ISD | Institute for Strategic Dialogue

Im Auftrag von:
Bundesnetzagentur

Studie im Auftrag der Bundesnetzagentur

Identifikation, Bewertung und Bekämpfung von systemischen Risiken im Kontext des Digital Services Act

Über das ISD

Die Institute for Strategic Dialogue gGmbH ist ein gemeinnütziger Think & Do Tank. Im Fokus der Arbeit des ISD Germany stehen die Erforschung der Verbreitung von Hassrede, Desinformation und Verschwörungserzählungen sowie die Entwicklung von Gegenstrategien. Dabei greift es auf branchenführende Analysekapazitäten und Erfahrungen im Umgang mit digitalen Diensten zurück.

Über den Bericht

Die Studie wurde von der Bundesnetzagentur beauftragt. Die inhaltliche Verantwortung liegt ausschließlich beim ISD Germany.

Herausgeberische Verantwortung:
Sarah Kennedy, Chief Operating Officer, ISD
Dr. Cornelius Adebahr, Interim Executive Director, ISD Germany

Danksagungen

Die Autor*innen bedanken sich bei folgenden Organisationen für ihre Kommentare, Ideen und Einschätzungen zur Risikobewertung im Online-Umfeld. Die in der Studie geäußerten Ansichten entsprechen nicht unbedingt den Ansichten der genannten Organisationen:

- Gesellschaft für Freiheitsrechte
- Coimisiún na Meán, Irish Digital Services Coordinator (Supervision team)
- Vereinte Nationen
- interface
- Reset Tech
- Universität Freiburg
- AlgorithmWatch
- Ofcom
- Mozilla
- Global Network Initiative
- Access Now
- Rada pre mediálne služby – Slovakian Digital Services Coordinator
- Democracy Reporting International
- AWO Agency
- Alexander von Humboldt Institut für Internet und Gesellschaft

www.isdgermany.org

Inhaltsverzeichnis

Kurzfassung

Die vorliegende Studie widmet sich der Entwicklung eines kohärenten und praktisch umsetzbaren Ansatzes für die Regulierungspraxis zur Identifikation, Bewertung und Minderung systemischer Risiken im Rahmen des Digital Services Act (DSA). Ziel ist es, einen umfassenden Überblick über systemische Risiken von Online-Plattformen zu geben, einschließlich Definitionen, Indikatoren, Messmethoden, Monitoring-Ansätzen und Risikominderungsmaßnahmen. Die Studie bietet konkrete Orientierung für die Akteur*innen der Multi-Stakeholder-Durchsetzungsstruktur, darunter Regulierungsbehörden, Plattformbetreiber*innen sowie Expert*innen aus Wissenschaft und Zivilgesellschaft.

Rechtliche Ausgangslage

Der DSA, der am 16. November 2022 in Kraft trat, verpflichtet sehr große Online-Plattformen (VLOPs) und sehr große Online-Suchmaschinen (VLOSEs) zu einer regelmäßigen Bewertung und Minderung sogenannter systemischer Risiken. Diese Verpflichtungen gelten seit dem 25. August 2023. Obwohl im Risikokategorien im DSA definiert werden, fehlen detaillierte Vorgaben zur Operationalisierung der Bewertungs- und Minderungsmaßnahmen. So bleibt der Begriff »systemisch« unpräzise. Außerdem gibt es keine standardisierten Risikomanagementansätze, die den komplexen und dynamischen Anforderungen des Online-Umfelds gerecht werden. Die Studie trägt dazu bei, diese konzeptionellen und methodischen Lücken zu schließen, indem sie eine solide und nachhaltige Operationalisierungsgrundlage schafft.

Methodisches Vorgehen und Ergebnisse

Im Rahmen dieser Studie entwickelt ISD einen integrierten Rahmen zur Risikobewertung und -minderung, der auf einer systematischen Analyse wissenschaftlicher Literatur, regulatorischer Leitlinien und Experteninterviews basiert. Ein zentraler Beitrag der Studie ist die Einführung einer differenzierten Konzeption systemischer Risiken. Die Bewertung systemischer Risiken erfolgt anhand der Kriterien »Auswirkungen« und »Wahrscheinlichkeit« über drei Risikoebenen hinweg: nutzungsbezogene Risiken, inhaltsbezogene Risiken und verhaltensbezogene Risiken. Die klare Trennung in diese drei Ebenen bietet eine systematische Grundlage, um sowohl Ursachen als auch Dynamiken systemischer Risiken differenziert zu analysieren. Der Bewertungsrahmen als Ganzes ermöglicht eine Analyse der Entstehungsbedingungen und Dynamiken der DSA-Risikokategorien und dient als Grundlage für die Ableitung allgemeiner Indikatoren und Messmethoden. Dies erlaubt eine gezielte Identifikation und Bewertung spezifischer systemischer Risiken sowie die Entwicklung passender Minderungsmaßnahmen für unterschiedliche VLOPs und VLOSEs.

Zusätzlich umfasst die Studie folgende Ergebnisse:

- Allgemeines Risikoprofil mit Leitfragenkatalog zur Identifikation potenzieller Kernrisiken
- Bewertungsindikatoren auf Basis des Bewertungsrahmens sowie eine Risikomatrix
- Beschreibung von Risikominderungsmaßnahmen sowie ein integriertes System zur Auswahl und Bewertung der Maßnahmen
- Konzepte für Risikofrühwarnsysteme zur Adressierung dynamischer Veränderungen
- Fallstudie zum Einsatz von Online-Bots mit einer Analyse ihres Einflusses auf das Auftreten systemischer Risiken auf einer VLOP
- Fallstudie zu Telegram mit der Identifikation potenzieller Kernrisiken und Vorschlägen für Anpassungen des DSA-Rechtsrahmens

Glossar

API: Softwareschnittstelle (engl. Application Programming Interface), die eine Kommunikation zwischen zwei Anwendungen ermöglicht. APIs haben eine riesige Vielzahl an Verwendungszwecken, aber im Zusammenhang mit diesem Bericht ermöglichen sie Forscher*innen den Zugriff auf bestimmte Daten von Online-Plattformen über Datenanfragen. Als zwischengeschaltete Instanz stellen APIs eine zusätzliche Sicherheitsebene bereit, indem sie keinen direkten Zugriff auf Daten zulassen und das Volumen und die Häufigkeit der Anfragen protokollieren, verwalten und überwachen.

Backlinks (Rückverweise): Links auf einer Webseite, die auf eine andere Webseite verweisen.

BERT-Modell: Eine Art Künstliche-Intelligenz-Modell, das hilft, die Bedeutung von Text besser zu verstehen. Es verarbeitet Wörter im Zusammenhang mit den Wörtern davor und danach, um den Sinn eines Satzes genauer zu erfassen (engl. Bidirectional Encoder Representations from Transformers, BERT). Dadurch eignet sich BERT besonders gut für Aufgaben wie das Beantworten von Fragen, das Verstehen von Texten oder die automatische Übersetzung.

CSAM: bezeichnet Material über sexuellen Kindesmissbrauch (engl. Child Sexual Abuse Material).

Dark Patterns: auf Social-Media-Plattformen implementierte Benutzeroberflächen und Nutzererfahrungen, die Nutzer*innen veranlassen sollen, unbeabsichtigte, ungewollte und potenziell schädliche Entscheidungen bezüglich der Verarbeitung ihrer personenbezogenen Daten zu treffen (European Data Protection Board, 2022).

Deplatforming: beschreibt die Entfernung von Konten und Gruppen aus Online-Plattformen.

Deep-Learning: Teilbereich des Maschinellen Lernens, der darauf abzielt, Computer so zu trainieren, dass sie wie ein menschliches Gehirn Daten verarbeiten, Muster erkennen und Entscheidungen treffen können. Dies geschieht durch den Einsatz mehrerer Schichten, sogenannter neuronaler Netze, die aus großen Datenmengen »lernen« können (Europäische Kommission, 2024a).

Desinformation: Verbreitung falscher oder irreführender Inhalte, die der Öffentlichkeit schaden können, in der Absicht, andere zu täuschen oder wirtschaftlich oder politisch daraus Kapital zu schlagen (Europäische Kommission, 2020).

DoS: Cyberangriff, bei dem ein System, Netzwerk oder eine Anwendung gezielt überlastet wird (engl. Denial of Service, DoS), um sie für legitime Nutzer*innen unzugänglich zu machen.

Doomscrolling: Phänomen, bei dem Menschen übermäßig lange und zwanghaft negative Nachrichten oder Inhalte auf Online-Plattformen konsumieren (Sharma et al., 2022).

Doxing: Veröffentlichung von privaten Informationen, wie beispielsweise der Wohnadresse, einer Person online gegen ihren Willen oder ohne deren Zustimmung.

DSC: Koordinierungsstelle für Digitale Dienste (engl. Digital Services Coordinator).

Empfehlungssystem: vollständig oder teilweise automatisiertes System, das von einer Online-Plattform verwendet wird, um auf ihrer Online-Schnittstelle den Nutzer*innen bestimmte Informationen vorzuschlagen oder diese Informationen zu priorisieren, auch infolge einer von dem*der Nutzer*in veranlassten Suche, oder das auf andere Weise die relative Reihenfolge oder Hervorhebung der angezeigten Informationen bestimmt (Art. 3s) DSA).

Erwägungsgrund: Beweggründe und Ziele eines Gesetzes oder einer Verordnung, die zwar nicht rechtlich bindend sind, aber einen hohen Stellenwert bei der rechtlichen Auslegung haben.

Fehlinformationen: falsche oder irreführende Inhalte, die ohne vorsätzliche Schädigungsabsicht weitergegeben werden, deren Auswirkungen jedoch schädlich sein können, z.B. wenn Personen falsche Informationen gutgläubig an Freunde und Familienangehörige weitergeben (Europäische Kommission, 2020).

Filter Bubble: Filterblase, die durch Personalisierung der den Nutzer*innen auf Online-Plattformen angebotenen Informationen durch Algorithmen entsteht (Deutscher Bundestag, 2022).

FIMI: Ausländische Informationsmanipulation und Einflussnahme (engl. Foreign Information Manipulation and Interference) meint ein meist nicht rechtswidriges Verhaltensmuster, das Werte, Verfahren und politische Prozesse bedroht oder das Potenzial hat, diese negativ zu beeinflussen. Eine derartige Aktivität hat einen manipulativen Charakter und wird von staatlichen oder nichtstaatlichen Akteur*innen, einschließlich ihrer Stellvertreter*innen innerhalb und außerhalb ihres eigenen Hoheitsgebiets, absichtlich und in koordinierter Weise durchgeführt (European External Action Service, 2023).

Flash Sales: zeitlich stark begrenzte Verkaufsaktion, bei der Produkte oder Dienstleistungen zu stark reduzierten Preisen angeboten werden.

Freemium Games: Videospiele, die kostenlos heruntergeladen und gespielt werden können, aber kostenpflichtige Zusatzinhalte oder Funktionen anbieten.

Grooming: gezielte Anbahnung von sexuellen Kontakten mit Kindern und Jugendlichen im Internet (Bundeskriminalamt, 2024).

High-demand (Hohe Nachfrage): Situation, in der die Nachfrage nach einem bestimmten Gut oder einer Dienstleistung das verfügbare Angebot übersteigt.

Hostingdienst: Dienst, der im Auftrag von Nutzer*innen bereitgestellte Informationen speichert (Art. 3g) DSA).

Low-supply (Geringes Angebot): Situation, in der eine begrenzte Menge eines bestimmten Guts oder einer Dienstleistung verfügbar ist.

Nudify App: Anwendungen, die mithilfe künstlicher Intelligenz (KI) Bilder so manipulieren, dass bekleidete Personen scheinbar nackt dargestellt werden.

Nutzer*in: jede natürliche oder juristische Person, die einen Vermittlungsdienst in Anspruch nimmt, insbesondere um Informationen zu erlangen oder zugänglich zu machen (Art. 3b) DSA).

Online-Plattform: Hostingdienst, der im Auftrag einer*eines Nutzer*in Informationen speichert und öffentlich verbreitet, sofern es sich bei dieser Tätigkeit nicht nur um eine unbedeutende und reine Nebenfunktion eines anderen Dienstes oder um eine unbedeutende Funktion des Hauptdienstes handelt, die aus objektiven und technischen Gründen nicht ohne diesen anderen Dienst genutzt werden kann, und sofern die Integration der Funktion der Nebenfunktion oder der unbedeutenden Funktion in den anderen Dienst nicht dazu dient, die Anwendbarkeit dieser Verordnung zu umgehen (Art. 3i) DSA).

Online-Suchmaschine: Vermittlungsdienst, der es Nutzer*innen ermöglicht, in Form eines Stichworts, einer Spracheingabe, einer Wortgruppe oder einer anderen Eingabe Anfragen einzugeben, um prinzipiell auf allen Websites oder auf allen Websites in einer bestimmten Sprache eine Suche zu einem beliebigen Thema vorzunehmen und Ergebnisse in einem beliebigen Format, in dem Informationen im Zusammenhang mit dem angeforderten Inhalt zu finden sind, angezeigt zu bekommen (Art. 3j) DSA).

Online-Troll: Person, die absichtlich provokante, beleidigende oder störende Kommentare und Inhalte online veröffentlicht, um andere Nutzer*innen zu verärgern, Diskussionen zu sabotieren oder Streit zu verursachen. Dieses Verhalten fällt oft nicht unter rechtswidrige Handlungen.

Peer-to-peer grooming: Beschreibt den Vorgang, bei dem ein junger Mensch einen anderen jungen Menschen zum Zwecke der sexuellen Ausbeutung manipuliert und kontrolliert (Ashurst & McAlinden, 2015).

Rabbit Holes: Erfahrung oder der Zustand, in dem sich jemand tief in ein Thema, eine Aktivität online hineinzieht, oft ungeplant und mit unerwarteter Dauer oder Tiefe.

Rechtswidrige Inhalte: alle Informationen, die als solche oder durch ihre Bezugnahme auf eine Tätigkeit, einschließlich des Verkaufs von Produkten oder der Erbringung von Dienstleistungen, nicht im Einklang mit dem Unionsrecht oder dem Recht eines Mitgliedstaats stehen, ungeachtet des genauen Gegenstands oder der Art der betreffenden Rechtsvorschriften (Art. 3h) DSA).

Revenge Porn: eine Unterform von bildbasierter sexualisierter Gewalt, nämlich die nicht-einvernehmliche Verbreitung privater, sexueller Bilder durch eine*n Ex-Partner*in (McGlynn et al., 2017).

Self-Engagement Bots: Social-Bots, die die Interaktionen eines Accounts auf Sozialen Medien künstlich erhöhen damit der Anschein erweckt wird, dass bestimmte Beiträge besonders beliebt sind.

Sextortion: Drohung, intime, sexuelle Materialien zu verbreiten, es sei denn, das Opfer erfüllt bestimmte, oft finanzielle, Forderungen (O'Malley & Holt, 2022).

Social-Bots: Computerprogramme, die nach ihrer Aktivierung ohne menschliches Zutun automatisiert auf Online-Plattformen agieren und kaum von menschlichen Nutzer*innen zu unterschieden sind, da sie sich ähnlich wie Menschen verhalten und beispielsweise dazu eingesetzt werden, um Beiträge zu teilen, zu liken oder zu kommentieren (Bundesamt für Sicherheit in der Informationstechnik, 2024).

SSB: Social-Bots, die betrügerische Aktivitäten durchführen (engl. Social Scam Bots, SSB) wie beispielsweise die Verbreitung von Desinformationen.

SVM: Algorithmus aus der KI, der hilft, Daten in Kategorien einzuteilen (Support Vector Machine). SVM versucht, eine klare Grenze (wie eine Linie oder Fläche) zwischen zwei oder mehr Gruppen von Datenpunkten zu finden. Das Ziel ist, diese Grenze so zu wählen, dass die Gruppen möglichst weit voneinander entfernt sind, um Fehler zu vermeiden. Auch bei komplexen Daten, die sich schwer trennen lassen, kann SVM durch spezielle Tricks gute Ergebnisse liefern.

VLOP/VLOSE: sehr große Online-Plattform (engl. Very Large Online Platform) oder sehr große Online-Suchmaschine (engl. Very Large Online Search Engine) mit durchschnittlich mindestens 45 Millionen monatlich aktiven Nutzer*innen in der EU.

1. Einleitung

1.1 Digital Services Act

Digitale Dienste wie Online-Marktplätze, App-Stores, Online-Plattformen der kollaborativen Wirtschaft oder Social-Media-Plattformen haben in den letzten 20 Jahren einen atemberaubenden Aufstieg erlebt. Sie sind mittlerweile fester Bestandteil des Alltags von Verbraucher*innen auf der ganzen Welt geworden, insbesondere in den Industrieländern. Mit dieser Entwicklung gingen aber auch negative Entwicklungen einher. So stellt der Global Digital Compact der Vereinten Nationen fest:

> »[...] digital and emerging technologies can facilitate the manipulation of and interference with information in ways that are harmful to societies and individuals and negatively affect the enjoyment of human rights and fundamental freedoms as well as the attainment of the Sustainable Development Goals.« (Vereinte Nationen, 2024)

Angesichts dieser Herausforderung haben zahlreiche Länder weltweit neue Vorschriften für digitale Dienste eingeführt, um gegen rechtswidrige Inhalte und gesellschaftliche Risiken vorzugehen (Global Online Safety Regulators Network, 2024). In Deutschland verpflichtete das 2017 in Kraft getretene Netzwerkdurchsetzungsgesetz (NetzDG) zum Beispiel Anbieter*innen von sozialen Netzwerken dazu, bestimmte rechtswidrige Inhalte innerhalb eines festen Zeitraums zu entfernen, sobald sie davon Kenntnis erlangen. Andere Staaten wie Frankreich oder Österreich führten ebenso eigene Gesetze ein. Diese lieferten zwar wichtige Beiträge zur Bekämpfung rechtswidriger Inhalte, doch entstand so auch ein Flickenteppich regulatorischer Vorschriften in Europa. Vor diesem Hintergrund verabschiedete die Europäische Union (EU) im Jahre 2022 das Gesetz über digitale Dienste (engl. Digital Services Act, DSA). Die Verordnung, die am 17. Februar 2024 vollständig in Kraft trat, harmonisiert die geltenden Vorschriften für die Anbieter*innen von Vermittlungsdiensten im europäischen Binnenmarkt, um ein sicheres, berechenbares und vertrauenswürdiges Online-Umfeld zu garantieren.

Für die Anbieter*innen von sehr großen Online-Plattformen (engl. Very Large Online Platforms, VLOPs) und sehr großen Online-Suchmaschinen (engl. Very Large Online Search Engines, VLOSEs) mit durchschnittlich mindestens 45 Millionen aktiven monatlichen Nutzer*innen in der EU (Art. 33 Abs. 1 DSA) sieht die Verordnung dabei zusätzliche Verpflichtungen in Bezug auf den Umgang mit sogenannten »Systemische Risiken« vor. Konkret müssen sie mindestens einmal jährlich Risikobewertungen (Art. 34 Abs. 1 DSA) und Risikominderungsmaßnahmen (Art. 35 Abs. 1 DSA) umsetzen, die in einer Multi-Stakeholder-Durchsetzungsstruktur unter Einbeziehung von betroffenen Gruppen (z.B. Frauen-, Kinder- und Migrantenorganisationen, Verbraucherschutzzentralen) sowie unabhängigen Sachverständigen und zivilgesellschaftlichen Organisationen (z.B. Menschenrechtsorganisationen) umgesetzt werden sollten (Erwägungsgrund 90 DSA). Dafür erhalten nicht nur die Europäische Kommission und die jeweiligen nationalen Koordinierungsstellen für digitale Dienste (engl. Digital Services Coordinators, DSCs) weitreichenden Datenzugang (Art. 40 Abs. 1 DSA), sondern unter bestimmten Voraussetzungen auch Forscher*innen aus Wissenschaft und Zivilgesellschaft (Art. 40 Abs. 4 DSA, Abs. 12 DSA).

Die Durchführung von Risikobewertungen ist keine neue Erfindung, sondern in Bereichen wie dem Bankensektor seit Jahren fest verankert. Trotzdem lassen sich die bisherigen Risikomanagementansätze aus anderen Industriesektoren nicht ohne Weiteres auf das komplexe und dynamische Online-Umfeld übertragen. Gleichzeitig gibt der DSA keine Klarheit in Bezug auf die Definition bestimmter Rechtsbegriffe wie »Systemische Risiken« oder in Bezug auf Methoden, Maßstäbe und Standards, die für die praktische Durchführung von Risikobewertungen herangezogen werden könnten.

Folglich bedarf es weiterer Konkretisierung, wie insbesondere das Konzept der »systemischen Risiken« in der Praxis umgesetzt und Risikobewertungen durchgeführt werden sollten (Ebert et al., 2023). Es ist klar, dass es keine Einheitslösung für alle designierten VLOPs und VLOSEs geben kann. Dennoch gibt es einen Bedarf an einer anleitenden Grundstruktur, die den beteiligten Stakeholder*innen helfen kann, in einem »Lernprozess« Risiken zu identifizieren, zu bewerten und zu bekämpfen (Marsh, 2024).

1.2 Ziel und Struktur der Studie

Die vorliegende Studie begegnet der konzeptionellen und praktischen Leerstelle im DSA bei der Risikobewertung, indem sie Methoden und Maßnahmen zur Identifikation, Bewertung und Bekämpfung von systemischen Risiken entwickelt und zusammenstellt. Sie leistet damit eine konkrete Hilfestellung für die beteiligten Akteur*innen der Multi-Stakeholder-Durchsetzungsstruktur – von Regulierungsbehörden über Anbieter*innen von VLOPs und VLOSEs bis zu Forscher*innen aus Wissenschaft und Zivilgesellschaft. Damit trägt die Studie zur nachhaltigen und effektiven DSA-Durchsetzung im Bereich der Risikobewertung und -minderung bei.

Basierend auf einer systematischen Analyse der bestehenden wissenschaftlichen Erkenntnisse, Leitlinien und Berichte sowie Interviews mit Expert*innen aus Regulierungsbehörden, Wissenschaft und Zivilgesellschaft entwickelt die Studie einen kohärenten und praktisch umsetzbaren Ansatz für die Regulierungspraxis. Damit bietet sie eine solide und nachhaltige Grundlage für zukünftige Entscheidungen und Maßnahmen zur DSA-Durchsetzung und trägt zur Harmonisierung entsprechender Konzepte in der EU bei. Zudem gibt sie einen umfangreichen Überblick über systemische Risiken von Online-Plattformen einschließlich Definitionen, Indikatoren, Messmethoden und Monitoring-Ansätzen. Dies hilft, die Fähigkeit von Regulierungsbehörden und anderen zu verbessern, systemische Risiken frühzeitig zu erkennen und zu bewerten. Darüber hinaus erfolgt eine vergleichende Darstellung und Bewertung von bestehenden und zukünftigen Risikominderungsstrategien auf Online-Plattformen.

Die Studie ist in acht Kapitel strukturiert. Nach dieser Einleitung (Kapitel 1) bietet Kapitel 2 einen grundlegenden Überblick über das Konzept »Systemische Risiken«: Es erläutert und interpretiert die Risikokategorien im DSA, ordnet den Begriff »systemisch« ein und entwickelt eine Arbeitsdefinition. Kapitel 3 leitet einen allgemeinen Bewertungsrahmen her, der auf bisherigen Risikomanagementansätzen aufbaut. Dazu werden ein Bewertungsprozess, ein Risikoprofil zur Risikoidentifikation bzw. -ermittlung sowie spezifische Indikatoren definiert und umfassend beschrieben.

In Kapitel 4 werden potenzielle und praktische Bewertungsindikatoren für jede Art von Risiko zur Bewertung von Auswirkungen und Wahrscheinlichkeit sowie eine Risikomatrix als Priorisierungsinstrument vorgestellt. Kapitel 5 befasst sich mit den Risikominderungsmaßnahmen, die für systemische Risiken kartiert wurden. Dabei stellt es eine Kategorisierung dieser Maßnahmen zusammen und schlägt einen konzeptionellen Rahmen zur Unterstützung ihrer Umsetzung und zur Bewertung der Wirksamkeit vor. Grundzüge und Anforderungen eines Frühwarnsystems für systemische Risiken werden in Kapitel 6 eingeführt. Dazu werden Anforderungen und mögliche Datenpunkte beschrieben. Zudem werden Möglichkeiten der automatisierten Datenanalyse erörtert.

Kapitel 7 stellt eine Fallstudie zu Nicht-VLOPs am Beispiel Telegram vor. Diese arbeitet anhand einer vereinfachten Risikoermittlung potenzielle spezifische systemische Risiken heraus, die von Telegram und seiner Nutzung ausgehen, und regt Ergänzungen und Anpassungen des DSA-Rechtsrahmens an. Kapitel 8 fasst die Studienergebnisse zusammen, zeigt Limitationen des in dieser Studie entwickelten Bewertungsrahmens und zugehöriger Indikatoren und Risikominderungsmaßnahmen auf. Außerdem liefert es einen Ausblick auf künftige Entwicklungen bei der Implementierung und Durchsetzung der Sorgfaltspflichten zum Umgang mit systemischen Risiken.

2. Systemische Risiken

2.1 Risikokategorien im DSA

Art. 34 Abs. 1 des DSA listet systemische Risiken auf, die sich für die EU aus der Konzeption, dem Betrieb oder der Nutzung von Diensten und ihren verbundenen Systemen, einschließlich algorithmischer Systeme, ergeben. Hintergrund ist, dass die Art und Weise der Nutzung von VLOPs und VLOSEs großen Einfluss auf die Online-Sicherheit, die öffentliche Meinungsbildung, den öffentlichen Diskurs sowie den Online-Handel haben können (Erwägungsgrund 79 DSA). Zwar erläutert der DSA den Begriff »systemisch« nicht, doch werden in der Verordnung vier grundsätzliche Risikokategorien aufgeführt. Zur Auslegung dieser vier Kategorien gilt es, die Erwägungsgründe des DSA und zusätzliche Leitlinien zu berücksichtigen, insbesondere die Leitlinien der Europäischen Kommission zur Minderung systemischer Risiken in Wahlprozessen (Europäische Kommission, 2024c).

