm-us-court-filings-allege-2025-11-23/ (letzter Zugriff: 31.01.2026).

7 Conger, Kate; Freedman, Dylan & Thompson, Stuart A. (2026): Musk's Chatbot Flooded X With Millions of Sexualized Images in Days, New Estimates Show, New York Times, https://www.nytimes.com/2026/01/22/technology/grok-x-ai-elon-musk-deepfakes.html (letzter Zugriff: 31.01.2026).

8 McHugh, Bridget Christine; Wisniewski, Pamela; Rosson, Mary Beth & Carroll, John M. (2018): When social media traumatizes teens: The roles of online risk exposure, coping, and post-traumatic stress. Internet Research, 28(5), 1169–1188. DOI: 10.1108/IntR-02-2017-0077.

9 Das Nettz (2025): Between Click and Consequence: An Evaluation of Platform Reporting Procedures under the Digital Services Act, https://www.das-nettz.de/sites/default/files/2025-11/ENG_NETTZ_Meldewege_18.pdf (letzter Zugriff: 16.12.2025).

10 HateAid (2025): Rights Without Reach. The DSA Put to the Test, https://hateaid.org/wp-content/uploads/2025/12/hateaid-dsa-rights-without-reach-2025.pdf (letzter Zugriff: 31.01.2026).

11 CCDH (2026): Grok floods X with sexualized images of women and children, https://counterhate.com/research/grok-floods-x-with-sexualized-images/ (letzter Zugriff: 31.01.2026).

12 Bouchau, Paul (2026): Grok Generating Flood of Sexualized Images of Women and Minors, AI Forensics, https://aiforensics.org/work/grok-unleashed (letzter Zugriff: 31.01.2026).

13 Watkins, Ali (2026): Malaysia and Indonesia Block Access to Grok Because of Sexually Explicit Content, New York Times, https://www.nytimes.com/2026/01/11/world/asia/malaysia-indonesia-grok-ban.html (letzter Zugriff: 31.01.2026).

14 Lyons, Emmet (2026): X, Grok AI still allow users to digitally undress people without consent, as EU announces investigation, CBS News, https://www.cbsnews.com/news/x-grok-ai-imagery-elon-musk-eu-uk-us-regulation/ (letzter Zugriff: 31.01.2026).

15 Woods, Lorna (2024): Safety by Design, https://www.onlinesafetyact.net/analysis/safety-by-design/ (letzter Zugriff: 31.01.2026).

16 Sinders, Caroline; Valencia, Antonia & Smith, Sam (2026): Safety by Design: A comprehensive methodology and exploration on design, policy and technology interventions to generate better user safety on platforms, S. 22, https://hateaid.org/wp-content/uploads/2026/03/safety-by-design-research-report-platform-safety-interventions-caroline-sinders.pdf (letzter Zugriff: 31.01.2026).

17 eSafety Commissioner (2025): Safety by Design, https://www.esafety.gov.au/industry/safety-by-design (letzter Zugriff: 31.01.2026).

18 The Secretary of State for Science, Innovation and Technology (2025): Final Statement of Strategic Priorities for Online Safety, www.gov.uk/government/publications/statement-of-strategic-priorities-for-online-safety/final-statement-of-strategic-priorities-for-online-safety#safety-by-design (letzter Zugriff: 31.01.2026).

19 Vgl. Art. 35(1)(a), DSA.

20 Vgl. Art. 16 ff., DSA.

21 Jankowicz, Nina; Hunchak, Jillian; Pavliuc, Alexandra; Davies, Celia; Pierson, Shannon & Kaufmann, Zoë (2021): Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women Online, Wilson Center, https://www.wilsoncenter.org/publication/malign-creativity-how-gender-sex-and-lies-are-weaponized-against-women-online (letzter Zugriff: 31.01.2026).

22 Jigsaw (2022): Technology to help women journalists document and manage online abuse, Medium, https://medium.com/jigsaw/technology-to-help-women-journalists-document-and-manage-online-abuse-5edcac127872 (letzter Zugriff: 31.01.2026).

23 Thomson Reuters Foundation (2025): TRFilter, https://www.trfilter.org/ (letzter Zugriff: 31.01.2026).

24 Instagram aktiviert automatisch den privaten Modus für alle minderjährigen Nutzenden, um Grooming zu verhindern.

25 Chumsky, Susan (Hrsg.) (2021): No Excuse for Abuse. What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users, PEN America, https://pen.org/report/no-excuse-for-abuse/ (letzter Zugriff: 31.01.2026).

26 Maxwell, Thomas (2025): X Creates a New Parody Label for Accounts, Solving a Problem Elon Created, Gizmodo, https://gizmodo.com/x-creates-a-new-parody-label-for-accounts-solving-a-problem-elon-created-2000548621 (letzter Zugriff: 31.01.2026).

27 BBC (2018): YouTube to label government and public-funded clips, https://www.bbc.com/news/technology-46139189 (letzter Zugriff: 31.01.2026).

28 Kofman, Ava (2019): YouTube Promised to Label State-Sponsored Videos But Doesn't Always Do So, ProPublica, https://www.propublica.org/article/youtube-promised-to-label-state-sponsored-videos-but-doesnt-always-do-so (letzter Zugriff: 31.01.2026).