Die erste Kategorie betrifft solche Risiken, die sich aus der »Verbreitung rechtswidriger Inhalte über ihre Dienste« (Art. 34 Abs. 1a) DSA) ergeben können. Hierzu gehören z.B. die Darstellungen von sexuellem Missbrauch von Kindern, rechtswidrige Hassreden oder andere Arten des Missbrauchs der Dienste für Straftaten. Rechtswidrige Aktivitäten wie der Verkauf von Waren oder Dienstleistungen, die nach Unionsrecht oder nationalem Recht verboten sind, sowie gefährliche oder gefälschte Waren oder rechtswidrig gehandelte Tiere fallen ebenfalls in diese Kategorie.

Die zweite Risikokategorie umfasst die möglichen negativen Auswirkungen auf die Grundrechte. Gemäß Art. 34 Abs. 1b), »etwaige tatsächliche oder vorhersehbare nachteilige Auswirkungen auf die Ausübung der Grundrechte, insbesondere des in Artikel 1 der Charta verankerten Grundrechts auf Achtung der Menschenwürde, des in Artikel 7 der Charta verankerten Grundrechts auf Achtung des Privat- und Familienlebens, des in Artikel 8 der Charta verankerten Grundrechts auf Schutz personenbezogener Daten, des in Artikel 11 der Charta verankerten Grundrechts auf die Meinungs- und Informationsfreiheit, einschließlich Medienfreiheit und -pluralismus auf das in Artikel 21 der Charta verankerte Grundrecht auf Nichtdiskriminierung, die in Artikel 24 der Charta verankerten Rechte des Kindes und den in Artikel 38 der Charta verankerten umfangreichen Verbraucherschutz«.

In Erwägungsgrund 81 DSA heißt es, dass diese Risiken häufig aus der Gestaltung algorithmischer Systeme resultieren, die Diskriminierung oder Manipulation fördern könnten, aus der Nutzung des Dienstes für die Übermittlung missbräuchlicher Nachrichten oder aus anderen Methoden zur Verhinderung der freien Meinungsäußerung, oder aus der Behinderung des Wettbewerbs. In dem Erwägungsgrund wird auch betont, dass Online-Plattformen Minderjährige schützen müssen, indem sie ihre Dienste so gestalten, dass junge Nutzer*innen sicher mit diesen umgehen können. Online-Plattformen sollten es Minderjährigen leicht machen, die Gestaltung und Funktionsweise des Dienstes zu verstehen. Sie sollten darüber aufklären, wie ihr Dienst es vermeiden kann, minderjährige Nutzer*innen Inhalten auszusetzen, die ihrer Gesundheit, ihrer körperlichen, geistigen oder moralischen Entwicklung schaden könnten.

Die dritte Kategorie umfasst »alle tatsächlichen oder absehbaren nachteiligen Auswirkungen auf die gesellschaftliche Debatte und auf Wahlprozesse und die öffentliche Sicherheit« (Art. 34 Abs. 1c) DSA). Den Leitlinien zufolge sind Beispiele für Risiken dieser Kategorie die Bedrohungen im Zusammenhang mit »Ausländischer Informationsmanipulation und Einflussnahme« (engl. Foreign Information Manipulation and Interference, FIMI); die Verbreitung gewalttätiger extremistischer Inhalte mit der Absicht, Menschen zu radikalisieren; die Verbreitung von Inhalten, die durch neue Technologien wie generative künstliche Intelligenz durch Empfehlungssysteme erzeugt werden; sowie Online-Belästigung von politischen Kandidat*innen oder Amtsinhaber*innen, Journalist*innen, Wahlhelfer*innen oder anderen am Wahlprozess Beteiligten. Durch ihr Design und ihre Funktionalität können Online-Plattformen dazu beitragen, die freie Meinungsbildung zu gefährden, insbesondere wenn algorithmische Systeme Inhalte begünstigen, die die gesellschaftliche Debatte verzerren. Forschungsberichte und wissenschaftliche Arbeiten weisen hier auf weitere Risiken hin, insbesondere durch koordiniertes Verhalten und gezielte Desinformationskampagnen (Calabrese & Reich, 2024).

Die vierte Risikokategorie betrifft »alle tatsächlichen oder absehbaren nachteiligen Auswirkungen in Bezug auf geschlechtsspezifische Gewalt, den Schutz der öffentlichen Gesundheit und von Minderjährigen sowie schwerwiegende nachteilige Folgen für das körperliche

und geistige Wohlbefinden einer Person« (Art. 34 Abs. 1d) DSA). Diese Kategorie bezieht sich auf die Gestaltung von Online-Plattformen, deren Inhalte potenziell schwerwiegende Auswirkungen auf das körperliche und geistige Wohlbefinden der Nutzer*innen haben können, z.B. Sucht-Verhaltensmuster verstärken. Weitere Risiken sich aus koordinierten Desinformationskampagnen in Bezug zu öffentlichen Gesundheitsthemen oder aus der Gestaltung von den Benutzeroberflächen der Online-Plattformen, die eine Verhaltenssucht der Nutzer*innen fördern können.

Nachdem die vier Kategorien systemischer Risiken im Rahmen des DSA vorgestellt wurden, stellt sich die Frage, wie diese Risiken in der Praxis definiert, erkannt und adressiert werden können. Dabei ergeben sich jedoch eine Reihe von Herausforderungen. Zwar geben der DSA und seine Erwägungsgründe den Regulierungsbehörden, VLOPs und VLOSEs und Forschenden eine erste Anleitung, welche Arten von Risiken als systemisch anzusehen sind. Jedoch schafft die fehlende Klarheit in der Definition von »systemisch« Herausforderungen für die Art und Weise, wie diese Akteur*innen sie identifizieren und abschwächen sollten (Hendrix & Jahangir, 2024). Dieses Szenario wirft praktische Umsetzungsprobleme auf, da sich bestimmte Risiken ändern können, was die Ermittlung und Messung dieser Risiken erschwerten.

Darüber hinaus verlangen dynamische und sich schnell verändernde Online-Risiken neue Entwicklungen in die Risikobewertungsrahmen aufzunehmen, um eine proaktive Risikoerkennung so wie reaktive Maßnahmen auf neu auftretende Bedrohungen einzuleiten (Djeffal, 2022). Zudem überschneiden sich viele der in den Erwägungsgründen und Leitlinien genannten Risiken gegenseitig (Allen, 2022). Beispielsweise kann geschlechtsspezifische Gewalt im Online-Umfeld (engl. Online Gender Based Violence, OGBV) ein Spektrum umfassen, das sowohl rechtswidrige Hassreden als auch legale, aber schädliche Inhalte umfasst.

Doch selbst wenn Online-Plattformen oder Regulierungsbehörden in der Lage sind, diese Risiken zu erkennen, bleibt die Herausforderung bestehen, sie zu messen. Aus diesem Grund sollten spezifische Indikatoren und Schwellenwerte, die die Entscheidung unterstützen, ob ein Risiko systemisch geworden ist, ständig überprüft und überwacht werden, um sich an das schnelllebige Online-Umfeld anzupassen.

2.2 Was bedeutet »systemisch«?

Obgleich Art. 34 Abs. 1 unterschiedliche Kategorien systemischer Risiken beschreibt, fehlt es an einer eindeutigen Definition im DSA, was unter »systemischen« Risiken zu verstehen ist. Zwar könnte argumentiert werden, dass die Risikokategorien (s. 2.1) selbst bereits den Begriff »Systemische Risiken« bestimmen. Außerdem führt Art. 34 Abs. 1 DSA die Kriterien »Schwere« und »Wahrscheinlichkeit« von systemischen Risiken als Bewertungsmaßstab ein. Dennoch bleibt weiterhin offen, welche konkreten Elemente den »systemischen« Charakter von Risiken ausmachen. Folglich eröffnen sich Interpretationsspielräume für die Auslegung von systemischen Risiken.

Im Entstehungsprozess des DSA spielte insbesondere die Verwendung des Begriffs im Bankensektor eine relevante Rolle. Hier ist das Verständnis von systemischen Risiken eng an den Begriff »Systemisches Ereignis« geknüpft. Darunter wird die Veröffentlichung einer negativen Nachricht verstanden, die eine*n der Akteur*innen des Systems betrifft und sich negativ auf mindestens eine*n andere*n Akteur*in desselben Systems auswirkt. Dabei basiert ein solches Ereignis immer auf einem oder mehreren Schocks, die ein oder mehrere Banken betreffen und entweder interner oder externer Natur sind. Im Hinblick auf das Online-Umfeld erweist sich das Konzept des Schocks trotz Übertragungsmöglichkeiten jedoch als limitiert, da die Risiken hier insbesondere von den Funktionalitäten der Dienste und ihrer Nutzung beeinflusst werden sowie von einer Akkumulation einzelner Schäden herrühren (Broughton Micova & Calef, 2023).

Vor diesem Hintergrund bleibt es den Anbieter*innen von VLOPs und VLOSEs, den Regulierungsbehörden, der forschenden Zivilgesellschaft oder der Wissenschaft bislang weitgehend selbst überlassen, wie sie den Begriff »Systemische Risiken« bei der Durchführung von Risikobewertungen interpretieren. Daraus resultieren unterschiedliche, teilweise kongruente Sichtweisen. Diese unterteilen sich gemeinhin in zwei Strömungen von Auslegungen: Risiken, die im Sinne des Verständnisses im Bankensektor weitreichende Auswirkungen auf »Systeme« haben und Risiken, die durch die Plattformsysteme verursacht oder verschärft werden (GNI & DTSP, 2023).

Die erste Auslegungsrichtung fokussiert sich auf Online-Schäden, die über einen digitalen Dienst hinausgehen, indem sie sich über mehrere Dienste erstrecken; viele Nutzer*innen eines Dienstes betreffen; sich gleichzeitig auf die Ausübung mehrerer Schutzgüter auswirken oder sich auf der Makroebene auf die Gesellschaft auswirken. Bei letzterer Auslegung geht es gerade darum, dass die Dienste und ihre Nutzung ursächlich dafür sind, dass einzelne Gefährdungen regelmäßig eine große Bandbreite entwickeln oder dass sich aus einer Vielzahl individueller Schäden ein Risiko ergibt, das über die Summe der Einzelfälle hinaus eine systemische Bedeutung erlangt (Müller-Terpitz et al., 2024). Dies steht im Einklang mit den Leitprinzipien der Vereinten Nationen für Wirtschaft und Menschenrechte (UN Guiding Principles on Business and Human Rights, UNGP) (DGCN, 2014), die auch in Erwägungsgrund 47 DSA explizit genannt werden. So fordert Prinzip 13 (b) der Leitprinzipien ein, »dass Wirtschaftsunternehmen bemüht sind, negative Auswirkungen auf die Menschenrechte zu verhüten oder zu mindern, die auf Grund einer Geschäftsbeziehung mit ihrer Geschäftstätigkeit, ihren Produkten oder Dienstleistungen unmittelbar verbunden sind, selbst wenn sie nicht zu diesen Auswirkungen beitragen« (DGCN, 2014). In anderen Worten: Die Anbieter*innen tragen auch dann eine Verantwortung, wenn sich ihr Dienst indirekt auf bestimmte Schutzgüter auswirkt.

Die zweite Auslegungsrichtung setzt die erste Auslegung voraus. Gleichzeitig rückt sie aber den Einflussgrad der Funktionalitäten digitaler Dienste und vorsätzliche Manipulation in den Vordergrund. Dies kann folgende Aspekte umschließen: Die Art und Weise, wie Nutzer*innen durch Plattformsysteme wie beispielsweise Empfehlungssysteme schädlichen Inhalten ausgesetzt werden; die potenzielle Schadensamplifikation durch die Interkation mehrerer unterschiedlicher, aber miteinander verbundener Funktionalitäten (umfasst auch die Funktionalitäten kleinerer Dienste); oder vorsätzliche Manipulationen, die Moderationssysteme oder andere Sicherheitsmaßnahmen der Dienste gezielt unterlaufen. Damit orientiert sich diese Auslegungsrichtung eng an Art. 34 Abs. 2 DSA. Dieser gibt vor, dass bei der Durchführung von Risikobewertungen insbesondere berücksichtigt werden soll, ob und wie bestimmte Faktoren systemische Risiken beeinflussen. Hierzu gehören Faktoren wie die Gestaltung von Empfehlungssystemen und anderer relevanter algorithmischer Systeme, Systeme zur Moderation von Inhalten oder die anwendbaren allgemeinen Geschäftsbedingungen und ihre Durchsetzung. Außerdem soll bei der Bewertung analysiert werden, ob und wie die Risiken beeinflusst werden, etwa durch vorsätzliche Manipulation eines Dienstes, oder die Verstärkung – das heißt, rasche und weite Verbreitung – von rechtswidrigen Inhalten und von Informationen, die mit den Nutzungsbedingungen unvereinbar sind.

Vor dem Hintergrund beider Auslegungsrichtungen stellt diese Studie eine integrierte, verallgemeinerte und praxisorientierte Arbeitsdefinition von systemischen Risiken im DSA-Kontext auf.

Arbeitsdefinition »Systemische Risiken« im DSA-Kontext

Systemische Risiken bezeichnen potenzielle Online-Schäden (z.B. virale Desinformation), die wesentlich durch Funktionalitäten von Diensten, deren Nutzung oder vorsätzliche Manipulation verursacht werden. Ihre Wahrscheinlichkeit ist hoch und ihre gesellschaftlichen Auswirkungen gehen weit über individuelle Schäden hinaus.

Zentrales Element ist der Gattungsbegriff »Online-Schäden« (engl. Online Harms). Dieser erlangte in den letzten Jahren auf nationalstaatlicher und internationaler Ebene zunehmend an Bedeutung und erlaubt eine praxisorientierte Interpretation systemischer Risiken. In Anlehnung an eine Typologie der Global Coalition for Digital Safety sollen Online-Schäden in dieser Studie als potenzielle Schäden im Online-Umfeld auf den Ebenen »Inhalt«, »Verhalten« und »Nutzung« (s. 3.1) verstanden werden (World Economic Forum, 2023b). Diese stellen wiederum ein systemisches Risiko dar, wenn a) ihre Wahrscheinlichkeit hoch ist und b) ihre gesellschaftlichen Auswirkungen weit über individuelle Schäden hinausgehen (z.B. virale Desinformation). Außerdem ist die Schadensursache wesentlich mit den Funktionalitäten oder deren Nutzung, oder vorsätzlicher Manipulation eines Dienstes verbunden.

2.3 Bisherige Risikomanagementansätze

Neben der Frage nach der Definition des Begriffs »Systemische Risiken« ist die Entwicklung eines allgemein

übertragbaren und praktisch anwendbaren Bewertungs- und Minderungsprozesses zentral für die Implementierung der zusätzlichen Verpflichtungen des DSA zum Umgang mit systemischen Risiken. Der DSA selbst bestimmt zwar, dass »Risikobewertungen mindestens einmal jährlich, in jedem Fall aber vor der Einführung von Funktionen« durchgeführt werden müssen (Art. 34 Abs. 1 DSA). Gleichzeitig werden die konkreten Indikatoren, Messmethoden, Prozessschritte und die Verankerung im Unternehmen weitgehend offengelassen. Vor diesem Hintergrund können bisherige Bewertungsprozesse aus unterschiedlichen Sektoren und Kontexten eine wichtige Hilfestellung für die Implementierung geben.

Schon seit langem beschäftigen sich Unternehmen mit der Bewertung von Risiken. Die ISO-Norm DIN ISO 31000 legt dafür sektoren- und branchenunabhängige allgemeine Leitlinien für das Behandeln von Risiken fest, denen Organisationen ausgesetzt sind (Din Media, 2024). Konkret wird dabei ein Top-Down-Ansatz verfolgt, der auf den Säulen »Grundsätze«, »Rahmenwerk« und »Prozess« aufbaut (Brühwiler & Romeike, 2010). Hierbei geht es darum, Werte zu schaffen und zu erhalten, indem Risikomanagement auf der Leitungsebene einer Organisation verankert wird. Beim Bottom-Up-Ansatz, wie er eher beim Risikomanagement als Resultat von regulatorischen Vorgaben wie im Bankensektor vorzufinden ist, geht es hingegen um die prozessspezifischen Einzelteile einer Organisation oder eines bestimmen Systems (Brühwiler & Romeike, 2010, S. 88). Zentral ist hier die Bestimmung von risikotragenden Eigenschaften und zugehörigen Risikominderungsmaßnahmen. Während das Risikomanagement in den meisten Unternehmen auf das Begrenzen möglicher Haftungsansprüche sowie notwendige Anpassungen an Umfeldveränderungen zielt, stehen indirekte Auswirkungen des Geschäftsmodells auf das öffentliche Interesse nicht explizit im Vordergrund.

Diese Leerstelle wird durch die bereits unter 2.2 erwähnten UNGPs oder die OECD-Leitsätze für multinationale Unternehmen zu verantwortungsvollem unternehmerischem Handeln geschlossen. Diese verlangen von Unternehmen die Durchführung risikobasierter Sorgfaltsprüfungen, damit deren Unternehmensaktivitäten weder direkt noch indirekt die Menschenrechte verletzen. Prinzip 15 UNGP legt dabei fest, dass Wirtschaftsunternehmen über ein Verfahren zur Gewährleistung ihrer menschenrechtlichen Sorgfaltspflicht verfügen sollten. Dieses soll die Auswirkungen der Unternehmenstätigkeit auf die Menschenrechte ermitteln und etwaige Schäden mindern sowie Rechenschaft ergriffene Maßnahmen ablegen. Darüber hinaus wird nach Prinzip 15 UNGP ein Verfahren zur Ermöglichung der Wiedergutmachung menschenrechtlicher Auswirkungen eingefordert. Der DSA spiegelt einige der Kernelemente der UNGPs wider, wenn es um die Bewertung von Risiken, die Transparenz und die Einbindung von Stakeholder*innen in die Unternehmenspraktiken geht. Trotzdem lassen sich die die UNGPs nicht ohne Weiteres auf die Herausforderungen des komplexen Online-Umfeldes übertragen.

In der Folge haben sich neben dem Projekt »B-Tech« des Büros des Hochkommissars für Menschenrechte (OHCHR) auch Organisationen der forschenden Zivilgesellschaft und Wissenschaft darauf fokussiert, die die Spezifika digitaler Geschäftsmodelle oder konkret die DSA-Vorgaben mit menschenrechtlichen Sorgfaltsprüfungen zu verbinden. So hat The Danish Institute for Human Rights bereits einen Leitfaden für eine menschenrechtsbasierte Folgeabschätzung digitaler Aktivitäten entwickelt, der eine praktische Anleitung für den gesamten Bewertungsprozess beinhaltet (Bloch Veiberg, 2023). Access Now und das European Center for Not-for-Profit Law haben wiederum Mindestanforderungen an eine grundrechtsbasierte Folgeabschätzung (engl. Fundamental Rights Impact Assessment, FRIA) im DSA-Kontext aufgestellt, einschließlich spezifischer Risiken und Prozessschritten und Vergleichsmaßstäben für die Evaluation von Risikobewertungen (AccessNow & ECNL, 2023). Andere Forschungsarbeiten legen wiederum einen Schwerpunkt auf Messmethoden (s. Anhang 1 für eine weiterführende Literatur-Datenbank). Diese knüpfen etwa an ereignisbasierte Ansätze an, bei denen die Messung der Wahrscheinlichkeit von Risikoereignissen und ihren Auswirkungen im Mittelpunkt steht (Loi, 2023). Oder sie verfolgen Szenario-basierte Ansätze, welche abstrakte Risiken in messbare Hypothesen übersetzen, um der Pluralität und Dynamik der Dienste zu entsprechen (Meßmer & Degeling, 2023). Diese Ansätze lassen sich in adaptierte menschenrechtliche Sorgfaltsprüfungen einbetten.

Außerhalb des DSA-Kontextes existieren bereits konkretere Risikomanagementansätze zur Sicherstellung eines sicheren Online-Umfeldes. Diese weisen aber meistens keinen Bezug zu menschenrechtsbasierten Folgeab-

schätzungen auf. Die britische Regulierungsbehörde Ofcom hat zum Beispiel einen Leitfadenentwurf zur Bewertung rechtswidriger Online-Schäden veröffentlicht (Ofcom, 2023). Hierfür wurden generische Risikoprofile für User-to-User-Dienste und Suchdienste erarbeitet, die spezifische Risikofaktoren und damit verbundene Kernrisiken umfassen. Dieser Ansatz greift die Bewertungsmaßstäbe »Wahrscheinlichkeit« und »Auswirkung« auf, hat aber einen starken Fokus auf rechtswidrige Inhalte und nicht auf Auswirkungen auf die Menschenrechte oder die gesellschaftliche Debatte. Einen weiter reichenden Vorschlag macht wiederum die Global Coalition for Digital Safety. Sie empfiehlt basierend auf anderen Risikomanagementansätzen einen umfassenden Risikobewertungsrahmen, einschließlich der Schadensbehebung (World Economic Forum, 2023a). Darüber hinaus hat die Allianz konkrete Messgrößen in den Kategorien »Wahrscheinlichkeit«, »Auswirkung« und »Prozess« zusammengetragen (World Economic Forum, 2024). Die bisherigen Risikomanagementansätze liefern hilfreiche Elemente für die Operationalisierung von Risikobewertungen im DSA-Kontext. Allerdings greifen sie oft zu kurz, da sie primär Symptome adressieren, ohne die zugrunde liegenden Ursachen und Dynamiken systemischer Risiken ausreichend zu erfassen. Angesichts der Komplexität und des ständigen Wandels der digitalen Landschaft bedarf es eines umfassenderen Ansatzes. Im nächsten Kapitel wird daher ein konzeptioneller Rahmen entwickelt, der über die symptomorientierte Betrachtung hinausgeht. Dieser Rahmen fokussiert auf die Wechselwirkungen zwischen Nutzer*innen und digitaler Infrastruktur, um die Ursachen systemischer Risiken präzise zu analysieren und geeignete Indikatoren sowie Minderungsstrategien abzuleiten. Er bietet eine fundierte Grundlage, die Online-Plattformen, Forscher*innen und Regulierungsbehörden gleichermaßen als Instrument zur Priorisierung und Bewertung systemischer Risiken dient.

3. Einführung eines Bewertungsrahmens

3.1 Herleitung der ISD-Risikoebenen (Nutzung, Inhalt, Verhalten)

Die Risikobewertung durch Online-Plattformen ist entscheidend für die wirksame Ermittlung und Minderung systemischer Risiken. VLOPs und VLOSEs sind daher verpflichtet, systematische Prüfungen durchzuführen, um potenzielle Risiken zu identifizieren, die sich aus der spezifischen Gestaltung ihrer Dienste sowie einem möglichen Missbrauch durch Nutzer*innen ergeben. Dabei ist zu berücksichtigen, dass die Verantwortlichkeiten digitaler Online-Plattformen eng mit den von ihnen bereitgestellten Möglichkeiten und Diensten sowie ihrer digitalen Infrastruktur verknüpft sind. Online-Plattformen können nicht für jeden Missbrauch oder unbeabsichtigte Folgen von Online-Interaktionen verantwortlich gemacht werden, sondern gezielt für Risiken, die durch ihre Gestaltung ermöglicht oder verstärkt werden. Insbesondere tragen sie Verantwortung für Risiken, die sich direkt aus den von ihnen geschaffenen und geförderten Merkmalen, Funktionalitäten und Möglichkeiten ergeben.

Um die Natur systemischer Risiken umfassend zu verstehen, wird im Rahmen dieser Studie ein konzeptioneller Rahmen entwickelt, der die Dynamiken zwischen Nutzer*innen und digitaler Infrastruktur analysiert, um jene Schnittstellen zu identifizieren, an denen Risiken entstehen. Dieser Rahmen soll nicht nur digitalen Online-Plattformen als Instrument dienen, ihre Verantwortung besser wahrzunehmen, sondern auch Forschenden eine Grundlage bieten, systemische Risiken detaillierter zu untersuchen. Gleichzeitig kann er die Arbeit der Regulierungsbehörden unterstützen, ob bei der Priorisierung von Risiken oder bei der Beurteilung, welche VLOPs und VLOSEs wirksame Minderungsmaßnahmen implementiert haben.

Der neue konzeptionelle Rahmen basiert auf einer neuen Kategorisierung systemischer Risiken und einer Neugruppierung der vier in Art. 34 DSA definierten systemischen Risikokategorien. Diese neue Kategorisierung verfolgt drei zentrale Ziele:

- Erstens, die zugrunde liegenden Ursachen der Risiken besser zu verstehen;
- zweitens, gezielte Indikatoren für den Risikobewertungsprozess zu entwickeln (s. 3.2 ff.);
- und drittens, die Verknüpfung spezifischer Maßnahmen zur Risikominderung mit den identifizierten Risiken zu erleichtern (s. Kapitel 5).

Die unter 2.1 beschriebenen DSA-Risikokategorien, wie etwa die Verbreitung rechtswidriger Inhalte (Art. 34 Abs. 1a DSA) oder die Beeinträchtigung der gesellschaftlichen Debatte (Art. 34 Abs. 1c DSA), sind rechtlich relevant und bilden die Grundlage für die Definition und Regulierung systemischer Risiken. Aus analytischer Perspektive wird jedoch deutlich, dass diese Kategorien vor allem Symptome adressieren und die zugrunde liegenden Ursachen nicht hinreichend differenziert betrachten. Eine neue Konzeptualisierung erweist sich daher als notwendig, um die Komplexität und Ursprünge dieser Risiken besser zu verstehen. Diese Neuausrichtung erlaubt die gezielte Entwicklung von Indikatoren und Maßnahmen, die nicht nur Symptome, sondern auch strukturelle und dynamische Ursachen adressieren. Damit wird eine präzisere Risikobewertung und wirksamere Risikominderungsstrategien ermöglicht.

Der neue Bewertungsrahmen bleibt bewusst flexibel, indem er keine festen Grenzwerte für die entwickelten Indikatoren definiert. Dieser Ansatz wird unter 8.1 ausführlich erläutert. Die Entscheidung, auf Grenzwerte zu verzichten, basiert auf zwei zentralen Überlegungen: Erstens erschweren sowohl die Herausforderungen bei der Messung von Indikatoren als auch der oft fehlende wissenschaftliche Konsens über die exakte Definition von Grenzwerten eine präzise und allgemein gültige Festlegung. Solche Grenzwerte setzen eindeutig messbare Indikatoren und standardisierte Messmethoden voraus, die in der Praxis häufig nicht verfügbar sind. Zweitens unterliegt das Online-Umfeld einem ständigen Wandel durch technologische Innovationen und den globalen Wettbewerb. Diese Dynamik würde feste Grenzwerte schnell obsolet machen und ihre Relevanz für die Risikobewertung stark einschränken.

Darüber hinaus besteht bei einer übermäßigen Quantifizierung der Indikatoren das Risiko, dass qualitative Aspekte, die für ein umfassendes und wirksames Risikomanagement essenziell sind, vernachlässigt werden. Ein solcher Ansatz könnte dazu führen, dass Bewertungen sich vornehmlich auf leicht messbare oder vorteilhafte Grenzwerte konzentrieren, während komplexere und schwerer messbare Risiken unzureichend berücksichtigt werden. Die bewusste Entscheidung, auf feste Grenzwerte zu verzichten, gewährleistet daher eine größere Flexibilität und Anpassungsfähigkeit des Rahmens, um den Anforderungen eines dynamischen und vielfältigen digitalen Umfelds gerecht zu werden.

Diese Überlegungen unterstreichen, dass ein konzeptioneller Rahmen klare Indikatoren liefern sollte, ohne jedoch durch starre Grenzwerte eingeschränkt zu werden. Der vom ISD eigens entwickelte konzeptionelle Rahmen mit seinen drei Hauptkategorien – nutzungsbezogene Risiken, inhaltsbezogene Risiken und verhaltensbezogene Risiken ermöglicht die Herleitung von spezifischen Indikatoren. Diese Systematik (s. Anhang 2 für eine umfassende Übersicht) erlaubt eine differenzierte Betrachtung der Ursachen und Dynamiken systemischer Risiken und bildet so die Grundlage für präzisere Bewertungs- und Minderungsstrategien. Im Folgenden wird diese Systematik näher erläutert.

1. Nutzungsbezogene Risiken: Nutzungsbezogene Risiken entstehen aus der Art und Weise, wie Online-Plattformen genutzt werden, unabhängig davon, ob das Design der Plattform diese intendiert oder nicht. Während Online-Plattformen rechtlich für ihr Design, ihre Funktionen und die daraus resultierenden Möglichkeiten verantwortlich sind, können Risiken auch aus einer Nutzung resultieren, die nicht explizit vorgesehen war. Der DSA erkennt an, dass Risiken sowohl durch die Gestaltung als auch durch die Nutzung bedingt sein können, was die Verantwortung der Online-Plattformen zur Risikominderung nicht ausschließt.

Nutzungsbezogene Risiken variieren je nach individuellen Gewohnheiten und Anfälligkeiten der Nutzer*innen. So können Funktionen wie Auto-Play oder unendliches Scrollen süchtig machendes Verhalten fördern, wobei die tatsächliche Risikobelastung sowohl von der Plattformgestaltung als auch von der Nutzung abhängt. Solche Beispiele verdeutlichen, dass nutzungsbezogene Risiken häufig an der Schnittstelle zwischen designbedingten Faktoren und individuellen Nutzungspraktiken entstehen.

Nutzungsbezogene Risiken

Systemisches Risiko	Risiko-Unterkategorien
Negative Auswirkungen auf die Grundrechte	Auswirkungen auf die Menschenwürde durch die Gestaltung des algorithmischen Systems, die die freie Meinungsäußerung verhindert.
	Auswirkungen auf die Ausübung der Freiheit der Meinungsäußerung und Informationsfreiheit, einschließlich der Freiheit und des Pluralismus der Medien, durch die Gestaltung des algorithmischen Systems, die die freie Meinungsäußerung verhindert.
	Auswirkungen auf die Ausübung des Rechts auf Privatleben durch die Gestaltung des algorithmischen Systems, die die freie Meinungsäußerung verhindert oder zur Behinderung des Wettbewerbs zurückführen.
	Auswirkungen auf die Ausübung des Rechts auf Datenschutz durch die Gestaltung des algorithmischen Systems.
	Auswirkungen auf die Ausübung des Rechts auf Nichtdiskriminierung durch die Gestaltung des algorithmischen Systems.
	Auswirkungen auf die Ausübung des Verbraucherschutzes durch die Gestaltung des algorithmischen Systems.
	Auswirkungen auf die Ausübung des Rechts des Kindes durch nicht verständliche Funktionsweise des Dienstes für Minderjährige.
	Auswirkung auf die Ausübung des Rechts des Kindes durch die Aussetzung Minderjähriger zu Inhalten, die ihre Gesundheit oder ihre körperlichem geistige oder sittliche Entwicklung beinträchtigen können.
	Auswirkungen auf die Ausübung des Rechts des Kindes durch die Gestaltung von Online-Schnittstellen, die absichtlich oder unabsichtlich die Schwächen und Unerfahrenheit von Minderjährigen ausnutzen.
	Auswirkung auf die Ausübung des Rechts des Kindes durch die Gestaltung von Online-Schnittstellen, die absichtlich oder unabsichtlich zu Suchtverhalten führen kann.
Alle tatsächlichen oder absehbaren nachteiligen Auswirkungen in Bezug auf geschlechtsspezifische Gewalt, den Schutz der öffentlichen Gesundheit und von Minderjährigen sowie schwerwiegende nachteilige Folgen für das körperliche und geistige Wohlbefinden einer Person	Tatsächliche oder absehbare negative Auswirkungen auf den Schutz der öffentlichen Gesundheit in Bezug auf die Gestaltung, die Funktionsweise oder die Nutzung von VLOPs und VLOSEs.
	Tatsächliche oder absehbare negative Auswirkungen auf den Schutz Minderjähriger in Bezug auf die Gestaltung, die Funktionsweise oder die Nutzung von VLOPs und VLOSEs.
	Tatsächliche oder absehbare negative Auswirkungen auf das körperliche und geistige Wohlbefinden einer Person in Bezug auf die Gestaltung, die Funktionsweise oder die Nutzung von VLOPs und VLOSEs.
	Tatsächliche oder absehbare negative Auswirkungen auf geschlechtsspezifische Gewalt, in Bezug auf die Gestaltung, die Funktionsweise oder die Nutzung von VLOPs und VLOSEs.

Tabelle 1: Übersicht zu nutzungsbasierten Risiken

Mit dieser ersten Dimension und Prämisse können wir zwei weitere Hauptkategorien betrachten:

2. Inhaltsbezogene Risiken: Inhaltsbezogene Risiken beziehen sich auf Risiken, die durch die Erstellung, Verbreitung oder Verstärkung rechtswidriger Inhalte entstehen. Dazu zählen beispielsweise rechtswidrige Hassreden, Material über sexuellen Missbrauch von Kindern oder terroristische Inhalte, die weitreichende gesellschaftliche Schäden oder Rechtsverletzungen verursachen können.

Der enge Fokus dieser Kategorie auf rechtswidrige Inhalte ergibt sich aus ihrem erheblichen gesellschaftlichen Spaltungspotenzial, insbesondere aufgrund von Assoziationen mit Zensur sowie der strikten Pflicht zur Abwägung mit anderen Grundrechten wie der Meinungsfreiheit. Gleichzeitig sind inhaltsbezogene Risiken eng mit den härtesten Minderungsmaßnahmen verknüpft, insbesondere der Entfernung von Inhalten, die als rechtswidrig eingestuft werden. Dies stellt Online-Plattformen und Regulierungsbehörden vor erhebliche Herausforderungen, da sie einerseits rechtlichen Vorgaben entsprechen und andererseits gesellschaftliche Debatten über Meinungsfreiheit berücksichtigen müssen. Inhaltsbezogene Risiken entstehen somit an der Schnittstelle von Plattformarchitektur, rechtlichen Anforderungen und gesellschaftlichen Dynamiken.

3. Verhaltensbezogene Risiken: Verhaltensbezogene Risiken konzentrieren sich auf Akteur*innen, die Schwachstellen oder Nutzungsbedingungen von Online-Plattformen ausnutzen, um rechtswidrige oder schädliche Aktivitäten durchzuführen. Dazu gehören Straftaten wie der Verkauf verbotener Waren, die Verbreitung von Desinformation sowie missbräuchliche Praktiken, die darauf abzielen, Meinungsäußerungen zu unterdrücken oder Grundrechte wie die Menschenwürde, Privatsphäre, das Diskriminierungsverbot oder den Schutz der öffentlichen Gesundheit zu verletzen.

Verhaltensbezogene Risiken haben oft erhebliche gesellschaftliche Auswirkungen, da sie nicht nur individuelle Rechte verletzen, sondern auch die öffentliche Sicherheit und das Vertrauen in Online-Plattformen gefährden. Der Fokus auf verhaltensbezogene Risiken unterstreicht, dass Anbieter*innen nicht nur für Gestaltung ihrer Dienste, sondern auch für die Vermeidung von missbräuchlicher Nutzung durch Dritte Verantwortung tragen. Dies macht sie zu einer besonders dynamischen und schwer vorhersehbaren Risikokategorie.

Inhaltsbezogene Risiken

Systemisches Risiko	Risiko-Unterkategorien
Rechtswidrige Inhalte	Verbreitung von rechtswidriger Hassrede
	Verbreitung von Darstellung von sexuellem Missbrauch von Kindern
	Verbreitung rechtswidriger terroristischer Inhalte

Tabelle 2: Übersicht zu inhaltsbasierten Risiken

Verhaltensbezogene Risiken

Systemisches Risiko	Risiko-Unterkategorien
Rechtswidrige Inhalte	Missbrauch von VLOPs und VLOSEs für Straftaten sowie rechtswidrige Tätigkeiten wie ein nach Unions- oder nationalem Recht untersagter Verkauf von Waren oder Dienstleistungen, wie z.B. gefährlicher oder gefälschter Güter oder rechtswidrig gehandelter Tiere.
	Verbreitung rechtswidriger Inhalte über Konten mit besonders großer Reichweite oder andere Möglichkeiten der Verstärkung, die rechtswidrige Inhalte rasch und weit verbreitet.
Negative Auswirkungen auf die Grundrechte	Auswirkungen auf die Ausübung der Menschenwürde durch den Missbrauch ihres Dienstes durch die Übermittlung missbräuchlicher Mitteilungen.
	Auswirkungen auf die Ausübung der Freiheit der Meinungsäußerung und Informationsfreiheit, einschließlich der Freiheit und des Pluralismus der Medien, durch den Missbrauch von VLOPs und VLOSEs für die Übermittlung missbräuchlicher Nachrichten.
	Auswirkungen auf die Ausübung des Rechts auf Privatleben durch den Missbrauch der Dienste von VLOPs und VLOSEs für die Übermittlung missbräuchlicher Nachrichten.
	Auswirkungen auf die Ausübung des Rechts auf Datenschutz durch den Missbrauch der Dienste von VLOPs und VLOSEs.
	Auswirkungen auf die Ausübung des Rechts auf Nichtdiskriminierung durch den Missbrauch der Dienste von VLOPs und VLOSEs.
	Auswirkungen auf die Ausübung des Rechts auf Verbraucherschutz durch den Missbrauch der Dienste von VLOPs und VLOSEs.
Alle tatsächlichen oder absehbaren nachteiligen Auswirkungen auf die gesellschaftliche Debatte und auf Wahlprozesse und die öffentliche Sicherheit	Verbreitung von rechtswidriger Hassrede im Internet durch die Verstärkung und potenziell schnelle Verbreitung durch Empfehlungssysteme.
	Bedrohungen im Zusammenhang mit ausländischer Informationsmanipulation und -beeinflussung (»FIMI«) und das umfassendere Phänomen der Desinformation.
	Bedrohungen im Zusammenhang mit der Manipulation und Beeinflussung ausländischer Informationen (»FIMI«) und dem umfassenderen Phänomen der Desinformation, das durch Empfehlungssysteme verstärkt wird.
	Die Verbreitung von (gewalttätigen) extremistischen Inhalten mit dem Ziel, Menschen zu radikalisieren.
	Die Verbreitung von Inhalten, die durch neue Technologien wie generative künstliche Intelligenz erzeugt werden.
	Die Verbreitung von Inhalten die durch neue Technologien wie generative künstliche Intelligenz erzeugt werden, durch Empfehlungssysteme.
	Die Verstärkung und potenziell schnelle und weite Verbreitung von Inhalten, die nach europäischem oder mitgliedstaatlichem Recht rechtswidrig sind, z.B. Drohungen, gewalttätige extremistische und terroristische Inhalte.
	Rechtswidrige Hassreden oder Online-Belästigungen gegen politische Kandidat*innen oder Amtsinhaber*innen, Journalisten*innen, Wahlhelfer*innen oder andere am Wahlprozess Beteiligte.
Alle tatsächlichen oder absehbaren nachteiligen Auswirkungen in Bezug auf geschlechtsspezifische Gewalt, den Schutz der öffentlichen Gesundheit und von Minderjährigen sowie schwerwiegende nachteilige Folgen für das körperliche und geistige Wohlbefinden einer Person	Auswirkungen auf den Schutz der öffentlichen Gesundheit durch die Nutzung der VLOPs und VLOSEs, unter anderem durch Manipulation und Desinformationskampagnen.
	Schwerwiegende negative Folgen für das körperliche und geistige Wohlbefinden einer Person durch die Nutzung der VLOPs und VLOSEs, einschließlich durch Manipulation und koordinierte Kampagnen.
	Geschlechtsspezifische Gewalt durch die Nutzung der VLOPs und VLOSEs, auch durch Manipulation und koordinierte Kampagnen.

Tabelle 3: Übersicht zu verhaltensbasierten Risiken

Diese drei konzeptionellen Hauptkategorien – nutzungsbezogene Risiken, inhaltsbezogene Risiken und verhaltensbezogene Risiken – sollen Online-Plattformen, Regulierungsbehörden und Forscher*innen helfen, präzise Indikatoren zu entwickeln, um die Art von Risiken auf Online-Plattformen besser zu identifizieren und zu bewerten. Dies ermöglicht zusätzlich eine klarere Einschätzung, welche Minderungsmaßnahmen angemessen und wirkungsvoll eingesetzt werden können. Dabei wird davon ausgegangen, dass alle Arten von Risiken in der Online-Umgebung durch das Design von Online-Plattformen beeinflusst werden.

In den folgenden Abschnitten wird gezeigt, wie der Ansatz Online-Plattformen dabei unterstützt, zu beurteilen, ob ihr Design systemische Risiken verstärken oder abschwächen kann. Gleichzeitig wird verdeutlicht, dass das Verbinden von Risiken und Risikominderungsmaßnahmen eine anspruchsvolle Aufgabe ist, die eine ständige Abwägung zwischen verschiedenen Werten erfordert. Ein Beispiel für eine solche Werteabwägung liefert die Frage nach dem richtigen Maß an Plattformmoderation. So könnte eine Plattform, die anonyme Nutzerinteraktionen ohne Moderationsinstrumente zulässt, für die potenzielle Verstärkung schädlichen Verhaltens zur Verantwortung gezogen werden, da Anonymität in Verbindung mit unzureichender Aufsicht Belästigungen oder rechtswidrige Inhalte begünstigen kann. Im Gegensatz dazu könnte eine Plattform mit robusten Moderationsfunktionen und Transparenzmechanismen solche Risiken wirksamer eindämmen und wäre weniger für den Missbrauch verantwortlich, der sich diesen Kontrollen entzieht. Allerdings birgt auch dies auch Risiken: Maßnahmen wie eine Klarnamenpflicht oder übermäßig strenge Moderation könnten das Recht auf freie Meinungsäußerung beeinträchtigen.

Online-Plattformen müssen daher erkennen, wie ihr Design die Nutzung, den Inhalt und das Nutzerverhalten prägt. Nur dann können sie proaktiv die vorhersehbaren Risiken ihrer Infrastruktur adressieren, ohne dabei grundlegende Rechte zu gefährden. Diese Abwägungsfragen werden im weiteren Verlauf deutlicher, wenn wir den Risikobewertungsprozess erläutern.

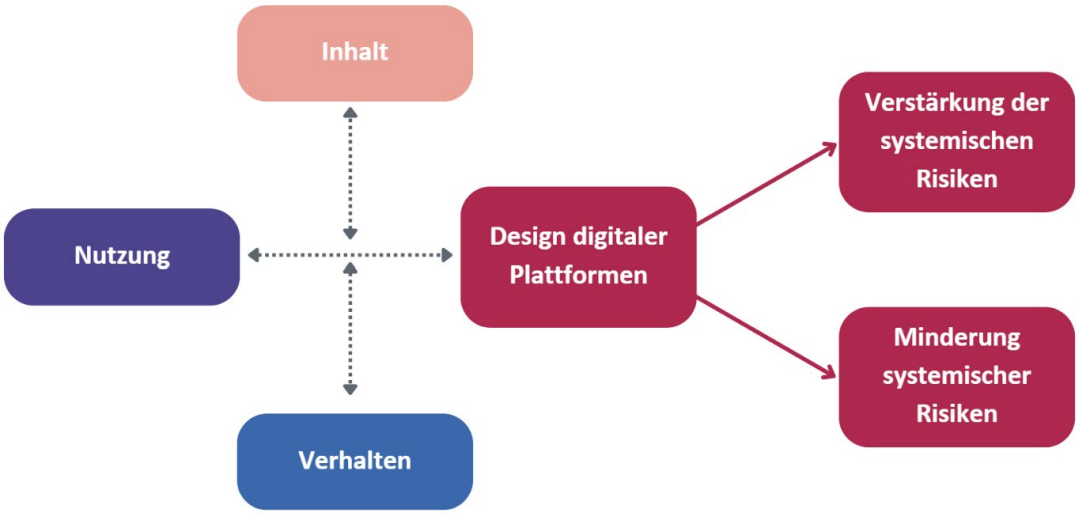

Abbildung 1: Beziehung zwischen Risikoebenen, Plattformdesign und systemischen Risiken

Fallstudie: Online-Bots

Um die theoretischen Schritte des Risikobewertungsprozesses nachzuvollziehen, wird in der vorliegenden Studie eine Fallstudie zum Risiko durch Online-Bots auf der Plattform X herangezogen. Ziel ist es, aufzuzeigen, wie Online-Plattformen ein spezifisches Risiko basierend auf ihrem Risikoprofil identifizieren, Aussagen über dessen Ausmaß und Wahrscheinlichkeit treffen und entsprechende Risikominderungsmaßnahmen ableiten können.

Das vorliegende Risiko basiert auf einem realen Beispiel, das in der Form schon eingetreten ist: Ende 2023 deckten Ermittler*innen des Auswärtigen Amtes ein von Russland unterstütztes Netzwerk zur Informationsmanipulation auf, das sich auf »Doppelgänger-Medien« spezialisiert hatte und mutmaßlich automatisiert manipulative Beiträge in sozialen Netzwerken teilte. Diese Doppelgänger imitierten dabei nicht nur das Auftreten vertrauenswürdiger Nachrichtenseiten, sondern gaben sich teilweise auch als offizielle Account von deutschen Politiker*innen aus. Zwischen dem 20. Dezember 2023 und dem 20. Januar 2024 identifizierten Ermittler*innen über fünfzigtausend gefälschte Nutzerkonten, die mehr als eine Million deutschsprachige Beiträge veröffentlichten und amplifizierten (Rosenbach & Schult, 2024). Das Auftreten der Online-Bots auf der Plattform X verdeutlicht somit, dass aufgrund der Dienste und Merkmale von X das Risiko besteht, dass auch in Zukunft Online-Bots systematisch manipulierte Inhalte verbreiten. Im Risikobewertungsprozess werden im Folgenden die Auswirkungen und Wahrscheinlichkeit des Eintreffens dieses potentiellen Risikos analysiert, jedoch ohne auf die vom deutschen Außenministerium bereits aufgedeckten Informationen zurückzugreifen.

Online-Bots oder sog. Social-Bots können als »automatisierte Konten« definiert werden, »die künstliche Intelligenz nutzen, um Diskussionen zu lenken und bestimmte Ideen oder Produkte in sozialen Medien wie Twitter und Facebook zu bewerben. Für typische Social-Media-Nutzer, die ihre Feeds durchsuchen, können Social-Bots unbemerkt bleiben, da sie so gestaltet sind, dass sie dem Aussehen menschlicher Nutzer*innen ähneln (z.B. zeigen sie ein Profilfoto und geben einen Namen oder Standort an) und sich online ähnlich wie Menschen verhalten (z.B. retweeten oder zitieren sie andere Beiträge und liken oder unterstützen andere Tweets).« (Allem & Ferrara, 2018, S. 1005) Wie einflussreich Online-Bots auf Online-Plattformen sind, ist in der Forschung aufgrund von methodologischen Herausforderungen bei der Messung von Bot-Aktivitäten umstritten (Keller & Klinger, 2019). Gleichzeitig werden immer wieder Botnetzwerke mit einer teils sehr großen Reichweite aufgedeckt. Das verdeutlicht Schwachstellen von Online-Plattformen, die von Akteur*innen ausgenutzt werden können, um in großem Maße Inhalte zu verbreiten.

Die Verwendung von Online-Bots wird im Rahmen des DSA als Risiko eingestuft, da sie zur betrügerischen Nutzung eines Dienstes und zur Ausnutzung von Schwachstellen in einer Plattform verwendet werden können. Nach dem DSA fallen die Online-Bots aus dem vorliegenden Beispiel in das Risiko von »Bedrohungen im Zusammenhang mit ausländischer Informationsmanipulation- und Beeinflussung (»FIMI«) und das umfassendere Phänomen der Desinformation«. Da Akteur*innen im vorliegenden Fall gezielt die Funktionen der Online-Plattformen nutzten, um menschliche Nutzer*innen zu imitieren oder Inhalte zu amplifizieren, werden Online-Bots nach den ISD-Risikoebenen als verhaltensbezogenes Risiko kategorisiert.

3.2 Bewertungsprozess

Die Realisierung systemischer Risiken auf Online-Plattformen kann den Nutzer*innen der Dienste sowie der Gesellschaft schweren Schaden zufügen. Solche digitalen Schäden können sich auf verschiedene Weise äußern: So beeinträchtigen Datenschutzverletzungen, Desinformation, Online-Mobbing oder algorithmische Verzerrungen die Privatsphäre, Sicherheit und Autonomie der Nutzer*innen. Der kollektive Schaden für die Gesellschaft wiederum liegt darin, öffentliches Vertrauen zu untergraben, Wahlentscheidungen zu beeinflussen und zur gesellschaftlichen Polarisierung beizutragen (Europäische Kommission, 2024b).

Die Risikobewertung von VLOPs und VLOSEs muss daher aufgrund des tiefgreifenden Einflusses, den diese Online-Plattformen auf die öffentliche Debatte, die Privatsphäre und das individuelle Wohlergehen haben, grundsätzlich mit menschenrechtlichen Erwägungen verknüpft werden. Online-Plattformen operieren in einer Größenordnung, in der ihre Aktivitäten Millionen von Menschen weltweit betreffen, und ihre Geschäftspraktiken können direkt oder indirekt grundlegende Menschenrechte verletzen. Aus diesem Grund hat sich das ISD bei der Entwicklung eines umfassenden Risikobewertungsrahmens für VLOPs und VLOSEs von den UNGPs leiten lassen. Die UNGPs bieten einen weltweit anerkannten Rahmen, der Unternehmen hilft, ihre Praktiken mit Menschenrechtsstandards in Einklang zu bringen und Rechenschaftspflicht und Verantwortung in ihren Aktivitäten zu betonen.

Insbesondere das Konzept der menschenrechtlichen Sorgfaltspflicht (engl. Human Rights Due Diligence), wie es in den UNGPs dargelegt ist, dient als Eckpfeiler der Risikobewertungsmethodik des ISD (Office of the High Commissioner for Human Rights, 2021). Sie fordert von Unternehmen, Menschenrechtsrisiken aktiv zu erkennen, zu verhindern oder zu mindern und gleichzeitig die Verantwortung für die Beseitigung möglicher negativer Auswirkungen zu übernehmen. Dieser Sorgfaltsprozess umfasst einen strukturierten Ansatz, der mit der Beurteilung tatsächlicher und potenzieller menschenrechtlicher Auswirkungen beginnt und mit der Überwachung und Kommunikation von Maßnahmen zur Behebung dieser Probleme endet. Dieser dynamische Prozess nimmt die Unternehmen in die Pflicht, sicherzustellen, dass ihre Aktivitäten nicht gegen die Menschenrechte verstoßen, und bietet ihnen einen Rahmen für die Durchführung kontinuierlicher Bewertungen und die Ergreifung von Maßnahmen zum Schutz der Menschenwürde und für ethisches Geschäftsgebaren.