29 Sinders et al. (2026), S. 61.

30 Chumsky (2021).

31 Ebd.

32 Ebd.

33 Hagey, Keach & Horwitz, Jeff (2021): Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead., The Wall Street Journal, https://www.wsj.com/tech/facebook-algorithm-change-zuckerberg-11631654215 (letzter Zugriff: 31.01.2026).

34 Costanza-Chock, Sasha (Hrsg.) (2020): Design Justice: Community-Led Practices to Build the Worlds We Need, The MIT Press. DOI: 10.7551/mitpress/12255.001.0001.

35 Kerr, Dara (2025): TikTok to replace trust and safety team in Germany with AI and outsourced labor, The Guardian, https://www.theguardian.com/technology/2025/aug/10/tiktok-trust-safety-team-moderators-ai (letzter Zugriff: 31.01.2026).

36 European Union (2022): Regulation (EU) 2022/2065 (Digital Services Act), Art. 4–6, O.J. (L 277) 1.

37 Böswald, Lena-Maria; Dolezalek, Corinna; Jost, Pablo & Schmid, Ursula Kristin (2025): Between Click and Consequence: An Evaluation of Platform Reporting Procedures under the Digital Services Act, Das Nettz, https://www.das-nettz.de/sites/default/files/2025-10/ENG_Langfassung_DSA.pdf (letzter Zugriff: 31.01.2026).

38 HateAid (2025): Rights Without Reach. The DSA Put to the Test, https://hateaid.org/wp-content/uploads/2025/12/hateaid-dsa-rights-without-reach-2025.pdf (letzter Zugriff: 31.01.2026).

39 Denga, Michael (2025): Enforcement and optimization of social media platforms' responsibility, S. 18–19. https://hateaid.org/wp-content/uploads/2026/03/safety-by-design-expert-opinion-platform-responsibility-michael-denga.pdf (letzter Zugriff: 31/01/2026).

40 Holznagel, Daniel (2025): Shortcomings of the first DSA Audits — and how to do better, DSA Observatory, https://dsa-observatory.eu/2025/06/11/shortcomings-of-the-first-dsa-audits-and-how-to-do-better/ (letzter Zugriff: 31.01.2026).

41 Denga (2025), S. 10–12.

42 Ebd., S. 12.

43 European Commission (2024): TikTok commits to permanently withdraw TikTok Lite Rewards programme from the EU to comply with the Digital Services Act, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_4161 (letzter Zugriff: 31.01.2026).

44 Harfst, Jan-Ole, Mast, Tobias & Schulz, Wolfgang (2025): Independence as a Desideratum: DSA Enforcement by the EU Commission, Verfassungsblog, https://verfassungsblog.de/dsa-enforcement-commission/. DOI: 10.59704/50020240f8894397.

45 Denga (2025), S. 14.

46 Ebd., S. 6–7, 15–17.

47 Ebd., S. 17.

48 European Commission (2024): Commission Staff Working Document Fitness Check on EU consumer law on digital fairness (SWD(2024) 230 final), https://commission.europa.eu/document/download/707d7404-78e5-4aef-acfa-82b4cf639f55_en (letzter Zugriff: 31.01.2026).

49 European Parliament (2023): European Parliament resolution of 12 December 2023 on addictive design of online services and consumer protection in the EU single market (2023/2043(INI)), https://www.europarl.europa.eu/doceo/document/TA-9-2023-0459_EN.html (letzter Zugriff: 31.01.2026).

50 Council of the European Union (2025): The Jutland Declaration: Shaping a Safe Online World for Minors, https://www.digmin.dk/Media/638956829775203140/DIGMIN_The%20Jutland%20Declaration%20Shaping%20a%20Safe%20Online%20World%20for%20Minors%20101025.pdf (letzter Zugriff: 31.01.2026).

51 Denga (2025), S. 20–22.

52 Ebd., S. 9–10.

53 Convocation Research + Design (2026), Safety by Design Taxonomy, https://sbd-taxonomy.vercel.app/ (letzter Zugriff: 31/01/2026).

54 Denga (2025), S. 21.

55 Penfrat, Jan (2022): Everyone is on Mastodon now, but why?, EDRi, https://edri.org/our-work/everyone-is-on-mastodon-now-but-why/ (letzter Zugriff: 31.01.2026).

56 European Parliament (2026): European Parliament resolution of 22 January 2026 on European technological sovereignty and digital infrastructure (2025/2007(INI)), https://www.europarl.europa.eu/doceo/document/TA-10-2026-0022_EN.html (letzter Zugriff: 31.01.2026).

Impressum

HateAid gGmbH
Greifswalder Straße 4
10405 Berlin

Email: kontakt@hateaid.org
hateaid.org

Unternehmenssitz: Berlin
Registergericht: Amtsgericht Charlottenburg
Handelsregister-Nr.: HRB 203883 B
Umsatzsteuer-Identifikationsnr.: DE322705305

EU-Transparenzregister-ID: 802412042190-08

Geschäftsführung und inhaltlich Verantwortliche:
Anna-Lena von Hodenberg, Josephine Ballon