Das ISD hat den vierstufigen Ansatz des Rahmens für menschenrechtliche Sorgfaltspflicht an die spezifischen Bedürfnisse und Anforderungen des DSA angepasst und konzentriert sich dabei auf das Management von systemischen Online-Risiken im Zusammenhang mit VLOPs und VLOSEs. Angesichts des weit verbreiteten Einflusses dieser Anbieter*innen und der Größenordnung, in der sie operieren, konzentriert sich die Anpassung dieses Ansatzes für die Einhaltung des DSA auf die Art und Weise, wie diese Online-Plattformen ihren Betrieb gestalten und steuern, um negative Auswirkungen auf die Nutzer*innen und die Gesellschaft zu vermeiden. Durch die Anpassung des Ansatzes an den regulatorischen und praktischen Kontext des DSA versucht das ISD, ein strukturiertes und umsetzbares Modell bereitzustellen, das Transparenz, Rechenschaftspflicht und Menschenrechte in den Vordergrund stellt.

Die ISD-Methode zur Risikobewertung und -minderung im Zusammenhang mit VLOPs und VLOSEs im Rahmen des DSA unterscheidet gemeinhin drei Phasen (s. Ab-

Abbildung 2: Allgemeiner Bewertungsprozess im DSA-Kontext

bildung 2). Jede Phase soll Online-Plattformen dabei helfen, die ihren Diensten innewohnenden Risiken umfassend zu ermitteln und darauf zu reagieren – vom Verständnis potenzieller Schäden bis zur Umsetzung von Strategien zur Risikominderung. Dieser Ansatz gibt Online-Plattformen einen Fahrplan für ein ethisches, effektives Risikomanagement an die Hand, das sowohl mit den Erwartungen der Regulierungsbehörden als auch mit den Menschenrechtsprinzipien übereinstimmt und das proaktive Engagement und die Transparenz in jeder Phase der Risikobewertung betont (Erwägungsgrund 90 DSA).

1) Identifizierung der Risiken: Die erste Phase des ISD-Ansatzes konzentriert sich auf die Identifikation und Priorisierung von Risiken, so dass VLOPs und VLOSEs ein klares Bild von potenziellen Bedrohungen innerhalb des Ökosystems ihrer Plattform erhalten. Durch die Analyse des einzigartigen Designs und der Funktionalität ihrer Plattform können sie diejenigen Risiken identifizieren, die am wahrscheinlichsten auftreten – sei es aufgrund der algorithmischen Struktur, des Nutzungsverhaltens oder der Praktiken der Inhaltsmoderation.

2) Bewertung von Auswirkung und Wahrscheinlichkeit: Sobald potenzielle Risiken identifiziert sind, sollten VLOPs und VLOSEs die Auswirkung und Wahrscheinlichkeit jedes einzelnen Risikos bewerten. Dies beinhaltet die Bewertung des Schweregrads jedes Risikos (Ausmaß), die Identifizierung potenziell betroffener Nutzergruppen (Umfang) und die Bewertung der Leichtigkeit, mit der diese Risiken gemindert werden können (Abhilfefähigkeit). Darüber hinaus sollte die Wahrscheinlichkeit des Eintretens dieser Risiken auf der Grundlage historischer Daten, aktueller Trends und kontextbezogener Faktoren wie z.B. Änderungen der Rechtsvorschriften oder Änderungen des Nutzungsverhalten, bewertet werden.

3) Risikominderung und Evaluation: In der dritten Phase kombinieren die VLOPs und VLOSEs die Erkenntnisse aus der Risikoidentifikation und der Bewertung von Auswirkung, um Maßnahmen zur Risikominderung zu formulieren, umzusetzen und ihre Wirksamkeit zu bewerten. Online-Plattformen entwickeln auf Grundlage der Art, Auswirkungen und Wahrscheinlichkeit der identifizierten Risiken Präventionsstrategien, die auf ihren spezifischen betrieblichen Kontext zugeschnitten sind. Dabei gleichen sie die systemischen Risiken mit geeigneten Minderungsmaßnahmen ab und wählen unter Berücksichtigung der Verhältnismäßigkeit die angemessensten und wirkungsvollsten Ansätze aus.

Nach der Implementierung dieser Maßnahmen ist es entscheidend, ihre Effektivität zu prüfen. Online-Plattformen müssen bewerten, ob die Maßnahmen die gewünschten Ergebnisse erzielen und das jeweilige Risiko tatsächlich vermindern. Diese fortlaufende Evaluierung gewährleistet, dass Minderungsmaßnahmen optimal angepasst und bei Bedarf weiterentwickelt werden, um die Herausforderungen eines dynamischen digitalen Umfelds wirksam zu adressieren.

Laufende Schritte: Ein zentrales Element des ISD-Ansatzes ist die Betonung der kontinuierlichen Einbeziehung von Stakeholder*innen und die transparente Kommunikation von Risikomanagementmaßnahmen. VLOPs und VLOSEs werden ermutigt, sich regelmäßig mit Interessengruppen, einschließlich betroffener Gemeinschaften, der Zivilgesellschaft und Menschenrechtsexpert*innen, zu beraten, um ihre Bewertungen zu validieren und Feedback zu ihren Minderungsstrategien zu erhalten (Erwägungsgrund 90 DSA). Dieser kontinuierliche Dialog ermöglicht es den Online-Plattformen, ihren Ansatz als Reaktion auf Erkenntnisse aus der Praxis und Empfehlungen von Expert*innen zu verfeinern. Darüber hinaus fördern die Online-Plattformen das Vertrauen und die Rechenschaftspflicht, indem sie offen über ihre Bemühungen und Erfolge beim Risikomanagement in Bezug auf den DSA berichten und so ihr Engagement für einen verantwortungsvollen und transparenten Umgang mit digitalen Schäden unter Beweis stellen.

Auf der Grundlage des oben vorgestellten konzeptionellen Rahmens wurden die abstrakten Phasen dieser Studie in ein konkretes, schrittweises Verfahren zur Risikobewertung übersetzt. Dieser Prozess soll einen strukturierten und umsetzbaren Ansatz für VLOPs und VLOSEs bieten, um die im DSA, insbesondere in Art. 34 Abs. 2, festgelegten Anforderungen zu erfüllen. Basierend auf einer umfassenden Literaturrecherche und angereichert durch die Einbeziehung von Interessengruppen- und Experteninterviews soll die Methodik einen zielorientierten Prozess zur Identifizierung, Bewertung und Minderung von Risiken in Übereinstimmung mit den Erwartungen der EU-Regulierungsbehörden darstellen.

Abbildung 3: Spezifischer Bewertungsprozess im DSA-Kontext

Schritt 1: Risikoidentifikation

a) Bestimmung des Ausgangspunkts

Gemäß dem DSA sind Online-Plattformen, die als VLOPs und VLOSEs eingestuft sind, verpflichtet, innerhalb bestimmter Fristen umfassende Risikobewertungen durchzuführen. Art. 34 Abs. 2 DSA hebt hervor, dass diese Bewertungen bis zum in Art. 33 Abs. 6 DSA genannten Anwendungszeitpunkt durchgeführt werden müssen, wobei die nachfolgenden Überprüfungen mindestens einmal jährlich erfolgen müssen. Darüber hinaus sind Risikobewertungen immer dann erforderlich, wenn neue Funktionalitäten eingeführt werden, die die Risikolandschaft der Plattform wesentlich verändern können. Diese Anforderung stellt sicher, dass die Online-Plattformen einen proaktiven und dynamischen Ansatz für das Risikomanagement verfolgen.

Um den Prozess der Risikobewertung gemäß den Vorgaben des DSA einzuleiten, müssen VLOPs und VLOSEs einen klar definierten Ausgangspunkt festlegen. Der zentrale Ansatzpunkt sind die gesetzlich vorgeschriebenen jährlichen Risikoanalysen, die eine systematische Bewertung der Plattformaktivitäten sicherstellen. Ergänzend können jedoch auch andere Auslöser, wie die Einführung neuer Dienste oder Funktionen sowie signifikante Veränderungen im Nutzerverhalten, eine zusätzliche Überprüfung notwendig machen. Dieser Schritt ist entscheidend, um sicherzustellen, dass die Online-Plattformen ihre Prozesse sowohl im Einklang mit den DSA-Vorgaben als auch proaktiv auf potenzielle Risikofaktoren ausgerichtet durchführen.

b) Heranziehen des Risikoprofils und Einbeziehung externer Gruppen

Ein solides Verständnis der Kernrisiken ist von grundlegender Bedeutung für den Risikobewertungsprozess. Online-Plattformen sollten ihr Risikoprofil, das das Verständnis einer Organisation von potenziellen Schäden auf der Grundlage ihrer betrieblichen Merkmale, Nutzerinteraktionen und systemischen Schwachstellen zusammenfasst, systematisch überprüfen. Dieses Profil dient als Grundlage für die Identifizierung von Bereichen, in denen Schäden auftreten können, und für die Festlegung von Prioritäten bei den Bemühungen zur Risikominderung. Der Prozess der Risikoidentifizierung und -Priorisierung kann durch die Einbeziehung von Nutzer*innen, den möglicherweise von den Diensten betroffenen Gruppen sowie unabhängigen Sachverständigen und zivilgesellschaftlichen Organisationen unterstützt werden. Diese umfassende Konsultation von internen Daten, Nutzerfeedback und externen Sta-

keholder*innen stellt sicher, dass die Perspektiven aller relevanten Akteur*innen berücksichtigt werden (s. Erwägungsgrund 90 DSA).

In Anlehnung an die Ofcom-Leitlinien zur Risikobewertung von Diensten nutzt diese Studie einen strukturierten Fragenkatalog zur Analyse von Risikofaktoren (Ofcom, 2023). Dieser Rahmen legt den Schwerpunkt auf eine detaillierte Untersuchung der Dienste, z.B. von spezifischen Merkmalen wie Größe und Nutzerbasis, Funktionen für die Erstellung und den Austausch von Inhalten sowie Möglichkeiten der Nutzerinteraktion. Die Online-Plattformen sollten gezielte Fragen beantworten, die auf die Bestimmungen des DSA und anderer einschlägiger EU-Rechtsvorschriften abgestimmt sind, um diese Risikofaktoren umfassend zu erfassen (s. 3.3 für den vollständigen Fragenkatalog).

Die aus dieser Analyse gewonnenen Erkenntnisse ermöglichen es den Online-Plattformen, ihre Risikofaktoren systematisch zu erfassen und so ein detailliertes und umsetzbares Risikoprofil zu erstellen. Dieses Profil zeigt nicht nur potenzielle Schwachstellen auf, sondern bietet auch einen Leitfaden für die Priorisierung von Strategien zur Risikominderung.

Ergebnis von Schritt 1

Das Ergebnis dieses ersten Schritts ist eine umfassende Bestandsaufnahme potenzieller Risiken, die sich aus der einzigartigen Betriebs- und Nutzerdynamik der jeweiligen Plattform ergeben. Auf der Grundlage dieser Bestandsaufnahme können die Online-Plattformen die zu bewertenden Risiken priorisieren. Wie diese Priorisierung stattfindet, wird im folgenden Schritt erläutert.

Schritt 2: Risikobewertung

Aufbauend auf den Erkenntnissen der Risikoidentifikation erfolgt im nächsten Schritt eine gründliche Abschätzung der Auswirkungen und der Wahrscheinlichkeit der identifizierten Risiken. Dieser Prozess ist für die Klassifizierung potenzieller Schäden und Priorisierung der Ressourcen entscheidend, um eine wirksame Schadensbegrenzung zu ermöglichen. Erwägungsgrund 79 DSA bietet eine grundlegende Anleitung und betont, dass Online-Plattformen sowohl die Schwere potenzieller negativer Auswirkungen als auch die Wahrscheinlichkeit solcher systemischen Risiken berücksichtigen sollten.

a) Abschätzung der Auswirkungen

In der Phase der Abschätzung der Auswirkungen werden die Risiken anhand von drei Kerndimensionen bewertet, die aus dem UN-Rahmen für die Risikobewertung im Bereich der Menschenrechte abgeleitet wurden: **Ausmaß, Umfang** und **Abhilfemöglichkeit.** Diese Dimensionen lehnen sich eng an Erwägungsgrund 79 des DSA an und umfassen Folgendes:

- Das Ausmaß beschreibt die Intensität potenzieller Auswirkungen, wobei der Fokus auf der Schwere der Konsequenzen liegt. Berücksichtigt werden mögliche irreversible Schäden, wie etwa die Beeinträchtigung des Wohlbefindens, gesellschaftliche oder wirtschaftliche Folgen, sowie die Fähigkeit eines Risikos, weitreichende und dauerhafte negative Veränderungen zu bewirken.
- Der Umfang bezieht sich auf die Verbreitung der Auswirkungen eines Risikos. Dies umfasst die Anzahl der potenziell betroffenen Akteur*innen, die geografische und soziale Reichweite sowie die mögliche Ausbreitung über einen spezifischen Kontext hinaus. Ein größeres Ausmaß der Reichweite signalisiert ein höheres Risiko für breite gesellschaftliche oder globale Auswirkungen.
- Die Abhilfefähigkeit misst die Fähigkeit, den ursprünglichen Zustand wiederherzustellen. Sie umfasst die Bewertung der Effektivität von Maßnahmen, wie etwa die Unterstützung betroffener Akteur*innen, oder strukturelle Anpassungen. Die Umsetzbarkeit solcher Maßnahmen und deren Nachhaltigkeit spielen hierbei eine zentrale Rolle.

Anbieter*innen sollten potenzielle Risiken systematisch anhand dieser drei Dimensionen bewerten, um einen umfassenden Überblick über die Art und das Ausmaß eines möglichen Schadens zu erhalten. Bei der Bewertung von Risiken im Zusammenhang mit rechtswidrigen Hassreden können Online-Plattformen beispielsweise das Ausmaß des emotionalen und psychologischen Schadens, die Reichweite im Hinblick auf die betroffenen Bevölkerungsgruppen und die Durchführbarkeit der Wiederherstellung der Würde oder der Minderung von Schäden, die durch solche Inhalte verursacht wur-

den, analysieren. Diese Bewertungen helfen den Online-Plattformen, schwerwiegende, weit verbreitete und schwer zu behebende Risiken zu priorisieren. In Kapitel 4 dieses Berichts wird eine Liste möglicher Indikatoren für jede Art von Risiko und Bewertung von Auswirkung aufgeführt.

b) Abschätzung der Wahrscheinlichkeit

Parallel zur Abschätzung der Auswirkungen von Risiken müssen die Online-Plattformen auch die Wahrscheinlichkeit ihres Eintretens bewerten. Dies beinhaltet eine Analyse von drei Hauptaspekten: Plattformmerkmalen, historischen Daten und Trends sowie Kontextfaktoren.

- Plattformmerkmale: Diese Dimension steht in direktem Zusammenhang mit dem Risikoprofil der Plattform, das in der Phase der Risikoidentifizierung festgelegt wurde. Wenn man versteht, wie plattformspezifische Merkmale (z.B. die Größe der Nutzerbasis, das algorithmische Design oder interaktive Funktionen) mit identifizierten Risiken zusammenhängen, können Online-Plattformen die Wahrscheinlichkeit des Auftretens dieser Risiken einschätzen.
- Historische Daten und Trends: Online-Plattformen sollten sich auf ihre historische Bilanz stützen, um das erneute Eintreten bestimmter risikoreicher Ereignisse zu beurteilen. Dazu gehört auch die Analyse früherer Vorfälle, um festzustellen, ob die Risikominderungsmaßnahmen wirksam waren oder ob der Schaden trotz der Interventionen anhaltend ist. Wenn eine Plattform beispielsweise wiederholt Probleme mit der Moderation extremistischer Inhalte hatte, könnte dies ein Hinweis darauf sein, dass ähnliche Risiken mit hoher Wahrscheinlichkeit wieder auftreten. In diesem Zusammenhang hebt Erwägungsgrund 85 DSA die Relevanz der Verfügbarkeit historischer Daten hervor, die durch Belege für frühere Risikobewertungen gestärkt werden können.
- Kontextuelle Faktoren: Externe Ereignisse und breitere soziopolitische oder wirtschaftliche Zusammenhänge können die Wahrscheinlichkeit von Risiken erheblich beeinflussen. Die Online-Plattformen müssen berücksichtigen, wie sich Faktoren wie Wahlen, globale Krisen oder regionale Instabilität auf das Nutzerverhalten und die Risikolandschaft auswirken. Diese Dimension umfasst qualitative Forschung und Szenarienplanung, um zu antizipieren, wie spezifische Kontexte mit der Plattformdynamik interagieren können. So kann beispielsweise während einer Wahl die Wahrscheinlichkeit von Risiken wie politischer Desinformation oder direkter Belästigung erheblich steigen.

Zusammen ermöglichen diese Dimensionen den Online-Plattformen, die Wahrscheinlichkeit des Eintretens von Risiken genauer zu bewerten. Bei der Analyse von Risiken im Zusammenhang mit Desinformation während einer öffentlichen Gesundheitskrise können Online-Plattformen beispielsweise bewerten, wie ihre algorithmischen Verstärkungsmechanismen (Plattformmerkmale), frühere Vorfälle und Desinformationsnarrative (historische Daten) und aktuelle Trends im Nutzerverhalten (Kontextfaktoren) zusammenwirken und hierdurch die Wahrscheinlichkeit eines Schadens beeinflussen.

Ergebnis von Schritt 2

Die kombinierte Bewertung von Bewertung von Auswirkung und Wahrscheinlichkeit bietet Online-Plattformen einen soliden Rahmen für eine umfassende Risikobewertung. Indem sie sowohl den Schweregrad als auch die Wahrscheinlichkeit potenzieller Schäden verstehen, können Online-Plattformen die Risiken priorisieren, die die größten Bedrohungen für Nutzer*innen und die Gesellschaft darstellen. Dabei ist es essenziell, dass diese Bewertung immer nur für ein identifiziertes Risiko gleichzeitig erfolgt, um eine gezielte und gründliche Analyse zu gewährleisten. Diese beidseitige Analyse, die sich auf Fakten stützt und mit den Grundsätzen des DSA übereinstimmt, bildet die Grundlage für die nächsten Schritte der Risikominderung, einschließlich der Einführung von Risikominderungsmaßnahmen und der anschließenden Bewertung ihrer Wirksamkeit.

Schritt 3: Risikominderung

Der dritte Schritt des Risikobewertungsprozesses konzentriert sich auf die Umsetzung und Bewertung von Maßnahmen zur Minderung der identifizierten Risiken. Dieser Schritt ist von entscheidender Bedeutung, um sicherzustellen, dass potenzielle Risiken nicht zu Online-Schäden führen, in Übereinstimmung mit den im DSA festgelegten Verpflichtungen. Die Online-Plattformen müssen nicht nur geeignete Maßnahmen umsetzen, sondern auch deren Wirksamkeit bewerten und ihren Ansatz kontinuierlich verfeinern auf Grundlage der sich entwickelnden Risiken und der Dynamik der Plattform. Dabei sollten sie gemäß Erwägungsgrund 86 DSA die Verhältnismäßigkeit unter Berücksichtigung ihrer wirtschaftlichen Leistungsfähigkeit beachten, unnötige Beschränkungen für die Nutzung ihrer Dienste vermeiden und mögliche negative Auswirkungen auf die Grundrechte und andere schützenswerte Interessen angemessen berücksichtigen, insbesondere die Meinungsfreiheit.

c) Umsetzung von Minderungsmaßnahmen

Nachdem die potenziellen Risiken hinsichtlich ihrer Auswirkungen und Wahrscheinlichkeit gründlich bewertet wurden, können die Online-Plattformen zur praktischen Umsetzung von Minderungsmaßnahmen übergehen. Dies beinhaltet ein strukturiertes Vorgehen bei der Auswahl und Anwendung von Maßnahmen, die auf die Bewältigung spezifischer Risiken ausgerichtet sind.

- **Heranziehen des Fragenkatalogs der risikomindernden Maßnahmen:** Online-Plattformen sollten zunächst Ressourcen wie den hier erarbeiteten Katalog der Risikominderungsmaßnahmen (s.u.) heranziehen, der Risikominderungsmaßnahmen aus verschiedenen EU-Verordnungen auflistet. Dieser Katalog dient als umfassender Leitfaden, der es den Online-Plattformen ermöglicht, die aufsichtsrechtlich relevanten Maßnahmen zu identifizieren, die für die von ihnen identifizierten Risiken am besten geeignet sind. Maßnahmen zur Bekämpfung von Desinformation könnten beispielsweise eine verstärkte Moderation von Inhalten oder Transparenz bei Werbepraktiken umfassen, während Maßnahmen zum Jugendschutz strengere Protokolle zur Altersverifikation beinhalten könnten.
- **Anwendung der Verhältnismäßigkeitsprüfung:** Online-Plattformen müssen bei der Auswahl von Minderungsmaßnahmen den Grundsatz der Verhältnismäßigkeit beachten, indem sie prüfen, ob die Maßnahme geeignet, notwendig und im Verhältnis zum Nutzen steht, ohne unverhältnismäßige Belastungen für Betroffene zu verursachen. Kapitel 5 bietet eine detaillierte Betrachtung des Verhältnismäßigkeitsgrundsatzes und seiner Messgrößen.

Beispiel: Eine Plattform, die algorithmische Änderungen vornimmt, um die Verbreitung schädlicher Inhalte einzudämmen, muss sicherstellen, dass diese Änderungen die Meinungsfreiheit der Nutzer*innen nicht unverhältnismäßig einschränken oder bestimmte Gruppen in unfairer Weise benachteiligen. In einem solchen Fall wäre bei der Bewertung, ob eine Maßnahme angemessen, erforderlich und verhältnismäßig ist, ein besonderes Augenmerk auf das Recht auf freie Meinungsäußerung zu legen (Erwägungsgrund 90 DSA). Dieses Beispiel zeigt, wie Online-Plattformen durch die systematische Anwendung dieser Prinzipien gewährleisten können, dass ihre Maßnahmen sowohl geeignet als auch mit rechtlichen und ethischen Standards vereinbar sind.

d) Bewertung der Risikominderungsmaßnahmen

Die Bemühungen zur Risikominderung enden nicht mit der Umsetzung; die Online-Plattformen müssen die Wirksamkeit ihrer Maßnahmen zur Minderung der Risiken, die sie angehen wollen, kontinuierlich bewerten. Gemäß Art. 35 DSA und Erwägungsgrund 86 DSA müssen VLOPs und VLOSEs sicherstellen, dass die Risikominderungsmaßnahmen wirksam sind, und einen Rahmen für die laufende Bewertung und Verbesserung schaffen. Wenn eine Plattform beispielsweise strengere Moderationsrichtlinien einführt, um Hassrede zu bekämpfen, kann die Wirksamkeit anhand des Rückgangs der gekennzeichneten Inhalte oder der Nutzermeldungen über Vorfälle von Hassreden im Laufe der Zeit gemessen werden. Ein ausführlicherer Überblick über die möglichen Ansätze zur Bewertung der Wirksamkeit wird in Kapitel 5 dieser Studie gegeben.

- **Kontinuierliche Überwachung:** Minderungsmaßnahmen müssen dynamisch sein und sich in Reaktion auf neue Entwicklungen, Nutzerverhalten oder äußere Umstände weiterentwickeln. Regelmäßige Audits, wie in Art. 37 DSA vorgeschrieben, und Datenanalysen können Online-Plattformen dabei helfen, festzustellen, ob zuvor wirksame Maßnahmen an Wirkung verlieren oder unbeabsichtigt neue Risiken schaffen.
- **Identifizierung von Sekundärrisiken:** Ein entscheidender Aspekt in diesem Schritt ist die Frage, ob Risikominderungsmaßnahmen zu Sekundärrisiken geführt haben. Die Erkenntnisse aus der Einbindung der Stakeholder*innen zeigen, dass einige Maßnahmen zwar primäre Risiken bekämpfen, aber unbeabsichtigt neue Herausforderungen schaffen können. Ein Beispiel hierfür sind Maßnahmen gegen rechtswidrige Hassrede (Mchangama et al., 2023). Während Anbieter*innen darauf abzielen, rechtswidrige Inhalte zu reduzieren, können sie unbeabsichtigt in die freie Meinungsäußerung eingreifen, etwa durch übermäßiges Entfernen legitimer Inhalte oder durch die Einschränkung kontroverser, aber rechtmäßiger Debatten. Die Online-Plattformen sollten ein Verfahren einrichten, um solche sekundären Risiken umgehend zu bewerten und anzugehen, um einen ganzheitlichen Ansatz zur Risikominderung zu gewährleisten (Erwägungsgründe 86, 90 DSA).

Ergebnis von Schritt 3

Risikominderung ist ein vielschichtiger Prozess, der sorgfältige Planung, fundierte Entscheidungen und eine kontinuierliche Bewertung erfordert. Durch die Konsultation von aufsichtsrechtlichen Katalogen zur Risikominderung, die Anwendung des Verhältnismäßigkeitstests und die Überwachung der Wirksamkeit anhand von Schlüsselmetriken können Online-Plattformen sicherstellen, dass ihre Maßnahmen Risiken effektiv angehen und gleichzeitig unbeabsichtigte Folgen minimieren. Der iterative Charakter dieses Prozesses ermöglicht es ihnen, sich an neu auftretende Risiken anzupassen und die Anforderungen des DSA zu erfüllen, wodurch sie zu einer sichereren und vertrauenswürdigeren Online-Umgebung beitragen.

Fallstudie: Online-Bots

Schritt 1: Risikoidentifikation
In der Fallstudie wird davon ausgegangen, dass das Vorhandensein von Online-Bots auf X bereits als potenzielles Risiko identifiziert wurde. Das löst den Prozess der Risikobewertung aus.

Schritt 2: Risikobewertung
In Kapitel 4 wird das potenzielle Risiko von Online-Bots auf X anhand der Auswirkungen (s. 4.1) und der Wahrscheinlichkeit (s. 4.2) bewertet und anschließend in die Risikomatrix eingeordnet, um zu verstehen, ob das Risiko eine sehr hohe, hohe, mittlere oder geringe Auswirkung und Wahrscheinlichkeit hat (s. 4.3). In diese Bewertung fließen spezifische Indikatoren und Datenquellen ein, die von der Plattform bereitgestellt werden.

Schritt 3: Risikominderung
In Kapitel 5 werden die am besten geeigneten Maßnahmen zur Risikominderung ausgewählt, um das potenzielle Risiko der Präsenz von Online-Bots auf X in Deutschland zu mindern.

Nach der detaillierten Herleitung der ISD-Risikoebenen und der Darstellung der grundlegenden Schritte des Bewertungsprozesses stellt sich aus praktischer Anwendungssicht die Frage, wie die Identifikation von potenziellen Kernrisiken zu Beginn des Prozesses abläuft. Im folgenden Abschnitt wird daher das entwickelte Risikoprofil und dessen schrittweise Erstellung näher betrachtet.

3.3 Risikoprofil

Um Online-Risiken wirksam zu mindern und regulatorische Rahmenbedingungen wie den DSA einzuhalten, müssen Online-Plattformen ein umfassendes Verständnis ihres Risikoprofils entwickeln. Dieser grundlegende Schritt des ISD-Risikobewertungsrahmens ermöglicht es VLOPs und VLOSEs, potenzielle Risiken zu identifizieren, die sich aus der Gestaltung und Struktur ihrer Plattform, ihrem Betrieb und ihrer Nutzung ergeben (Art. 34 Abs. 1 DSA). Durch das Aufzeigen spezifischer Schwachstellen können die Online-Plattformen maßgeschneiderte

Maßnahmen zur Minderung dieser Risiken heranziehen. Dieser Ansatz, der sich an den Ofcom-Leitfaden zur Risikobewertung anlehnt (2023), gliedert sich in zwei zentrale Schritte:

a) **Identifizierung von Risikofaktoren:** Online-Plattformen beantworten zunächst eine Reihe detaillierter Leitfragen, um zu analysieren, wie ihre Dienste, Funktionalitäten und Designentscheidungen potenzielle Risiken begünstigen könnten. Dabei werden kritische Aspekte wie Mechanismen der Nutzerbindung, algorithmische Entscheidungsfindung und Verbreitungsmuster von Inhalten untersucht.

b) **Erstellung eines Risikoprofils:** Auf Basis des Verständnisses ihrer Risikofaktoren nutzen Online-Plattformen evidenzbasierte Erkenntnisse, um diese Faktoren systemischen Risiken zuzuordnen. Dieser Schritt verdeutlicht die Wechselwirkungen zwischen plattformspezifischen Eigenschaften und umfassenderen Schadensmustern und unterstützt eine fundierte, datengestützte Entscheidungsfindung.

c) **Risikofaktoren**

Die vom ISD erstellten Leitfragen basieren auf Art. 34 Abs. 2 DSA, in dem ein klarer Auftrag zur Bewertung systemischer Risiken durch die VLOPs oder VLOSEs erteilt wird. Im Zentrum stehen fünf Kernaspekte:

- Die Gestaltung von Empfehlungssystemen und anderer relevanter algorithmischer Systeme;
- Die Systeme zur Moderation von Inhalten;
- Die anwendbaren allgemeinen Geschäftsbedingungen und ihre Durchsetzung;
- Systeme zur Auswahl und Anzeige von Werbung;
- Die datenbezogene Praxis des Anbieters.

Aufbauend auf diesen Anforderungen hat das ISD den Umfang der Risikobewertung um drei zusätzliche Dimensionen erweitert, die für ein ganzheitliches Verständnis des Risikoprofils einer Plattform entscheidend sind: Die Art des Dienstes, dessen Größe und Nutzerbasis sowie sein Geschäftsmodell. Das ISD hat für jede dieser Dimensionen Leitfragen entwickelt, damit Online-Plattformen ihr Risikoprofil besser verstehen können. Hinsichtlich der Geschäftsbedingungen und ihrer Durchsetzung (Art. 34 Abs. 2c) DSA) hat das ISD diese Kategorie in zwei Bereiche unterteilt: einer fokussiert sich auf die Aspekte der Geschäftsbedingungen, der andere ausschließlich auf deren Durchsetzung. Jede Kategorie wird dabei klar definiert:

Art der Dienste: In dieser Kategorie wird analysiert, welche spezifischen Dienste ein VLOP oder VLOSE anbietet und wie diese potenziellen Risiken beeinflussen. Dieses Verständnis ermöglicht die Identifikation von Risiken, die mit den Funktionen und Nutzerinteraktionen der Plattform verbunden sind. Jede Dienstkategorie hat ihre eigene Dynamik, die entweder zur Entstehung oder zur Minderung von Online-Schäden beitragen kann. Die Art des Dienstes beeinflusst direkt die Art der Interaktionen, Funktionen und Risiken, die aus der Gestaltung und Nutzung der Plattform resultieren. Unterschiedliche Dienste wie Social-Media-Plattformen, Video-Sharing-Seiten oder Online-Marktplätze stellen besondere Herausforderungen dar, wie z.B. die Verbreitung schädlicher Inhalte, die Ermöglichung von Betrug oder die Bereitstellung rechtswidriger Materialien. Die Identifizierung der Art des Dienstes ermöglicht es Online-Plattformen, spezifische Schwachstellen zu ermitteln und Strategien zur Risikominderung entsprechend anzupassen.

Größe und Nutzerbasis: Die Größe wird auf der Grundlage der durchschnittlich monatlich aktiven Nutzer*innen des Dienstes in der EU bewertet. Die Nutzerbasis umfasst dabei nicht nur die Anzahl der Nutzer*innen, sondern auch deren demografische Zusammensetzung. Die Bewertung von Größe und Nutzerbasis hilft, das Ausmaß potenzieller Schäden einzuschätzen und die am stärksten gefährdeten Nutzergruppen zu identifizieren. Auch wenn alle VLOPs und VLOSEs die von der DSA definierte Schwelle von durchschnittlich mindestens 45 Millionen monatlich aktiven Nutzer*innen in der EU erreichen, kann deren demografische Zusammensetzung erheblich variieren. Diese Unterschiede wirken sich direkt auf die Art und das Ausmaß der Risiken aus, die mit jeder Plattform verbunden sind. Online-Plattformen mit einem höheren Anteil an schutzbedürftigen Nutzer*innen – etwa Minderjährigen oder Personen mit bestimmten geschützten Merkmalen (Erwägungsgrund 94 DSA) – sind beispielsweise einem erhöhten Risiko für Online-Schäden ausgesetzt. Obwohl sich diese Bewertung primär auf VLOPs und VLOSEs konzentriert, ist die

Analyse von Größe und Nutzerbasis auch für kleinere Online-Plattformen relevant, insbesondere wenn der Rechtsrahmen erweitert wird. Eine solche Analyse ermöglicht eine verhältnismäßige Bewertung und die Umsetzung von Schutzmaßnahmen, die auf Online-Plattformen mit unterschiedlichem Einfluss und variierender demografischer Dynamik zugeschnitten sind. Dadurch wird ein integrativer und flexiblerer Ansatz zur Risikominderung gefördert.

Geschäftsmodell: Diese Kategorie untersucht, wie eine VLOP oder VLOSE Einnahmen generiert und Wachstum anstrebt. Dazu gehören die Analyse der Einnahmequellen (z.B. Werbung, Abonnements oder In-App-Käufe) sowie Strategien zur Steigerung des Nutzerengagements und zur Erweiterung der Nutzerbasis. Ein zentraler Aspekt ist dabei die Bewertung, wie Designentscheidungen zur Optimierung von Einnahmen oder Wachstum unbeabsichtigt rechtswidrige oder schädliche Aktivitäten auf der Plattform fördern können. Das Verständnis des Geschäftsmodells ist entscheidend, da solche Designentscheidungen, die auf Einnahmen oder Wachstum abzielen, ungewollt Risiken wie Hassrede, Betrug oder Ausbeutung verstärken können. Beispielsweise kann die Bevorzugung von Inhalten, die das Engagement der Nutzer*innen erhöhen, unbeabsichtigt die Verbreitung schädlicher oder rechtswidriger Inhalte begünstigen. Ebenso können Funktionen wie gezielte Werbung oder das reibungslose Teilen von Inhalten durch Akteur*innen ausgenutzt werden, um schädliche Inhalte zu verbreiten oder Betrug zu begehen. Die Bewertung dieser Dimension ermöglicht es Online-Plattformen, potenzielle Schwachstellen zu identifizieren und gezielte Schutzmaßnahmen zu ergreifen, ohne dabei ihre Geschäftsziele zu gefährden.

Gestaltung von Empfehlungssystemen und anderen relevanten algorithmischen Systemen (Art. 34 Abs. 2a) DSA): Diese Kategorie untersucht, wie Online-Plattformen Algorithmen nutzen, um Nutzer*innen Inhalte, Produkte oder Interaktionen vorzuschlagen. Dabei werden die Mechanismen analysiert, die den Empfehlungen zugrunde liegen, ihre Anpassung an das Verhalten und die Vorlieben der Nutzer*innen sowie die integrierten Funktionen für Transparenz und Engagement. Dies ist besonders relevant, da Empfehlungssysteme nicht nur die Nutzererfahrungen prägen, sondern auch die Sichtbarkeit von Inhalten beeinflussen und direkt zur Verbreitung von rechtswidrigem oder schädlichem Material beitragen können. Durch die Analyse von Algorithmen, Feedback-Signalen und Transparenzoptionen können Online-Plattformen Schwachstellen identifizieren und gezielt beheben, um unbeabsichtigten Schäden durch ihre Systeme entgegenzuwirken.

Systeme zur Moderation von Inhalten (Art. 34 Abs. 2b) DSA): Diese Kategorie umfasst die Instrumente, Strategien und Maßnahmen, die Online-Plattformen nutzen, um schädliche oder rechtswidrige Inhalte und Verhaltensweisen zu erkennen, zu verwalten und darauf zu reagieren. Dazu gehören Mechanismen zur Regulierung von nutzergenerierten Inhalten und Nutzerkonten sowie zur Steuerung von Sichtbarkeit, Monetarisierung und proaktiven Erkennungstechnologien, um Schäden zu verhindern. Darüber hinaus umfasst diese Dimension die in Art. 16 DSA beschriebenen Melde- und Abhilfeverfahren, die sicherstellen, dass Nutzer*innen und Betroffene potenziell schädliche oder rechtswidrige Inhalte wirksam melden können. Ebenso umfasst sie die in Art. 20 DSA geforderten internen Beschwerdemanagementsysteme, die den Nutzer*innen ermöglichen, Entscheidungen der Online-Plattformen auf faire und transparente Weise anzufechten. Das Verständnis dieser Dimension ist entscheidend, da sie direkt bestimmt, wie Online-Plattformen schädliche oder rechtswidrige Inhalte und Verhaltensweisen erkennen, behandeln und minimieren und gleichzeitig die Rechte der Nutzer*innen wahren.

Geschäftsbedingungen (Art. 34 Abs. 2c) DSA): In dieser Kategorie geht es darum, wie die Regeln, Richtlinien und Funktionalitäten einer Plattform die Interaktionen der Nutzer*innen gestalten und die Risiken beeinflussen (Art. 14 DSA). Diese Dimension ist wichtig, um zu verstehen, ob die Dienste der Plattform für ihre Nutzer*innen, insbesondere für gefährdete Gruppen, zugänglich, sicher und angemessen reguliert sind. Untersucht wird auch die Art der Inhaltserstellung, der Austauschfunktionen, der Kommunikationsfunktionen und der Funktionen, die die Nutzer*innen gefährden oder die Wahrscheinlichkeit eines schädlichen Verhaltens erhöhen könnten. Die Gestaltung und Anwendung dieser Systeme ist von zentraler Bedeutung für die Abschwächung systemischer Risiken, insbesondere für gefährdete Nutzer*innen und sensible Inhaltskategorien (Erwägungsgründe 45, 48 DSA). VLOPs und VLOSEs müssen eine robuste Durchsetzung mit Transparenz, Fairness und Verhältnismäßigkeit in Einklang bringen, um das Vertrauen aufrechtzuerhal-

ten, rechtliche Standards einzuhalten und gleichzeitig ein sicheres Online-Umfeld zu fördern.

Durchsetzung der Geschäftsbedingungen (Art. 34 Abs. 2c) DSA): Diese Kategorie konzentriert sich darauf, wie Online-Plattformen ihre Richtlinien umsetzen und aufrechterhalten, um ein sicheres und reguliertes Umfeld zu schaffen. Untersucht werden Mechanismen für die Verwaltung der Nutzeridentität, die Kontosicherheit und Maßnahmen zur Einhaltung der Vorschriften, um gefährdete Nutzer*innen wie Minderjährige vor Schaden zu bewahren. Diese Kategorie ist eine Fortsetzung der Kategorie »Allgemeine Geschäftsbedingungen«, da der Aspekt der Durchsetzung sicherstellt, dass die erklärten Richtlinien der Plattform vorhanden sind und in der Praxis effektiv angewandt werden, um die Risiken zu mindern.

Systeme zur Auswahl und Anzeige von Werbung (Art. 34 Abs. 2d) DSA): In dieser Kategorie geht es darum, wie Online-Plattformen Werbesysteme implementieren, einschließlich gezielter und politischer Werbung und nutzergesteuerter Monetarisierung wie Influencer-Promotion (Erwägungsgründe 26 und 39 DSA). Werbesysteme beeinflussen die Plattformdynamik, indem sie das Engagement und die Einnahmen steigern, aber sie können auch Risiken wie Fehlinformationen, Betrug oder unethisches Targeting verstärken. So kann gezielte Werbung sensible Nutzerdaten ausnutzen oder schädliche Inhalte für bestimmte Bevölkerungsgruppen bewerben. Politische Werbung birgt das Risiko der Manipulation und Beeinflussung, während nutzergenerierte Werbung die traditionellen Schutzmaßnahmen umgehen kann. Die Bewertung dieser Systeme verdeutlicht ihre Rolle bei der Gestaltung des Nutzererlebnisses und der Bewältigung potenzieller Schäden.

Datenbezogene Praktiken (Art. 34 Abs. 2e) DSA): In dieser Kategorie wird untersucht, wie Online-Plattformen mit personenbezogenen Daten umgehen, insbesondere mit sensiblen Informationen wie ethnische Herkunft, politische Meinungen, religiöse Überzeugungen, Gesundheitsdaten oder sexuelle Orientierung. Diese Praktiken sind von entscheidender Bedeutung, da ein unsachgemäßer Umgang mit Daten zu Verletzungen der Privatsphäre, diskriminierender Ansprache oder Ausbeutung gefährdeter Nutzer*innen führen kann. Wenn man versteht, wie Daten gesammelt, gespeichert und verwendet werden, kann man die Risiken in Bezug auf den Schutz der Privatsphäre der Nutzenden, die Datensicherheit und die Einhaltung von Vorschriften wie die Datenschutz-Grundverordnung (DSGVO) besser einschätzen. Es zeigt auch auf, wie Datenpraktiken zu systemischen Risiken beitragen können, wie z.B. die Ermöglichung von schädlichem Verhalten durch Profiling oder unangemessenes Targeting.

Fragenkatalog zu Risikofaktoren

Risikofaktoren	Leitfragen	Optionen
Art der Dienste	Welche Dienste werden angeboten?	a. Social-Media-Plattform b. Video-Sharing-Plattform c. Erwachsenen-Plattform d. Online-Marktplatz e. App-Store f. Reise- und Unterkunftsplattform g. Online-Enzyklopädie h. Suchmaschine
Größe und Nutzerbasis	Wie viele durchschnittlich monatlich aktive Nutzer*innen in der EU hat der Dienst?	Kontinuierliche Messung
	Welche demografischen Merkmale charakterisieren die Nutzer*innen?	a. Alter (z.B. Kinder, Jugendliche, Erwachsene, Senioren) b. Geschlecht (z.B. überwiegend Männer, Frauen ohne Angabe und divers oder ausgewogen) c. Region (z.B. Herkunft aus spezifischen Mitgliedsstaaten, städtischen, ländlichen Regionen)
Geschäftsmodell	Verwendet der Dienst eine der folgenden Gestaltungsmöglichkeiten, um Einnahmen zu erzielen?	a. Werbung b. Abonnements c. In-App-Käufe
Gestaltung von Empfehlungssystemen und anderen relevanten algorithmischen Systemen	Folgende Fragen müssen für jedes Produkt (z.B. Hauptseitenempfehlungen; Suchempfehlungen; Feed Empfehlungen) beantwortet werden:	
	Enthält der Dienst Empfehlungssysteme (z.B. Feed, Suche)?	a. Inhaltsbasierte Filterung b. Kollaborative Filterung
	Berücksichtigt der Dienst Feedback-Signale? (Thorburn, 2023)	a. Implizit b. Explizit c. Weder noch
	Erzeugt der Dienst personalisierte oder nicht-personalisierte Empfehlungen?	a. Personalisiert b. Nicht-personalisiert c. Weder noch
	Beruht der Dienst bei seinen Empfehlungen von Inhalten, Produkten oder Dienstleistungen auf einer Strategie der Ausbeutung oder der Erkundung? Wenn ja, sind die Empfehlungen:	a. Diversifiziert (Erkundung) b. Personalisierend (Ausbeutung) c. Begrenzt (Ausbeutung) d. Entdeckungsorientiert (Erkundung) e. Engagementmaximierend (Ausbeutung)

Risikofaktoren	Leitfragen	Optionen
	Bietet der Dienst die folgenden benutzerspezifischen Transparenzoptionen? (Bengani et al., 2022)	a. Erklärungen b. Abgeleitete Interessen und Attribute
	Bietet der Dienst die folgenden systemspezifischen Transparenzoptionen an? (Bengani et al., 2022)	a. Priorisierte Parameter und Signale b. Inhaltsspezifische Ranking-Entscheidungen und Interventionen c. Transparenz-Berichte d. Änderungsprotokoll zum Produktdesign
	Verfügt der Dienst über spezielle UI-Funktionen, die die Benutzerfreundlichkeit erhöhen sollen?	a. Unendliches Scrollen oder automatisches Abspielen von Videos b. Push-Benachrichtigungen c. Verfolgung der Bildschirmzeit
Systeme zur Moderation von Inhalten	Enthält der Dienst die folgenden Maßnahmen zur Inhaltemoderation (Goldman, 2021)?	a. Entfernung von Inhalten b. Vorübergehende Inhaltssperrung c. Verlagerung von Inhalten d. Redigieren von Inhalten e. Interstitielle Warnung f. Deaktivierung der Kommentarfunktion g. Hinzufügen von Kennzeichnungen h. Gegenrede hinzufügen
	Enthält der Dienst die folgenden Maßnahmen zum Umgang mit Konten (Goldman, 2021)?	a. Kündigung des Kontos b. Sperrung des Kontos c. Aussetzung von Posting-Rechten d. Entzug von Glaubwürdigkeitsplaketten e. Reduzierte Dienstebenen f. Bloßstellung von Fehlverhalten
	Enthält der Dienst die folgenden Maßnahmen zur Verringerung der Sichtbarkeit (Goldman, 2021)?	a. Shadowban b. Aus dem externen Suchindex entfernen c. Herabstufung der internen Suchsichtbarkeit d. Kein Auto-Suggest e. Keine/reduzierte interne Werbung f. Reduzierte Viralität g. Altersbeschränkung h. Anzeige nur für eingeloggte Nutzende
	Beinhaltet der Dienst Maßnahmen zur Einschränkung der Monetarisierung?	a. Verfall der aufgelaufenen Verdienste b. Beendigung des zukünftigen Verdienstes c. Aussetzung künftiger Einkünfte d. Geldstrafe/Auferlegung von Schadensersatz

Risikofaktoren	Leitfragen	Optionen
	Enthält der Dienst die folgende andere Moderationsmaßnahmen?	a. Strikes/Warnungen zuweisen b. Outing/Enttarnung c. Meldung an die Strafverfolgungsbehörden d. Nutzer/Inhalt auf Blockliste setzen e. Gemeinschaftsdienst f. Wiedergutmachungsmaßnahmen
	Verwendet der Dienst eines der folgenden Modelle/Tools, um rechtswidrige und schädliche Inhalte zu erkennen (Gorwa et al., 2020)?	a. Algorithmen zum Abgleich/Hashing von Inhalten (wie GIFTC oder PhotoDNA) b. Vorhersagealgorithmen (z.B. Perspective API)
	Sind menschliche Moderator*innen an den Entscheidungen zur Inhaltsmoderation beteiligt (Rieder & Skop, 2021)? Wenn ja, wie?	a. Die Moderation von Inhalten beruht ausschließlich auf menschlichen Moderator*innen b. Algorithmische Systeme unterstützen die menschlichen Moderator*innen durch c. Empfehlungen an menschliche Moderator*innen geben
Allgemeine Geschäftsbedingungen	Erlaubt der Dienst gefährdeten Verbraucher*innen den Zugang zu seinen Diensten (Sajn, 2021)?	a. Minderjährige b. Andere Gruppen von schutzbedürftigen Verbrauchern
	Unterstützt der Dienst Funktionen zur Erstellung und Freigabe von Inhalten, mit denen Nutzer*innen Medien erstellen, freigeben oder hochladen können?	a. Textbeiträge (entweder offene oder geschlossene Kanäle) b. Bilder oder Videos (entweder offene oder geschlossene Kanäle) c. Audioaufnahme und -freigabe d. Veröffentlichen oder Senden von Standortinformationen e. Wiederveröffentlichen und Weiterleiten von Inhalten f. Erneutes Posten oder Teilen auf anderen Online-Plattformen g. Cross-Posting zwischen verknüpften Konten h. Veröffentlichen oder Senden von Standortinformationen
	Erlaubt der Dienst Inhalte für Erwachsene?	a. Zugriff auf solche Inhalte b. Freigabe solcher Inhalte c. Erstellung solcher Inhalte
	Verfügt der Dienst über eine der folgenden Funktionen, die es den Benutzenden ermöglichen, miteinander zu kommunizieren?	a. Livestreaming (entweder offene oder geschlossene Kanäle) b. Direktnachrichten (einschließlich ephemerer Direktnachrichten) c. Verschlüsselte Nachrichten d. Kommentierung von Inhalten e. Veröffentlichen oder Senden von Bildern oder Videos (entweder offene oder geschlossene Kanäle)
	Erlaubt der Dienst den Nutzer*innen, Waren und Dienstleistungen zum Verkauf anzubieten?	a. Ja b. Nein

Risikofaktoren	Leitfragen	Optionen
Durchsetzung der allgemeinen Geschäftsbedingungen	Verfügt der Dienst über eine der folgenden Funktionen, die sich darauf beziehen, wie sich Nutzer*innen untereinander identifizieren?	a. Nutzer*innen können identifizierende Informationen über ein Nutzerprofil anzeigen, das von anderen eingesehen werden kann (z.B. Bilder, Nutzernamen, Alter) b. Nutzer*innen können Inhalte anonym teilen (z.B. anonyme Profile oder Zugang ohne Konto)
	Bietet der Dienst eine der folgenden Sicherheitsmaßnahmen?	a. Überprüfte Konten b. Zwei-Faktor-Authentifizierung oder Login-Warnungen
Systeme zur Auswahl und Anzeige von Werbung	Bietet der Dienst Werbe- und Monetarisierungs-Möglichkeiten?	a. Ja b. Nein
	Erlaubt der Dienst eine der folgenden Funktionen in Bezug auf Werbung?	a. Gezielte Werbung b. Politische Werbung c. Influencer-/Nutzerwerbung
Datenbezogene Praktiken	Arbeitet der Dienst mit personenbezogenen Daten nach der DSGVO?	a. Daten, aus denen die ethnische Herkunft hervorgeht b. Politische Meinungen c. Religiöse oder philosophische Überzeugungen d. Gewerkschafts-Zugehörigkeit e. Genetische Daten f. Biometrische Daten g. Gesundheitsdaten h. Daten über das Sexualleben oder die sexuelle Ausrichtung einer natürlichen Person

Tabelle 4: Leitfragen für Online-Plattformen zum Verständnis ihrer Risikofaktoren

Die Beantwortung der Fragen im oben angeführten Fragenkatalog zu Risikofaktoren stellt lediglich den ersten Schritt der Analyse durch Online-Plattformen dar, um potenzielle Risiken zu identifizieren, denen sie ausgesetzt sind. Um jedoch zu verstehen, welche spezifischen Risiken aus den identifizierten Risikofaktoren resultieren, ist eine detaillierte Analyse des Risikoprofils erforderlich. Dieses wird im nächsten Abschnitt anhand eines konkreten Risikofaktors – der Art der Dienste erläutert. Das Risikoprofil dient als entscheidende Brücke zur Übersetzung von allgemeinen Risikofaktoren in konkrete Kernrisiken.

b) Risikoprofil

Unter Berücksichtigung der neun Risikofaktoren lassen sich für jede der oben genannten Leitfragen systematisch potenzielle Risiken identifizieren, denen eine digitale Plattform aufgrund ihrer Konzeption, ihres Serviceangebots und ihrer betrieblichen Praktiken ausgesetzt sein könnte. Dieser Ansatz hilft Online-Plattformen dabei, besser zu verstehen, wie ihre spezifischen Funktionalitäten zu bestimmten Schwachstellen und systemischen Risiken beitragen. Auf dieser Grundlage können gezielte Strategien zur Risikominderung entwickelt und umgesetzt werden.

Durch die Prüfung verfügbarer Forschungsergebnisse und regulatorischer Richtlinien kann das Risikoprofil erstellt werden. Diese dient den Online-Plattformen als Hilfsmittel, um Risiken zu identifizieren und zu bewerten, die mit ihrer Struktur und ihren Diensten verbunden sind. Die Matrix gleicht Risikofaktoren mit potenziellen Schäden ab und bietet einen klaren und umsetzbaren Überblick darüber, wo eine Plattform mit Problemen konfrontiert sein könnte, z.B. mit der Verbreitung rechtswidriger Inhalte, der Gefährdung vulnerabler Nutzer*innen oder der unzureichenden Moderation von Inhalten.

Als praktischen Ansatz hat das ISD ein vorläufiges Risikoprofil für den ersten Risikofaktor – die Art des Dienstes – entwickelt, um zu veranschaulichen, wie dieser Rahmen angewendet werden kann. Die Matrix enthält eine Spalte für den identifizierten Risikofaktor (z.B. »Art der Dienste«) und eine weitere für potenzielle Risiken, die auf spezifische Plattformfunktionen zugeschnitten sind. Dieser strukturierte Ansatz bietet eine skalierbare und reproduzierbare Methode für die Bewertung von Risiken in allen neun Kategorien und hilft Online-Plattformen, Prioritäten für Maßnahmen zu setzen, Vorschriften einzuhalten und Nutzer*innen effizienter zu schützen.

VLOPSs und VLOSEs mit relevanten Merkmalen sollten bei ihrer Risikobewertung Folgendes berücksichtigen

Social-Media-Plattformen

Mögliche systemische Risiken für Social-Media-Plattformen umfassen verschiedene Aspekte, die sich aus der Art der Online-Plattformen ergeben.

Nutzungsbezogene Risiken beinhalten Verhaltenssucht (Müller et al., 2016), wie etwa durch unendliches Scrollen, »Doomscrolling« (Satici et al., 2023) oder automatische Wiedergabe von Inhalten, sowie die Aussetzung gegenüber schädlichen Interaktionen, die die psychische Gesundheit der Nutzer*innen beeinträchtigen können.

Einige Studien legen nahe, dass sich die Nutzung von Social-Media-Plattformen negativ auf die mentale Gesundheit besonders von Minderjährigen und jungen Erwachsenen auswirken könnte (Haidt et al., o. J.). Dabei könnte eine starke Nutzung der Online-Plattformenmit schlechtem Schlaf, geringem Selbstwertgefühl (Barthorpe et al., 2020; Kelly et al., 2018; Woods & Scott, 2016), Unzufriedenheit mit dem eigenem Körper (Kelly et al., 2018), wahrgenommener sozialer Isolation (Primack, Shensa, Sidani, et al., 2017), allgemeiner psychologischer Belastung (Sampasa-Kanyinga & Lewis, 2015), sowie depressiven Stimmungen (Barthorpe et al., 2020; Hawes et al., 2020; Kelly et al., 2018; M. Liu et al., 2022; Merrill et al., 2022) sowie Angstsymptomatiken (Brailovskaia & Margraf, 2023; Hawes et al., 2020; Primack, Shensa, Escobar-Viera, et al., 2017) und Verhaltensproblemen (McNamee et al., 2021) einhergehen.

Ob und in welche Richtung hier eine Kausalität besteht, ist jedoch umstritten. Ob die Nutzung von Social-Media-Plattformen einen kausalen Effekt auf diese Symptomatiken hat oder, ob Nutzer*innen mit diesen Symptomatiken lediglich besonders häufig dazu neigen, Social-Media-Plattformen intensiv zu nutzen (Aalbers et al., 2019) spielt eine wichtige Rolle in der Bewertung. Beispielsweise könnte ein Zusammenhang zwischen bestehenden Depressionen und dem Bedürfnis nach

»sozialen Belohnungen« wie Likes, Views oder Shares bestehen (Mirea et al., 2024). Einige Studien weisen darauf hin, dass ein reziprokes Verhältnis bestehen könnte, welches sich gegenseitig verstärkt (Miljeteig & von Soest, 2022; Wang et al., 2018).

Wiederum andere Studien fanden keinen signifikanten Effekt auf das mentale Wohlbefinden (Coyne et al., 2020; Gov.uk, 2019; Kreski et al., 2021; Rozgonjuk et al., 2020) oder sogar einen positiven Effekt (Boer, 2022), welcher jedoch in unterschiedlichen Gruppen unterschiedlich ausgeprägt ist (Beyens et al., 2020; Griffioen et al., 2023). Die Frage ob Social-Media-Plattformen einen positiven oder negativen Einfluss auf das mentale Wohlbefinden hat, scheint wesentlich abhängig von vermittelnden Faktoren und der Art der Nutzung zu sein (Alberto et al., 2022; Burke & Kraut, 2016; Burnell et al., 2019; Faelens et al., 2019; Fardouly et al., 2020; Gomez et al., 2022; Hanna et al., 2017; Jarman et al., 2021; Lee, 2022; Samra et al., 2022; Schettino et al., 2023; Spitzer et al., 2023; Steers et al., 2014). Diese vermittelnden Faktoren und Nutzungsweisen sollten daher besonders berücksichtigt werden in der Risikobewertung und minderung, um vulnerable Gruppen besonders zu schützen. Dabei kann das Plattformdesign auch eine entscheidende Rolle spielen. Beispielsweise scheint Instagram die Nutzer*innen mehr zu Vergleichen anzuregen als z.B. Facebook (Engeln et al., 2020). Auch weisen einige Studien darauf hin, dass gerade passive Nutzung einen negativen Effekt auf das mentale Wohlbefinden haben kann (Ozimek & Bierhoff, 2020; Verduyn et al., 2015).

Inhaltsbezogene Risiken umfassen die Verstärkung und Verbreitung von rechtswidrigen Inhalten, von legaler Hassrede und von Material über sexuellen Kindesmissbrauch (engl. Child Sexual Abuse Material, CSAM).

Des Weiteren kann zum Beispiel der Kontakt mit Inhalten zu Selbstverletzung auf Social-Media-Plattformen zu Selbstmordgedanken, Selbstverletzung und emotionalen Störungen beitragen (Arendt et al., 2019). Außerdem kann der Kontakt mit manipulierten Fotos auf Social Media-Plattformen könnte einen negativen Effekt auf die eigene Zufriedenheit mit dem eigenen Körper haben (Kleemans et al., 2018).

Verhaltensbezogene Risiken ergeben sich durch die Ausnutzung der Online-Plattformen durch Akteur*innen, beispielsweise für Belästigung, Desinformationskampagnen oder gezielte Angriffe auf gefährdete Nutzergruppen.

Forschung zeigt, dass Social-Media-Plattformen besonders häufig für »Grooming« missbraucht werden können. Hierbei gilt es sowohl peer-to-peer grooming (Ashurst & McAlinden, 2015) als auch pädophile Online-Aktivitäten zum Kontakt Minderjähriger (Cano et al., 2014) als Risiken zu berücksichtigen. Hierbei werden oft Strategien angewandt, um die Filtermechanismen der Social-Media-Plattformen zu umgehen (Lykousas & Patsakis, 2021).

Video-Sharing-Plattformen

Mögliche systemische Risiken für Video-Sharing-Plattformen umfassen verschiedene Gefahren, die aus der Nutzung und dem Umgang mit den angebotenen Online-Plattformen entstehen.

Nutzungsbezogene Risiken beinhalten die Förderung einer übermäßigen Bildschirmnutzung durch Empfehlungssysteme oder automatische Wiedergabefunktionen, was Verhaltenssucht begünstigen kann.

Su et al. zeigen, wie empfohlene Videos auf TikTok, basierend auf personalisierten Empfehlungssystemen, die Gehirnaktivitäten junger Nutzer*innen beeinflussen und zu unerwünschten psychischen Symptomen (z.B. geringere Fähigkeit zur Selbstkontrolle) und damit zu schädlichen Nutzungsformen führen könnten (2021).

Inhaltsbezogene Risiken ergeben sich aus der Bereitstellung oder Empfehlung rechtswidriger Inhalte.

Beispielsweise zeigte eine Untersuchung des slowakischen DSCs, dass schädliche und potenziell rechtwidrige Inhalte nach dem Terroranschlag in Bratislava leicht erreichbar waren (Council for Media Services & Trust Lab, 2023). Daneben besteht die Gefahr, dass Nutzer*innen durch Interaktionen mit bestimmten Inhalten immer mehr ähnliche, extremere oder radikalere Empfehlungen erhalten (Hussein et al., 2020; Reed et al., 2019; Whittaker et al., 2021). Diese Dynamiken werden in der Forschung oft als »Rabbit Holes« oder »Filterblasen« bezeichnet, sind in ihrer Ausprägung jedoch äußerst umstritten (Brown et al., 2022; Chen et

al., 2023; Haroon et al., 2023; Ledwich & Zaitsev, 2020; Ribeiro et al., 2023a).

Verhaltensbezogene Risiken umfassen die Ausnutzung von Video-Sharing-Plattformen durch Algorithmen, die gezielt Desinformation oder andere schädliche Inhalte verbreiten.

Zum Beispiel können Akteur*innen YouTube durch eine neue Art von betrügerischer Werbemethode ausnutzen, die als »Social Scam Bots« (SSBs) bekannt ist (Na et al., 2023). Diese Bots imitieren echtes Nutzerverhalten, indem sie Kommentare posten und mit anderen Nutzer*innen interagieren, wobei sie oft als Kommentare mit Top-Bewertung erscheinen. Von den analysierten YouTube-Kommentare identifizierten die Forscher*innen 1.134 SSBs und deckten 72 Betrugskampagnen auf, die 31% der Videos betrafen. Die Ergebnisse deuten darauf hin, dass SSBs raffinierter sind als herkömmliche Bots, da sie effektiv auf bestimmte Zielgruppen abzielen, indem sie ihre Betrugsversuche auf relevante Videoinhalte abstimmen und so das Empfehlungssystem von YouTube ausnutzen. Zusätzliche Strategien wie »Self-Engagememt« Bots helfen die Sichtbarkeit der SSBs Kommentare zu steigern, indem Videoaufrufzahlen künstlich gesteigert werden in dem die Bots automatisch Kanäle ansehen, liken, kommentieren und abonnieren. Dies führt zu einer wahrgenommenen Popularität und Glaubwürdigkeit, und verzerrt die Analysewerte von YouTube, was gegebenenfalls Gegenmaßnahmen schwerer machen.

Erwachsenen-Plattformen

Mögliche systemische Risiken für Erwachsenen-Plattformen umfassen spezifische Herausforderungen, die aus der Natur der angebotenen Inhalte und deren Nutzung entstehen.

Nutzungsbezogene Risiken beinhalten die Aussetzung gegenüber grafischem oder explizitem Material, das das psychische Wohlbefinden der Nutzer*innen beeinträchtigen kann.

Es stellt zum Beispiel ein Risiko dar, wenn Minderjährige Erwachsenen-Dienste nutzen und somit pornografischen Inhalten ausgesetzt werden. Es besteht ein Risiko, wenn Minderjährige Erwachsenen-Plattformen nutzen und pornografischen Inhalten ausgesetzt werden. Eine Studie der britischen Kinderkommission zeigt, dass das Durchschnittsalter für den ersten Kontakt mit Online-Pornografie bei 13 Jahren liegt, und manche Kinder bereits ab 9 Jahren solchen Inhalten begegnet sind. Zudem sind 79 % der Jugendlichen vor ihrem 18. Lebensjahr mit Gewaltpornografie in Kontakt gekommen (Children's Commissioner, 2023). Weitere Forschungserkenntnisse zeigen, dass das Ansehen von Pornografie teilweise zu Depressionen, Angstzuständen und Stress führen kann (Bernstein et al., 2023; Borgogna et al., 2018; Camilleri et al., 2021).

Inhaltsbezogene Risiken ergeben sich aus der Bereitstellung oder Weitergabe rechtswidriger Inhalte für Erwachsene, einschließlich nicht einvernehmlicher intimer Bildaufnahmen oder CSAM.

In einer der größten Studien über Online-Pornografie wurde festgestellt, dass einer aus acht Beiträgen auf den Startseiten von Erwachsenen-Plattformen sexuell gewalttätige Inhalte, einschließlich bildbasierten sexuellen Missbrauchs, beschreibt (Vera-Gray et al., 2021). Ebenso werden Erstnutzer*innen direkt solche gewalttätigen Inhalte vorgeschlagen, was darauf hinweist, dass Erwachsenen-Dienste, einschließlich deren Empfehlungssysteme, solche Inhalte aktiv anbietet und Nutzer*innen vorschlägt (McGlynn, 2021). Diese bereitgestellten Inhalte stellen ein Risiko für alle Nutzergruppen da, insbesondere minderjährige Erstnutzer*innen.

Verhaltensbezogene Risiken entstehen durch die Ausnutzung der Plattform durch Akteur*innen, z.B. für rechtswidrigen Handel, Missbrauch oder Erpressung.

Forschungen sowie Fallrecht verdeutlichen, wie Täter*innen die Upload-Funktionen nutzen, um nicht einvernehmliche intime (synthetische) Bildmaterialien zu teilen. Die Beweggründe dahinter können zum Beispiel Rache (»revenge porn«) (European Institute for Gender Equality, 2024a, 2024b), Erpressung (»sextortion«) (Edwards & Hollely, 2023) oder die sexuelle Befriedigung der*des Täter*in sein (Henry & Beard, 2024). Zusätzlich werden in manchen Fällen Beschreibungsfunktionen oder Kommentarspalten genutzt, um personenbezogene Daten der dargestellten Person ohne Einwilligung zu teilen (»Doxing«).

Online- Marktplätze

Mögliche systemische Risiken von Online-Marktplätzen ergeben sich aus der Art ihrer Nutzung und den angebotenen Inhalten.

Nutzungsbezogene Risiken umfassen die Gefährdung durch gefälschte Angebote oder unsichere Produkte, die die Sicherheit und das Vertrauen der Nutzer*innen beeinträchtigen können.

Eine Studie von Lamis et al. (2022) deutet drauf hin, wie die Nutzung von Online-Marktplätzen, Nutzer*innen für Impulskäufe anfällig machen kann. Die Ergebnisse deuten darauf hin, dass Faktoren wie Erregung und Freude einen erheblichen Einfluss auf Impulskäufe haben können. Insbesondere Aspekte wie begrenzte Verfügbarkeit und Zeit schaffen ein Gefühl der Dringlichkeit und erhöhen das Erregungsniveau, während Elemente wie Information, Unterhaltung und wahrgenommene wirtschaftliche Vorteile sowohl zur Erregung als auch zur Freude beitragen. Die Studie unterstreicht auch, dass die Art und Weise, wie Nutzer*innen »Flash-Sales« wahrnehmen ihre Einstellung zu Impulskäufen beeinflusst. Diese kognitiven Neigungen machen Nutzer*innen besonders anfällig für Online-Marktplätze, da der Kontext der Verkäufe durch die erzeugte Aufregung rationale Entscheidungsprozesse außer Kraft setzen kann, was zu impulsivem Verhalten führt, das von Dringlichkeit und emotionaler Reaktion bestimmt wird.

Inhaltsbezogene Risiken entstehen durch den Verkauf rechtswidriger Waren, gefälschter Artikel oder unsicherer Produkte, die gesetzliche und ethische Standards verletzen.

Dies ist beispielsweise auch von der deutschen Verbraucherzentrale erkannt worden, welche davor warnte bei Temu einzukaufen, da bei elektronischen Produkten oftmals die CE-Zertifizierung fehlen würde (Verbraucherzentrale, 2024). Wenn die Inhalte, die für solche Produkte, werben es nicht für Nutzer*innen erkennbar macht, dass solche Zertifizierungen fehlen, stellt das einen vermutlich täuschenden Inhalt und im Falle eines Kaufs des irreführenden Produkts, ein Sicherheitsrisiko, dar.

Verhaltensbezogene Risiken beinhalten die Ausnutzung der Plattform durch Betrüger*innen, um Nutzer*innen zu täuschen oder finanziellen Betrug zu begehen.

Es besteht das Risiko, dass Online-Plattformen gezielt durch den Einsatz sogenannter »Dark Patterns« vulnerable Nutzergruppen ausgebeutet werden (European Parliament, 2021). Der DSA definiert Dark Patterns als »Praktiken, mit der darauf abgezielt oder tatsächlich erreicht wird, dass die Fähigkeit der Nutzer*innen, eine autonome und informierte Auswahl oder Entscheidung zu treffen, maßgeblich verzerrt oder beeinträchtigt wird« (Erwägungsgrund 67). Solche Muster kommen trotz Verbot im DSA großflächig zum Einsatz, wie sich nicht zuletzt am Fall Temu zeigte (BEUC, 2024), bei dem die Europäische Kommission ein laufendes Verfahren einleitete (Europäische Kommission, 2024e). Dark Patterns beruhen dabei häufig auf der Ausübung von Druck, operativem Zwang und gezielten Hindernissen, die zusätzlichen Aufwand auf der Seite der Nutzer*innen nötig machen, heimliche Änderungen wie versteckte Kosten oder gezielte Irreführungen (Martini et al., 2021, S. 52).

App-Stores

Mögliche systemische Risiken für App-Stores ergeben sich aus der Art der angebotenen Anwendungen und deren Nutzung.

Nutzungsbezogene Risiken umfassen die Gefährdung durch schädliche Apps, die Daten missbräuchlich verwenden oder Geräte beschädigen können.

Durch sogenannte »Freemium Games«, welche sich in den Appstores besonders gut vermarkten lassen (C. Z. Liu et al., 2014) können vulnerable Nutzer*innen besonders hohen finanziellen Risiken ausgesetzt sein. Die Einnahmen von den Freemium Games, basieren wesentlich auf den Ausgaben eines ein kleinen Prozentsatz der Nutzer*innen, welcher für einen Großteil der Einnahmen verantwortlich ist. Besonders Spieler*innen die eine Sucht nach dem Spiel entwickelt haben, verbringen nicht nur sehr viel Zeit in diesen Spielen, sondern geben auch besonders viel Geld aus (Dreier et al., 2017). Durch gezielte Werbung und prominente Platzierung können Appstores dazu beitragen, dass die

Aufmerksamkeit vulnerabler Nutzer*innen auf solche Apps gelenkt wird.

Inhaltsbezogene Risiken entstehen durch die Verbreitung von Apps, die rechtswidrige Inhalte enthalten.

Ein Beispiel von Apps, welche rechtswidrigen Inhalte zeigen oder gar erstellen lassen, sind so genannte »nudify« – (nudizifieren) Apps. Nudify-Apps verwenden »Deep-Learning« -Techniken, um Bilder von bekleideten Personen zu analysieren, ihre Körperformen und Posen zu erkennen und dann die Bilder digital zu verändern, um eine simulierte Nacktdarstellung der Person zu erzeugen.

Eine Studie von 2023 untersuchte 20 Apps dieser Nudify-Apps und stellte fest, dass 19 von ihnen sich speziell auf das Ausziehen von Frauen konzentrierten, wobei nur die Hälfte die Notwendigkeit einer Einwilligung anerkannte (Gibson et al., 2024). Dadurch bekommen Nutzer*innen Zugang zu nicht einvernehmlichen intime Inhalten (Gibson et al., 2024). Da diese Apps ebenso Minderjährigen zugänglich gemacht werden, vor allem, weil diese oft auf Social-Media-Plattformen beworben werden, stellt dies ein großes Risiko da, dass Kindern sexualisierten Inhalten ausgesetzt werden, oder sich (unwissend) CSAM aussetzen (Trendell, 2024).

Im Dezember 2023 ergab eine Analyse von 34 Nudify-App-Anbietern, dass diese Online-Plattformen allein im September über 24 Millionen Besucher*innen hatten, was auf eine beträchtliche Nutzerbasis für solche Anwendungen hinweist.

Verhaltensbezogene Risiken ergeben sich aus dem Missbrauch der Online-Plattform durch Akteur*innen, beispielsweise zur Verbreitung von schädlicher Software oder zur Begehung von Betrug.

Durch eine unzureichende Kontrolle oder Moderation von bereitgestellten Apps, können Apps in App-Stores gelangen, welche beispielsweise Hintertüren für Spionageaktivitäten bereitstellen, gezielt selbst Spionageaktivitäten betreiben, Betrug bei der Abrechnung von Mobilfunkrechnungen oder dem Handel (nicht existenter) Kryptowährungen betreiben »Denial of Service (DoS)« Attacken ausführen, andere schädliche Apps installieren, Phishing betreiben, die Installation von Ransomware, die Ausführung von Spam Aktivitäten (Suleman et al., 2021). Durch Akteur*innen, welche »Fake Bewertungen« (Martens & Maalej, 2019; Xie et al., 2016) für diese Apps inszenieren, können solche oder anderweitige Apps oftmals ihren Erfolg steigern.

Reise- und Unterkunftsplattformen

Mögliche systemische Risiken für Reise- und Unterkunftsplattformen ergeben sich aus der Natur der bereitgestellten Dienste und Inhalte.

Nutzungsbezogene Risiken umfassen die Aussetzung gegenüber betrügerischen oder unsicheren reisebezogenen Inhalten, die die Sicherheit und das Vertrauen der Nutzer*innen beeinträchtigen können (Sthapit & Björk, 2019) haben in ihrer Studie analysiert, was bei Nutzer*innen der Plattform Airbnb dazu führt, dass sich durch die Nutzung misstrauen entwickelt. Unter anderem wurde festgestellt, dass eine unzureichende Berteitstellung von effektiven Meldemöglichkeiten dazu führt, dass Nutzer*innen eine schlechte Erfahrung auf der Plattform machen.

Inhaltsbezogene Risiken entstehen durch gefälschte Inserate oder Bewertungen, die Nutzer*innen täuschen und zu Fehlentscheidungen führen können. Eine Studie von Schneider and Teubner, 2024 deutet darauf hin, dass Erstbewertungen auf Online-Plattformen wie Airbnb oftmals von den Netzwerken der Gastgeber*innen kommen, um deren Inserate reputabler wirken zu lassen. Inwiefern dies direkt zu Fehlentscheidungen führen kann, ist nicht ausreichend erforscht. Die Forschung von Ert et al., (2016) deutet wiederum daraufhin hin, dass die hochgeladenen Bilder Nutzer*innen mehr als Bewertungen beeinflussen. Folglich könnten gerade hochgeladene Inhalte auf Reise- und Unterkunftsplattformen eine wichtige Rolle spielen und Nutzer*innen täuschen.

Verhaltensbezogene Risiken ergeben sich durch die Ausnutzung der Plattform für Betrügereien oder rechtswidrige Aktivitäten durch betrügerische Angebote oder Manipulierungen. Ähnlich wie bei den oben beschriebenen Online-Marktplätzen, sind Dark-Pattern-Taktiken auch auf Reise- und Unterkunftsplattformen eine Methode, um Nutzer*innen zu beeinflussen, was unter anderem zu kognitiven Verzerrungen

führen kann. Hierbei werden unter Anderem Fake-Rabatte, versteckte Kosten sowie »High-Demand«- und »Low-Supply«-Benachrichtigungen eingesetzt (Kim et al., 2021), um die Nutzer*innen gezielt zu beeinflussen.

Online-Enzyklopädien

Mögliche systemische Risiken für Online-Enzyklopädien ergeben sich aus der offenen und kollaborativen Struktur ihrer Inhalte.

Nutzungsbezogene Risiken beinhalten die Verbreitung von Fehlinformationen oder Vorurteilen durch kollaborative Bearbeitungen, die das Verständnis und die Bildung der Nutzer*innen beeinträchtigen können.

Zum Beispiel können Nutzer*innen gefährdet werden, wenn sie nicht weitere externe Quellen zu den Online-Enzyklopädien heranziehen, da die Freiwilligkeit der Beiträge zu einer ungleichmäßigen Berichterstattung führen kann. Die könnte zu einem verzerrten Verständnis verschiedene Themen führen (Denning et al., 2005).

Inhaltsbezogene Risiken entstehen durch die Bereitstellung rechtswidriger oder verleumderischer Inhalte aufgrund unzureichender Moderations- und Prüfmechanismen.

Kollaborative Online-Enzyklopädien sind häufig mit schädlichen und teils rechtswidrigen Inhalten konfrontiert. Jiang and Vetter (2020) zeigen auf wie Inhalte auf Wikipedia oftmals irreführende Informationen wiedergeben, wie zum Beispiel eine voreingenommene Darstellung von marginalisierten Gruppen. Außerdem können Interessenkonflikte die Neutralität der erstellten Inhalte beeinträchtigen. Grabowski and Klein (2023) argumentieren, wie Wikipedia Inhalte wiedergibt, die die Geschichte des Holocausts deutlich verzerren.

Verhaltensbezogene Risiken umfassen den Missbrauch der Plattform zur Desinformation oder zur Durchführung von Hasskampagnen.

Außerdem können Online-Enzyklopädien genutzt werden, um sich durch gezieltes Entfernen oder Hinzufügen kritischer Informationen einen Wettbewerbsvorteil zu verschaffen. Außerdem können Desinformation über Personen, Organisationen oder Unternehmen durch Online-Enzyklopädien besonders einflussreich verbreitet werden (Miller et al., 2022).

Durch gezielte Manipulationen können ebenso politische Debatten beeinflusst werden, mit Auswirkungen auf die öffentliche Debatte sowie die Informations- und Meinungsfreiheit. Zum Beispiel kann sich eine Gruppe an Redakteur*innen besonders stark auf ein kontroverses Thema fokussieren, sobald sie einen Admin-Status erhalten. So können die Meinungen und politischen Positionen einiger weniger Administrator*innen einen wesentlichen Einfluss auf die Darstellung eines Themas auf Wikipedia haben (Das et al., 2016). Daneben sind Online-Enzyklopädien auch der Gefahr von »Online-Trollen« ausgesetzt, welche aus Langeweile, Aufmerksamkeitsbedürfnis oder Rache getrieben sind und Manipulationen als Unterhaltungszweck betrachten (Shachaf & Hara, 2010).

Suchmaschinen

Mögliche systemische Risiken für Suchmaschinen ergeben sich aus ihrer Rolle als zentraler Zugangspunkt zu Informationen.

Nutzungsbezogene Risiken umfassen die Aussetzung gegenüber schädlichen oder irreführenden Suchergebnissen, die die Entscheidungen und das Vertrauen der Nutzer*innen beeinflussen können.

Zum Beispiel können Suchmaschinen bestimmte Inhalte depriorisieren, was dazu führen kann, dass Suchanfragen zu politischen Streitfragen zu verzerrten oder einseitigen Suchergebnisse führen (Kulshrestha et al., 2017). Dies kann wiederum einen direkten negativen Effekt auf die Informations- und Meinungsfreiheit sowie Medienpluralismus haben.

Inhaltsbezogene Risiken entstehen durch das Anzeigen und Verbreitung rechtswidriger oder schädlicher Inhalte wie CSAM oder Desinformation.

Dieses Risiko kann zum Beispiel durch den Einsatz neuer Technologien wie Generative KI verschärft werden. Ashraf et al. (2024) zeigen zum Beispiel Schwachstellen von Suchmaschinen auf, die Generative KI nutzen und den Nutzer*innen Inhalte empfehlen. Die Forschungs-

ergebnisse zeigen, wie rechtswidrige Online-Apotheken ungewollt empfohlen werden. Die angezeigten Inhalte können dabei nachteilige Auswirkungen auf den Schutz der öffentlichen Gesundheit haben.

Verhaltensbezogene Risiken ergeben sich durch die Ausnutzung der Plattform durch Akteur*innen, die Suchergebnisse manipulieren oder gezielt Desinformationen verbreiten.

Eine Studie von Williams & Carley (2023) analysierte, wie Kreml-nahen Organisationen versuchen, Suchmaschinen Ergebnisse zu manipulieren, indem sie »Backlink« und Schlüsselwortnetzwerke verschiedener Think Tanks analysiert. Dabei zeigt sich, dass kremlnahe Pseudo-Think Tanks durch minderwertige Websites, die Millionen von Backlinks generieren und auf verschwörungsbezogene Schlüsselwörter abzielen, stark verstärkt werden. Trotz dieser Manipulationsbemühungen scheint der Google-Suchalgorithmus russische und Pseudo-Think Tanks abzustrafen, was sich an ihren im Vergleich zu US-amerikanischen und europäischen Think Tanks niedrigeren durchschnittlichen Ranking-Positionen ablesen lässt.

Fallstudie: Online-Bots

Das Folgende soll zeigen, wie anhand der Risikofaktoren Bots als ein systemisches Risiko identifiziert werden können. In der Fallstudie wird davon ausgegangen, dass sich die Plattform X einer akuten Präsenz von Online-Bots nicht bewusst ist oder zumindest nicht weiß, dass Bots ein systemisches Risiko darstellen. Im Folgenden wird gezeigt, wie durch die Analyse von Risikofaktoren, Bot-Präsenz als Risikofaktor herausgearbeitet werden könnte.

- Art des Dienstes: Das Design von X als Austauschplattform von Echtzeit-Information mit offenen APIs erleichtert die schnelle Verbreitung von Inhalten und macht sie anfällig für die Ausnutzung durch automatisierte Konten zur Verbreitung von Desinformationen.
- Größe und Nutzerbasis: Mit einer großen globalen Nutzerbasis bietet X Botnetzwerken die Möglichkeit, gezielt bestimmte demografische Gruppen, wie beispielsweise deutschsprachige Nutzer*innen, anzusprechen, um Desinformationskampagnen zu verbreiten.
- Geschäftsmodell: Das auf Engagement ausgerichtete Einkommensmodell von X könnte unbeabsichtigt viralen Inhalten unabhängig von ihrer Authentizität Vorrang geben. Das kann möglicherweise dazu führen, dass von Bots generierte Desinformationen an Reichweite gewinnen.
- Gestaltung von Empfehlungssystemen und anderen relevanten algorithmischen Systemen: Die Algorithmen der Plattform, die darauf ausgelegt sind, Trendthemen zu fördern, könnten unbeabsichtigt Bot-gesteuerte Narrative verstärken, wenn sie nicht angemessen überwacht werden.
- Systeme zur Moderation von Inhalten: Die Entdeckung und Entfernung von über 50.000 gefälschten Konten von Online-Bots deutet auf aktive Moderationsbemühungen hin. Das schiere Volumen der Bot-Aktivitäten deutet jedoch auch darauf hin, dass die bestehenden Systeme möglicherweise nicht mit den ausgefeilten Desinformationstaktiken Schritt halten können und das weiterhin das potenzielle Risiko von Online-Bots besteht.
- Geschäftsbedingungen: X verbietet automatisierte Konten, die betrügerische Praktiken anwenden. Gleichzeitig ist unklar, wie wirksam die Richtlinien angesichts der sich ständig weiterentwickelnden Praktiken durchgesetzt werden können.
- Systeme zur Auswahl und Anzeige von Werbung: X bietet das Schalten von Werbung auf seinen Diensten an. Sollten Bots Werbemechanismen ausnutzen, um Desinformationen zu verbreiten, könnte dies die Verbreitung schädlicher Inhalte noch weiter verstärken.

Die analysierten Risikofaktoren verdeutlichen, dass es weitere potenzielle Risiken gibt, die die Plattform berücksichtigen sollte. In dieser Fallstudie liegt der Fokus jedoch auf der potenziellen Präsenz von Online-Bots.

4. Bewertungsindikatoren

4.1 Abschätzung der Auswirkungen

Um systemische Risiken, die von digitalen Online-Plattformen ausgehen, wirksam zu bekämpfen, ist es notwendig, ihre potenziellen Auswirkungen anhand klar definierter Indikatoren zu messen. Hierdurch können Online-Plattformen den Umfang, das Ausmaß und die Möglichkeiten zur Abhilfe von Risiken systematisch bewerten und so einen strukturierten sowie umfassenden Ansatz für das Risikomanagement gewährleisten. In Übereinstimmung mit Art. 34 DSA sind die Anbieter*innen verpflichtet, Risikobewertungen durchzuführen, die spezifisch auf ihre Dienstleistungen abgestimmt sind. Die Bewertung von Auswirkung konzentriert sich insbesondere auf die Schwere der Risiken und stellt sicher, dass die Online-Plattformen die Risiken priorisieren, die eine erhebliche Bedrohung für die Nutzer*innen und die Gesellschaft als Ganzes darstellen.

Erwägungsgrund 79 DSA unterstreicht außerdem die Bedeutung der Bewertung potenzieller negativer Auswirkungen und ihrer weiterreichenden Folgen. Die Online-Plattformen werden aufgefordert, sowohl die Schwere als auch die Wahrscheinlichkeit von systemischen Risiken zu berücksichtigen. Dazu gehört die Bewertung, ob ein Risiko eine große Anzahl von Personen betreffen könnte, die potenzielle Abhilfefähigkeit des Schadens und die Schwierigkeit, den Zustand wie vor dem Eintreten des Risikos wiederherzustellen. Ein solcher Ansatz stellt sicher, dass Online-Plattformen Risiken nicht nur auf individueller oder plattformspezifischer Ebene angehen, sondern auch im Hinblick auf ihre gesamtgesellschaftlichen Auswirkungen, so dass sie ihre rechtlichen und ethischen Verpflichtungen erfüllen können.

Bei der Abschätzung der Auswirkungen müssen die Online-Plattformen die Art und Schwere des Schadens berücksichtigen, die für Einzelpersonen entstehen könnten. Dazu gehören direkte Schäden für die Nutzer*innen, wie z.B. Aussetzung gegenüber rechtswidrigen Inhalten oder missbräuchliches Verhalten, und indirekte Schäden für Nicht-Nutzer*innen, wie z.B. Opfer von sexueller Ausbeutung, Missbrauch von Kindern oder Betrug.

Zwar sind Messgrößen für die Quantifizierung von Risiken und deren Auswirkungen notwendig, ebenso ist auch die Transparenz sowie qualitative Bewertung essenziell. Wenn man sich zu sehr auf starre Messgrößen verlässt, kann man die nuancierte und vernetzte Natur der Risiken übersehen, insbesondere wenn sie über eine einzelne Plattform hinaus auf breitere gesellschaftliche Strukturen übergreifen. Online-Plattformen müssen ein Gleichgewicht zwischen quantitativen Messgrößen und qualitativen Erkenntnissen sicherstellen, damit sie Risiken ganzheitlich angehen können und gleichzeitig das öffentliche Vertrauen und die Rechenschaftspflicht bewahren.

In diesem Abschnitt wird die Tabelle mit den Schlüsselindikatoren für die Bewertung des Umfangs, des Ausmaßes und der Abhilfemöglichkeiten von nutzungs-, inhalts- und verhaltensbezogenen Risiken untersucht, die einen detaillierten Rahmen für das Verständnis der Komplexität systemischer Risiken bietet.

Art des Risikos	Definition der Indikatoren pro Risikoart		
	Umfang	**Ausmaß**	**Abhilfemöglichkeiten**
Nutzungsbezogene Risiken	Bewertet, wie Designmerkmale und die Möglichkeiten der Plattform die Nutzer*innen, insbesondere gefährdete Bevölkerungsgruppen, Risiken aussetzen. Dazu gehört die Bewertung der demografischen und geografischen Verteilung der betroffenen Nutzer*innen, die Prävalenz schädlicher Inhalte, die auf gefährdete Personen abzielen, und wie das Design der Plattform zu schädlichen Ergebnissen beiträgt.	Misst die langfristigen und unmittelbaren Auswirkungen der Plattformnutzung auf Einzelpersonen und Gemeinschaften. Dazu gehören negative Auswirkungen auf die psychische Gesundheit, wirtschaftliche Auswirkungen auf die betroffenen Gruppen und sekundäre Schäden für Familien oder Gemeinschaften. Es wird berücksichtigt, ob die Designmerkmale der Plattform zu systemischen Problemen wie Suchtverhalten oder Ausnutzung führen.	Bewertet die Möglichkeit, den durch das Design der Plattform verursachten Schaden zu mindern. Dabei werden Maßnahmen in Betracht gezogen, wie die Bereitstellung psychosozialer Unterstützung, die Umsetzung von Schutzmaßnahmen für gefährdete Nutzer*innen, die Neugestaltung schädlicher Funktionen und das Angebot von Bildungsressourcen zur Sensibilisierung der Nutzer*innen.
Inhaltsbezogene Risiken	Klärt ab, wie weit rechtswidrige Inhalte verbreitet werden und welche Gemeinschaften davon betroffen sind. Bewertet wird das Ausmaß der Gefährdung, z.B. die Anzahl der Nutzer*innen sowie die betroffenen Herkunftsregionen oder Sprachgruppen. Es wird auch berücksichtigt, ob sich die rechtswidrigen Inhalte über die Plattform hinaus verbreitet haben und sich auf andere digitale oder physische Bereiche ausgewirkt haben.	Untersucht das Ausmaß des gesellschaftlichen und individuellen Schadens, der durch die Verbreitung rechtswidriger Inhalte entsteht. Dies umfasst die Bewertung von Rechtsverstößen, psychischen oder physischen Schäden für Nutzer*innen, Folgen für die öffentliche Gesundheit und breitere gesellschaftspolitische Auswirkungen, wie z.B. die Beeinflussung von Wahlen oder die Erschütterung des öffentlichen Vertrauens.	Bewertet die Fähigkeit der Plattform, den Ursprungszustand wiederherzustellen, nachdem die Nutzer*innen rechtswidrigen Inhalten ausgesetzt waren. Berücksichtigt werden Maßnahmen zur Wiederherstellung des Vertrauens der Nutzer*innen, die Unterstützung der betroffenen Personen sowie rechtliche oder institutionelle Maßnahmen zur Schadensbehebung.
Verhaltensbezogene Risiken	Untersucht, wie der Missbrauch von Funktionen der Online-Plattform durch Akteur*innen Schaden anrichtet. Dazu gehört die Anzahl der Nutzer*innen, die von missbräuchlichem Verhalten betroffen sind, die plattformübergreifende Reichweite schädlicher Kampagnen, die schädliches Verhalten verstärken. Es wird auch bewertet, wie diese Aktivitäten verschiedene Nutzergruppen oder Gemeinschaften durchdringen.	Berücksichtigt die Schwere des Schadens, der durch missbräuchliches Verhalten verursacht wird. Dazu gehört der Leidensdruck oder Schaden, den die Nutzer*innen erfahren, die rechtlichen Folgen missbräuchlicher Kampagnen, die Schädigung des Rufs der Plattform und die breitere soziale Spaltung oder Polarisierung, die aus solchen Aktivitäten resultiert.	Konzentriert sich auf die Fähigkeit der Plattform, den Missbrauch zu bekämpfen, indem versucht wird, den verursachten Schaden rückgängig zu machen. Bewertet die Wirksamkeit der Kontrolle von Narrativen, die Aktualisierung von Richtlinien zur Verhinderung von Wiederholungen und Unterstützungsmechanismen für betroffene Nutzer*innen, einschließlich Beratung und Rechtsbehelfe.

Tabelle 5: Definition der Indikatoren pro Risikoart in Bezug auf die Auswirkungen

Art des Risikos	Potenzielle Indikatoren für die Abschätzung von Auswirkungen		
	Umfang	**Ausmaß**	**Abhilfemöglichkeiten**
Nutzungsbezogen	• Anzahl der gefährdeten Nutzer*innen, die schädlichen Inhalten ausgesetzt sind (z.B. Anzahl der Expositionen, Prozentsatz der gefährdeten, betroffenen Nutzer*innen) • Plattformübergreifende Auswirkungen auf gefährdete Personen (z.B. Anzahl der Sekundär-Plattformen mit entsprechenden schädlichen Inhalten, Korrelation der Expositionsmetriken) • Demografische und geografische Verteilung der betroffenen Nutzer*innen (z.B. Prozentsatz der gefährdeten Nutzer*innen in verschiedenen Bevölkerungsgruppen/Regionen) • Auswirkungen auf verbundene Gemeinschaften (z.B. Anzahl der gemeldeten Vorfälle in der Gemeinschaft, sekundäre Auswirkungswerte)	• Schwere der Auswirkungen auf die psychische Gesundheit (z.B. Berichte über Angstzustände, Depressionsraten, Umfrageergebnisse zum psychischen Wohlbefinden) • Langfristige Auswirkungen auf gefährdete Nutzer*innen (z.B. % mit anhaltenden Schäden, Rückfallquoten) • Sozioökonomische Auswirkungen (z.B. wirtschaftliche Verluste in den betroffenen Gemeinden, prozentualer Anstieg der gemeldeten Fälle von Ausbeutung) • Sekundäre Auswirkungen auf Familien/ Gemeinschaften (z.B. Prozentsatz der betroffenen Familien, Erhebungen zum Stress in der Gemeinschaft)	• Verfügbarkeit von psychosozialer Unterstützung (z.B. Zugangsrate zu psychosozialen Diensten, Anzahl der Unterstützungsfälle) • Schnelligkeit und Einfachheit der Wiederherstellung der Sicherheit (z.B. durchschnittliche Zeit bis zur Wiederherstellung der Sicherheit, Zufriedenheitswerte der betroffenen Nutzer*innen) • Wirksamkeit der Bildungsressourcen (z.B. Nutzungsrate der Ressourcen, Verbesserung der Risikobewusstseinsdaten der Nutzer*innen)
Inhaltsbezogen	• Anzahl der Nutzer*innen, die mit rechtswidrigen Inhalten auf der Plattform in Berührung kommen (z.B. einmalige Aufrufe, Anzahl der Aufrufe, % der Gesamtnutzenden) • Umfang der rechtswidrigen Inhalte, die über Gemeinschaften oder Regionen hinweg geteilt werden (z.B. % der Regionen oder Sprachgruppen, die betroffen sind, semantische Ähnlichkeitswerte) • Verbreitung rechtswidriger Inhalte außerhalb der Plattform (z.B. Erwähnungen oder Wiederveröffentlichungen auf sekundären Online-Plattformen, Nachrichtenartikel mit Verweis auf Inhalte)	• Schwere der psychischen oder physischen Schädigung (z.B. Erhebungen über die wahrgenommene Schädigung, Berichte von psychosozialen Meldestellen, gemeldete Vorfälle von Schädigung) • Rechtliche Auswirkungen (z.B. Anzahl der behördlichen Maßnahmen, Geldbußen oder Gerichtsverfahren) • Auswirkungen auf die öffentliche Gesundheit/Sicherheit (z.B. Anzahl der Gesundheitswarnungen im Zusammenhang mit dem Inhalt, Berichte von Notdiensten) • Soziopolitische Auswirkungen (z.B. Einfluss auf Wahlergebnisse, Umfragen zum öffentlichen Vertrauen, % der betroffenen Regionen)	• Wiederherstellung des Vertrauens der Nutzer*innen (z.B. Umfragen nach Vorfällen, Metriken zur Nutzerbindung) • Unterstützung der psychologischen Genesung (z.B. Zugang zu Beratungsdiensten, Inanspruchnahme von Unterstützungsdiensten) Wiederherstellung des Rufs (z.B. Zeit bis zur Klärung rufschädigender Ansprüche) • Institutionelle/juristische Abhilfe (z.B. Anzahl und Ergebnisse der gelösten Fälle)

Art des Risikos	Potenzielle Indikatoren für die Abschätzung von Auswirkungen		
	Umfang	**Ausmaß**	**Abhilfemöglichkeiten**
Verhaltensbezogen	• Anzahl der Nutzer*innen, die von Belästigung oder missbräuchlichem Verhalten auf der Plattform betroffen sind (z.B. Anzahl der Berichte oder Beschwerden, Prozentsatz der betroffenen Nutzenden der Plattform) • Plattformübergreifende Reichweite von schädlichen Kampagnen (z.B. Anzahl der Wiederveröffentlichungen auf anderen Online-Plattformen, Verbreitungsgrad für verwandte Begriffe) • Ausmaß der Echokammern, die missbräuchliches Verhalten verstärken (z.B. Ähnlichkeitsindizes für Inhalte, Netzwerkanalyse von Nutzerclustern) • Anteil der Plattform, der von ruchlosem Verhalten betroffen ist (z.B. % der gesamten Inhalte oder Nutzer*innen, die an Trolling-Kampagnen beteiligt sind)	• Ausmaß der Belastung oder Schädigung (z.B. Umfrageergebnisse zur Belastung, Anzahl der Anträge auf psychosoziale Unterstützung) • Auswirkungen auf das öffentliche Vertrauen und den Ruf der Plattform (z.B. Ergebnisse der Stimmungsanalyse, Vertrauensindex der Nutzer*innen) • Rechtliche Auswirkungen von Missbrauchskampagnen (z.B. Anzahl der rechtlichen Schritte, Geldbußen) • Schwere der sozialen Spaltung (z.B. Veränderungen der Polarisierungsindizes, Umfrageergebnisse zum sozialen Zusammenhalt)	• Umkehrung der Auswirkungen von Kampagnen (z.B. Zeitaufwand für die Kontrolle der Erzählung, Änderung der Ausbreitungsrate) • Abschwächung künftiger Verhaltensweisen (z.B. Verringerung wiederkehrender Verhaltensweisen, Wirksamkeit von Richtlinienänderungen) • Verfügbarkeit von Support/Wiederherstellung für betroffene Nutzer*innen (z.B. Anzahl der Nutzer*innen, die Unterstützung in Anspruch nehmen, Zufriedenheitsbewertungen)

Tabelle 6: Potenzielle Indikatoren für die Abschätzung von Auswirkungen

Fallstudie: Online-Bots

Wie oben erläutert, sollten Online-Plattformen die drei Dimensionen Umfang, Ausmaß und Abhilfefähigkeit bewerten, um die Schwere der Auswirkungen zu verstehen. Im Folgenden werden einige der oben genannten Indikatoren zur Abschätzung der Auswirkungen der Präsenz von Online-Bots auf X in Deutschland angewandt. Dabei wird untersucht, welche Arten von Daten im Kontext von X zur Bewertung der verschiedenen Indikatoren erforderlich wären. Es ist zu berücksichtigen, dass Indikatoren und Schwellenwerte je nach Diensten, Nutzerbasis etc. und je nach Online-Plattform unterschiedlich ausfallen können.

Eine der größten Herausforderungen bei der Ermittlung von Metriken zur Abschätzung der Auswirkungen liegt in der vorausschauenden Natur der Bewertung. Im Fall von X wäre es relativ einfach, auf frühere Vorfälle zu verweisen, bei denen Bots identifiziert wurden, und die entsprechenden Daten zu analysieren. Die Bewertung potenzieller Risiken ohne Präzedenzfälle stellt jedoch eine Herausforderung dar. Ohne historische Fälle, auf die Anbieter*innen zurückgreifen können, müssen sie sich auf theoretische Modelle, simulierte Szenarien oder allgemeine Indikatoren stützen, die möglicherweise nicht vollständig mit der Dynamik der jeweiligen Plattform übereinstimmen. Dieser Mangel an Präzedenzfällen verdeutlicht die Schwierigkeit, Risiken zu quantifizieren, und unterstreicht die Bedeutung der Entwicklung flexibler, anpassungsfähiger Indikatoren, die sowohl bekannte als auch hypothetische Bedrohungen abdecken können. Das Gleichgewicht zwischen diesem prädiktiven Ansatz und der für verwertbare Erkenntnisse erforderliche Genauigkeit bleibt eine Herausforderung im Risikobewertungsprozess.

In der vorliegenden Fallstudie müsste die Plattform eine Reihe von Metriken untersuchen, um Umfang, Ausmaß und Abhilfemöglichkeiten zu bewerten. Einige relevante Daten zur Bewertung der Indikatoren sind im Folgenden aufgelistet:

Abmessung: Umfang

- **Anzahl der Nutzer*innen, die von Belästigung oder missbräuchlichem Verhalten auf der Plattform betroffen sind:** Bewertung der Reichweite und Sichtbarkeit von Bot-Aktivitäten durch Analyse von Impressionen, Interaktionen und der Verteilung von Bot-generierten Inhalten über die Nutzerbasis. Aus vorherigen Fällen geht hervor, dass mit Online-Bots ein erheblicher Teil des deutschsprachigen Publikums erreicht werden kann. Dies könnte ggf. das Vertrauen in die gesellschaftliche Debatte untergraben.
- **Plattformübergreifende Reichweite von schädlichen Kampagnen:** Bewertung der plattformübergreifenden Verbreitung von Bot-generierten Inhalten durch Überwachung von geteilten Links, eingebetteten Beiträgen und Trends innerhalb des Social-Media-Ökosystems. Im vorliegenden Fallbeispiel deuten vorherige Online-Bots Kampagnen darauf hin, dass das Risiko besteht, dass plattformübergreifend Desinformationen geteilt und amplifiziert werden könnten.
- **Ausmaß der Echokammern, die missbräuchliches Verhalten verstärken:** Durchführung von Netzwerkanalysen zur Ermittlung von Mustern koordinierter Verstärkung, Gruppierung von Konten und wiederholter Verwendung identischer Botschaften. Botnetzwerke sind oft an einer koordinierten Verstärkung beteiligt und bilden Echokammern, die wiederholt Desinformationen

verbreiten. Eine Netzwerkanalyse könnte Cluster dieser gefälschten Konten aufdecken, die im Falle des deutschen Beispiels pro-russische Narrative verbreiten.

- **Anteil der Plattform, der von Verhalten betroffen ist:** Das potenziell große Volumen der Bot-Aktivitäten könnte einen erheblichen Teil der deutschsprachigen Inhalte auf X ausmachen.

Abmessung: Ausmaß

- **Ausmaß der Belastung oder Schädigung:** Bewertung des durch Bot-Aktivitäten verursachten Schadens durch die Erhebung qualitativer und quantitativer Daten über die Wahrnehmung der Nutzer*innen, gemeldete Beschwerden und gekennzeichnete Inhalte.
- **Auswirkungen auf das öffentliche Vertrauen und den Ruf der Plattform:** Ermittlung des durch Bot-Aktivitäten verursachten Schadens durch die Erhebung qualitativer und quantitativer Daten über die Wahrnehmung der Nutzer*innen, gemeldete Probleme und gekennzeichnete Inhalte.
- **Rechtliche Auswirkungen von Missbrauchskampagnen:** Verfolgung von Rechts- und Compliance-Problemen im Zusammenhang mit Bot-gesteuerten FIMI, einschließlich Mitteilungen von Regierungsbehörden und Geldstrafen gemäß den einschlägigen Vorschriften.

Abmessung: Abhilfefähigkeit

- **Umkehrung der Auswirkungen von Kampagnen:** Bewertung der Wirksamkeit von Reaktionsstrategien durch Messung des Zeit- und Ressourcenaufwands, der erforderlich ist, um Bot-gesteuerten Desinformationen entgegenzuwirken, einschließlich Änderungen der Verbreitungsrate von Inhalten und der Reichweite des Publikums.
- **Abschwächung künftiger Verhaltensweisen:** Bewertung der Fähigkeit der Plattform, künftige Bot-Aktivitäten zu verhindern, durch Analyse der Wirksamkeit von Algorithmus-Updates, Richtlinienimplementierungen und proaktiven Überwachungstools.
- **Verfügbarkeit von Support/Wiederherstellung für betroffene Nutzer*innen:** Messung der Zugänglichkeit und Effektivität von Support-Mechanismen für Nutzer*innen, die von Bot-Aktivitäten betroffen sind, einschließlich Bewertungen der Nutzerzufriedenheit, Zeitrahmen für die Problemlösung und Einsatz von Schutz-Tools.

Neben der Analyse spezifischer Metriken sollte die Plattform auch ihre Richtlinien und Nutzungsbedingungen überprüfen, um diese Indikatoren effektiv zu bewerten. Klare Richtlinien zur Inhaltsmoderation, Bot-Erkennung und Nutzerunterstützung können die Fähigkeit der Plattform positiv beeinflussen, die Auswirkungen von Desinformationskampagnen rückgängig zu machen, künftige Verhaltensweisen einzudämmen und angemessene Wiederherstellungsmechanismen für betroffene Nutzer*innen bereitzustellen. Die Abschätzung der potenziellen Auswirkungen von Online-Bots auf X verdeutlicht, dass diese weitreichend sein und sowohl Nutzer*innen als auch das öffentliche Vertrauen und das Informationsökosystem im weiteren Sinne beeinträchtigen könnten. Das Risiko von Online-Bots auf X wird damit als Risiko mit **schwerwiegenden Auswirkungen** eingestuft.

4.2 Abschätzung der Wahrscheinlichkeit

Dieser Bewertungsmaßstab bezieht sich auf die Wahrscheinlichkeit, dass Nutzer*innen auf systemische Risiken wie rechtswidrige Inhalte, Desinformation oder koordiniertes unauthentisches Verhalten stoßen, und dass Online-Plattformen zur Förderung von Straftaten genutzt werden. Er verwendet einen strukturierten Ansatz, um die Wahrscheinlichkeit des Auftretens dieser Risiken auf der Grundlage plattformspezifischer Faktoren, historischer Trends und breiterer Kontextelemente zu bestimmen.

Eine solide Wahrscheinlichkeitsbewertung stützt sich auf eine Reihe von Dimensionen, um die Wahrscheinlichkeit von Risiken zu bewerten. Online-Plattformen sollten ein Gleichgewicht zwischen Messbarkeit und bedeutenden Auswirkungen herstellen (Broughton Micova & Calef, 2023). Zwar können Metriken wertvolle Erkenntnisse liefern, doch besteht die Gefahr, dass sich Online-Plattformen unverhältnismäßig stark auf leicht messbare, aber weniger einschneidende Risiken konzentrieren und dabei möglicherweise wichtige, aber schwerer zu quantifizierende Probleme übersehen. Beispielsweise können systemische Risiken, die neuartige Funktionen oder aufkommende Bedrohungen beinhalten, sich traditionellen Messrahmen entziehen. Dennoch erfordern diese Bereiche Aufmerksamkeit und Strategien zur Risikominderung auf der Grundlage verfügbarer Forschung und historischer Muster.

Wahrscheinlichkeitsbewertungen sind jedoch mit inhärenten Herausforderungen verbunden. Ein übermäßiger Rückgriff auf Kennzahlen kann dazu führen, dass man sich zu sehr auf die Einhaltung und Durchsetzung von Vorschriften konzentriert, anstatt sich mit den zugrunde liegenden Risiken zu befassen, die diese Kennzahlen darstellen sollen (Marsh, 2024). Dies unterstreicht die Bedeutung eines ganzheitlichen Ansatzes, der quantitative Daten mit qualitativen Erkenntnissen verbindet und sicherstellt, dass auch schwer messbare Risiken angemessene Aufmerksamkeit erhalten.

Letztendlich müssen Wahrscheinlichkeitsbewertungen ein Gleichgewicht zwischen Präzision und Anpassungsfähigkeit herstellen, indem sie plattformspezifische Merkmale, frühere Muster und kontextbezogene Faktoren nutzen, um die Wahrscheinlichkeit systemischer Risiken zu bewerten. Dieser Ansatz hilft nicht nur bei der Einhaltung regulatorischer Rahmenbedingungen wie dem DSA, sondern stellt auch sicher, dass Online-Plattformen proaktiv Risiken angehen, die weitreichende gesellschaftliche Folgen haben könnten. Auf der Grundlage der Literatur konzentriert sich der ISD-Rahmen für Wahrscheinlichkeitsbewertungen auf drei Dimensionen: Plattformmerkmale, historische Daten und Trends sowie Kontextfaktoren.

Art des Risikos	Definition der Indikatoren pro Risikoart		
	Plattformmerkmale	**Historische Daten und Trends**	**Kontextuelle Faktoren**
Nutzungsbezogene Risiken	Ergeben sich direkt aus dem Design und den Möglichkeiten einer Plattform. Diese Dimension bewertet Funktionen wie endloses Scrolling, Autoplay oder algorithmisches Targeting, die zu schädlichen Ergebnissen wie Sucht oder einer Verschlechterung der psychischen Gesundheit beitragen können. Online-Plattformen mit schlecht gestalteten Benutzeroberflächen oder unzureichenden Sicherheitsvorkehrungen für gefährdete Nutzer*innen haben eine höhere Wahrscheinlichkeit, dass Nutzer*innen diesen Risiken auszusetzt werden.	Helfen bei der Identifizierung von Schadensmustern im Zusammenhang mit der Nutzung von Online-Plattformen, wie z.B. wiederholte Vorfälle von Notlagen oder psychischen Krisen bei Nutzer*innen. Wenn bestimmte Plattformfunktionen oder Gruppen kontinuierlich Schäden verursachen, deutet dies auf eine hohe Wahrscheinlichkeit für weitergehende oder verschlimmernde Risiken hin.	Externe Ereignisse wie z.B. soziale Isolation während Pandemien oder politische Instabilität können nutzungsbedingte Risiken verschlimmern. So können sich beispielsweise gefährdete Nutzer*innen in Krisenzeiten stärker auf eine Plattform verlassen, wodurch sie vermehrt schädlichen Inhalten oder manipulativen Verhaltensweisen ausgesetzt sind. Online-Plattformen müssen solch kontextuelle Veränderungen und Ereignisse antizipieren, um die Risiken wirksam zu mindern.
Inhaltsbezogene Risiken	Entstehen aus Funktionen wie Empfehlungssystemen, Benutzeroberflächen und Instrumenten zur Inhaltserstellung, die das Produzieren, Verbreiten oder Verstärken rechtswidriger Inhalte erleichtern. So können Online-Plattformen mit minimalen Mechanismen zur Überprüfung von Inhalten oder freizügigen Werbesystemen unbeabsichtigt die Wahrscheinlichkeit erhöhen, dass rechtswidrige Inhalte geteilt und verbreitet werden wie z.B. rechtswidrige Hassrede oder Material zu Kindesmissbrauch.	Bewertet Muster vergangener Vorfälle wie z.B. die Häufigkeit des Hochladens rechtswidriger Inhalte oder die Prävalenz extremistischen Materials. Anhand historischer Daten lässt sich feststellen, welche Erfolge die Plattform bei der Moderation solcher Inhalte erzielt hat und ob die bestehenden Systeme wirksam sind, um deren Wiederholung zu verhindern.	Werden häufig von externen Faktoren wie politischen Ereignissen, Krisen oder kulturellen Empfindlichkeiten beeinflusst. Bei Wahlen oder in Zeiten von Unruhen kann beispielsweise die Erstellung und Verbreitung rechtswidriger Inhalte sprunghaft ansteigen. Die Online-Plattformen müssen diese Dynamik antizipieren, indem sie geografische und demografische Trends analysieren, um Schwachstellen zu erkennen.

Art des Risikos	Definition der Indikatoren pro Risikoart		
	Plattformmerkmale	**Historische Daten und Trends**	**Kontextuelle Faktoren**
Verhaltensbezogene Risiken	Beziehen sich darauf, wie Kommunikationswerkzeuge, algorithmische Designs oder mangelnde Kontosicherheit Akteur*innen die Ausnutzung von Schwachstellen ermöglichen. Beispielsweise können schwache Verifizierungsprozesse oder unregulierte Nachrichtensysteme Aktivitäten wie Trolling, Doxing oder koordinierte Desinformationskampagnen erleichtern.	Die Online-Plattformen sollten die Häufigkeit und die Muster schädlicher Verhaltensweisen wie Belästigung, Missbrauch oder Manipulationsversuche analysieren. Wiederholte Vorfälle, wie z.B. koordinierte Angriffe oder Desinformationskampagnen, deuten auf eine höhere Wahrscheinlichkeit hin, dass sich ein ähnliches Verhalten wiederholt, vor allem, wenn frühere Eindämmungsstrategien unzureichend waren.	Reagieren oft auf externe Ereignisse wie Wahlen, soziale Bewegungen oder globale Krisen. So können beispielsweise umstrittene politische Perioden oder aktuelle Online-Herausforderungen Trolling oder Belästigungen verstärken. Das Verständnis dieser kontextabhängigen Auslöser hilft den Online-Plattformen, sich auf einen Anstieg des schädlichen Verhaltens vorzubereiten.

Tabelle 7: Definition der Indikatoren pro Risikoart in Bezug auf die Wahrscheinlichkeit

Art des Risikos	Mögliche Indikatoren für die Wahrscheinlichkeitsabschätzung		
	Plattformmerkmale	**Historische Daten und Trends**	**Kontextuelle Faktoren**
Nutzungsbezogen	• Art der Dienste • Größe und Nutzerbasis • Geschäftsmodell • Gestaltung von Empfehlungssystemen und anderen relevanten algorithmischen Systemen • Systeme zur Moderation von Inhalten • Geschäftsbedingungen • Systeme zur Auswahl und Anzeige von Werbung • Datenbezogene Praktiken	• Muster, die darauf hinweisen, dass gefährdete Nutzende häufig auslösenden oder schädlichen Inhalten ausgesetzt sind. • Häufigkeit der gemeldeten psychischen Krisen, Notlagen oder Selbstverletzungen im Zusammenhang mit der Nutzung der Plattform. • Starkes Engagement für bestimmte schädliche Gemeinschaften (z.B. extremistische Gruppen, schädliche Subkulturen).	• Sozioökonomische oder politische Instabilität, die das Verhalten der gefährdeten Nutzenden beeinflusst. • Trends, die auf eine stärkere Nutzung der Plattform durch gefährdete Nutzende während einer Krise oder Instabilität hinweisen. • Demografische Anfälligkeit für Manipulation (z.B. Altersgruppen, die für bestimmte Arten von Desinformationsnarrativen anfällig sind). • Ereignisse, die zu sozialer Isolation führen, was gefährdete Nutzende anfälliger für die Beeinflussung durch die Plattform machen könnte.

Art des Risikos	Mögliche Indikatoren für die Wahrscheinlichkeitsabschätzung		
	Plattformmerkmale	**Historische Daten und Trends**	**Kontextuelle Faktoren**
Inhaltsbezogen	• Art der Dienste • Größe und Nutzerbasis • Geschäftsmodell • Gestaltung von Empfehlungssystemen und anderen relevanten algorithmischen Systemen • Systeme zur Moderation von Inhalten • Geschäftsbedingungen • Systeme zur Auswahl und Anzeige von Werbung • Datenbezogene Praktiken	• Häufigkeit und Umfang des Hochladens rechtswidriger Inhalte (z.B. Häufigkeit der markierten Beiträge). • Häufigkeit von extremistischen oder hasserfüllten Inhalten in der Vergangenheit • Rate der Auseinandersetzung mit rechtswidrigen Inhalten • Verhältnis zwischen den zur Moderation gekennzeichneten und den insgesamt eingestellten Inhalten • Vorhandensein von organisierten Hassgruppen oder -netzwerken auf der Plattform • Nutzung von Meldemechanismen	• Vermehrte Erstellung rechtswidriger Inhalte in Zeiten sozialer Unruhen oder bei Krisenereignissen. • Empfindlichkeit der Plattforminhalte gegenüber politischen oder kulturellen Faktoren, die zu schädlichen Verhaltensweisen anregen können. • Einfluss von externen Ereignissen (z.B. Wahlen, Pandemien) auf die Verbreitung rechtswidriger Inhalte. • Geografische und demografische Trends, die auf eine Zunahme rechtswidriger Inhalte in bestimmten Gebieten hindeuten. • Geografische, sprachliche und juristische Unterschiede (und das Verständnis der Nutzer*innen und Moderator*innen dafür). • Erhöhte Nachfrage nach rechtswidrigen Waren oder Dienstleistungen.
Verhaltensbezogen	• Art der Dienste • Größe und Nutzerbasis • Geschäftsmodell • Gestaltung von Empfehlungssystemen und anderen relevanten algorithmischen Systemen • Systeme zur Moderation von Inhalten • Geschäftsbedingungen • Systeme zur Auswahl und Anzeige von Werbung • Datenbezogene Praktiken	• Häufiges Auftreten von beleidigendem, belästigendem oder trollendem Verhalten. • Muster, die auf koordinierte Angriffe (z.B. Brigading, Doxxing) durch Gruppen von Nutzern hinweisen. • Aufgezeichnete Daten von Desinformationsverbreitung oder Manipulationsversuchen. • Häufige Fälle von Kontosperrungen aufgrund von Verhaltensverstößen. • Prävalenz und Häufigkeit missbräuchlicher Mitteilungen. • Schwachstellen, die es Bots oder automatisierten Konten in der Vergangenheit ermöglichten, unkontrolliert zu operieren.	• Externe Ereignisse, die missbräuchliches oder trollendes Verhalten verstärken (z.B. umstrittene politische Zeiten, Wahlen). • Einfluss aktueller Trends oder Online-Herausforderungen, die schädliche Verhaltensweisen fördern. • Koordinierte Desinformations-/ FIMI- oder Belästigungskampagnen als Reaktion auf aktuelle Ereignisse. • Soziopolitische Bewegungen oder Krisen, die feindseliges Online-Verhalten verschärfen könnten.

Tabelle 8: Potenzielle Indikatoren für die Abschätzung der Wahrscheinlichkeit

Fallstudie: Online-Bots

Wie zuvor festgestellt, könnten sich Online-Bots erheblich auf das öffentliche Vertrauen in das Informationsökosystem auswirken. Um abzuschätzen, wie wahrscheinlich das Eintreten des erkannten Risikos ist, ist eine detaillierte Untersuchung der Plattformmerkmale, historischer Daten und Trends sowie kontextueller Faktoren notwendig, die zu diesen Risiken beitragen. Im Fall des Risikos der Verbreitung von Desinformationen oder FIMI durch Online-Bots ist insbesondere die Untersuchung technischer und betrieblicher Strukturen, historischer Muster missbräuchlichen Verhaltens und externer soziopolitischer Kontexte notwendig, um die Wahrscheinlichkeit abzuschätzen. Im Folgenden werden spezifische, auf diese Wahrscheinlichkeitsindikatoren zugeschnittene Metriken entwickelt, um einen praktischen Rahmen für eine solche Bewertung vorzulegen.

Abmessung: Plattformmerkmale

Für die Abschätzung der Wahrscheinlichkeit ist es notwendig, die Plattformmerkmale zu berücksichtigen. Dafür können Anbieter*innen auf das erstelle Risikoprofil zurückgreifen, um zu verstehen, inwiefern die Plattformmerkmale sie erhöhen oder verringern.

Abmessung: Historische Daten und Trends

Häufiges Auftreten von beleidigendem, belästigendem oder trollendem Verhalten: Die Identifizierung eines groß angelegten Bot-Netzes, das Desinformation betreibt, spiegelt eine immer wiederkehrende Herausforderung für X bei der Bewältigung automatisierten missbräuchlichen Verhaltens wider.

Muster, die auf koordinierte Angriffe durch Gruppen von Nutzer*innen hinweisen: Die synchronisierte Aktivität tausender gefälschter Konten in vorherigen Fällen weist auf ein Muster koordinierter Desinformationsbemühungen hin, die erneut auftreten könnten.

Aufgezeichnete Raten von Desinformationsverbreitung oder Manipulationsversuchen: Die Menge der in vorherigen Fällen von Bots erstellten Beiträge unterstreicht die Anfälligkeit der Plattform für Manipulationsversuche, die darauf abzielen, die öffentliche Wahrnehmung zu beeinflussen.

Häufige Fälle von Kontosperrungen aufgrund von Verhaltensverstößen: Die beträchtliche Anzahl von Kontosperrungen in vorherigen Fällen unterstreicht die laufenden Bemühungen zur Bekämpfung von Richtlinienverstößen, auch wenn das Fortbestehen solcher Aktivitäten auf die Notwendigkeit verbesserter Aufdeckungsmechanismen hinweist.

Prävalenz und Häufigkeit missbräuchlicher Mitteilungen: Die Prävalenz von Bot-gesteuerten Desinformationskampagnen trägt zur Verbreitung schädlicher Mitteilungen bei und beeinträchtigt die Nutzererfahrung und das Vertrauen.

Schwachstellen, die es Bots oder automatisierten Konten in der Vergangenheit ermöglichten, unkontrolliert zu operieren: Frühere Vorfälle, bei denen Bots genutzt wurden, zeigen systemische Schwachstellen auf, die eine kontinuierliche Bewertung und Abhilfe erfordern.

Abmessung: Kontextuelle Faktoren

Externe Ereignisse, die missbräuchliches oder trollendes Verhalten verstärken: Geopolitische Spannungen schaffen ein Umfeld, in dem staatlich geförderte Einrichtungen Bots einsetzen können, um die öffentliche Meinung auf Online-Plattformen wie X zu beeinflussen.

Einfluss aktueller Trends oder Online-Herausforderungen, die schädliche Verhaltensweisen fördern: Die rasche Übernahme von generativen KI-Tools durch Akteur*innen ermöglicht die Schaffung ausgeklügelter Bot-Netzwerke, die in der Lage sind, Desinformationen in großem Umfang zu produzieren und zu verbreiten.

Koordinierte Desinformations-/FIMI- oder Belästigungskampagnen als Reaktion auf aktuelle Ereignisse: Die orchestrierte Natur von vorherigen

Desinformationskampagnen spiegelt eine strategische Bemühung wider, Narrative in kritischen Zeiten zu manipulieren.

Soziopolitische Bewegungen oder Krisen, die feindseliges Online-Verhalten verschärfen könnten: Aktuelle gesellschaftspolitische Themen bieten einen fruchtbaren Boden für Bots, um Spaltungen auszunutzen und Uneinigkeit durch gezielte Desinformation zu verstärken.

Der Fall der von Russland unterstützten Desinformationskampagne, die sich gegen deutsche Nutzer*innen auf X richtete, zeigt, wie anfällig die Plattform für Bot-gesteuerte Manipulationen ist. Das angeführte Beispiel deutet darauf hin, dass das Risiko sehr wahrscheinlich ist, dass Online-Bots zu anderen Zeitpunkten auf der Plattform präsent sind, wenn es nicht wirksam gemindert wird.

4.3 Risikomatrix

Eine Risikomatrix ist ein unverzichtbares Instrument für Online-Plattformen, um Risiken systematisch zu kategorisieren und zu priorisieren. Durch die Kombination der Ergebnisse von Bewertungen der Auswirkungen und der Wahrscheinlichkeit in einer einzigen visuellen Darstellung vereinfacht die Matrix die komplexe Risikodynamik. Sie ermöglicht es Online-Plattformen und Aufsichtsbehörden, fundierte Entscheidungen zu treffen hinsichtlich der Ressourcenzuweisung, der Strategien zur Risikominderung und der Bemühungen, um die Einhaltung des DSA. Dieses Tool bietet eine optimierte Möglichkeit, das Zusammenspiel zwischen den potenziellen Auswirkungen von Risiken und ihrer Wahrscheinlichkeit zu verstehen und zu kommunizieren, was letztlich die Transparenz und Verantwortlichkeit in Risikomanagementprozessen fördert.

Die Risikomatrix ist in der Regel so aufgebaut, dass die Wahrscheinlichkeit auf der einen und die Auswirkungen auf der anderen Achse aufgetragen werden. Jede Achse ist in Skalen unterteilt, die üblicherweise als niedrig, mittel, hoch oder schwer kategorisiert werden, um das relative Ausmaß des Risikos für jede Dimension darzustellen. Durch diese Struktur entsteht ein Raster, in dem die Risiken auf der Grundlage ihrer Bewertungen dargestellt werden können, so dass die Beteiligten ihre Risikolandschaft visualisieren können. Risiken, die in die untere rechte Ecke der Matrix fallen (sehr hohe Wahrscheinlichkeit, sehr hohe Auswirkung), sollten als kritisch eingestuft werden und erfordern sofortige Maßnahmen. Risiken in der oberen linken Ecke (niedrige Wahrscheinlichkeit, niedrige Auswirkung) hingegen erfordern möglicherweise nur minimale Maßnahmen.

Für Online-Plattformen schafft diese strukturierte Visualisierung Klarheit und Orientierung und ermöglicht es ihnen, ihre Strategien zur Risikominderung an der Schwere und Wahrscheinlichkeit der einzelnen Risiken auszurichten. Ebenso können Aufsichtsbehörden die Matrix als Instrument nutzen, um systemische Risiken schnell zu bewerten und zu beurteilen, worauf sich die Bemühungen zur Einhaltung der Vorschriften und zur Überwachung konzentrieren sollten. Die Matrix unterstützt auch die Priorisierung von Maßnahmen zur Risikominderung im Einklang mit den Ressourcen und konzentriert sich auf die Risiken, die am wahrscheinlichsten eintreten und erheblichen Schaden verursachen.

Zwar bietet die Risikomatrix einen praktischen Rahmen, doch ist es entscheidend zu beachten, dass Schwellenwerte für die Kategorisierung von Risiken nicht pauschal

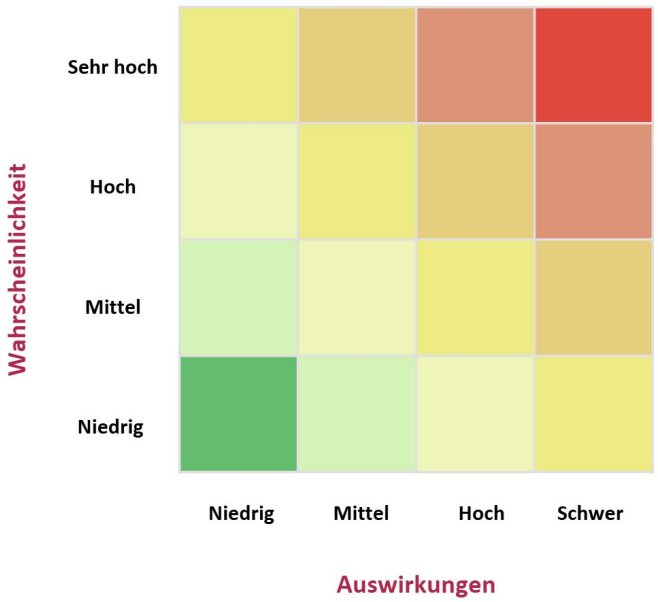

Abbildung 4: Risikomatrix mit Stufen der Auswirkungen und Wahrscheinlichkeit

von außen festgelegt werden können. Stattdessen müssen sie die spezifischen Eigenschaften der Plattform und den relevanten Kontext, in dem sie betrieben wird, berücksichtigen. Dafür müssen die Schwellenwerte mit wissenschaftlichen Daten abgeglichen und in Absprache mit den Beteiligten entwickelt werden, einschließlich der DSCs und externen Expert*innen.

Angesichts der dynamischen Natur von Online-Risiken sind Schwellenwerte und Kategorisierungen innerhalb der Risikomatrix nicht statisch. Die Online-Plattformen müssen ihre Risikobewertungen ständig überprüfen und anpassen, um sicherzustellen, dass sie relevant und wirksam bleiben. Dieser iterative Prozess umfasst die Einbeziehung von Interessengruppen mit regelmäßigen Konsultationen mit Regulierungsbehörden, der Zivilgesellschaft und Sachverständigen, um neue Daten, technologische Fortschritte und Erkenntnisse in die Matrix einzubeziehen. Des Weiteren geht es um eine wissenschaftliche Validierung, bei der die Schwellenwerte auf Grundlage neuer Forschungsergebnisse oder neuer Erkenntnisse über die Wahrscheinlichkeit und die Auswirkungen von Risiken aktualisiert werden.

Durch die Kombination von Abschätzungen der Auswirkungen und Wahrscheinlichkeit in einem einzigen Rahmen ermöglicht die Risikomatrix den Online-Plattformen eine effiziente Ressourcenzuweisung und die Einhaltung gesetzlicher Vorgaben. Der Nutzen der Matrix hängt jedoch von der sorgfältigen Kalibrierung der Schwellenwerte, der kontinuierlichen Einbeziehung der Interessengruppen und der Reaktion auf neue Daten und kontextuelle Veränderungen ab. Auf diese Weise wird die Risikomatrix zu mehr als nur einem Diagnoseinstrument – sie ist ein Eckpfeiler eines proaktiven, adaptiven Risikomanagements im digitalen Zeitalter.

Fallstudie: Online-Bots

Auf der Grundlage der Ergebnisse dieser Fallstudie kann das potenzielle Risiko, dass Online-Bots Desinformationen über X verbreiten, sowohl hinsichtlich der Auswirkungen als auch der Wahrscheinlichkeit als schwerwiegend eingestuft werden, was es in der kritischen oberen rechten Ecke der Risikomatrix platziert. Die Erfahrung mit vorherigen Fällen, in denen eine große Anzahl von Nutzer*innen mit Bot-generierter Desinformation konfrontiert wurde, sowie die Fähigkeit von Online-Bots, spaltende Narrative zu verstärken und das öffentliche Vertrauen zu untergraben, zeigt die tiefgreifenden potenziellen Auswirkungen dieses Risikos. Auch die Merkmale der Plattform – wie die Verbreitung von Inhalten in Echtzeit, historische Muster der Bot-Nutzung und kontextbezogene Schwachstellen wie geopolitische Spannungen – unterstreichen die hohe Wahrscheinlichkeit, dass sich solche Vorfälle wiederholen. Wichtig ist, dass die Einstufung eines Risikos innerhalb der Matrix von einer ständigen Überprüfung der Schwellenwerte abhängt, die sowohl durch wissenschaftliche Validierung als auch durch die kontinuierliche Zusammenarbeit mit Interessengruppen gestützt wird, einschließlich Regulierungsbehörden und der Zivilgesellschaft.

5. Minderungsmaßnahmen

Im dritten Schritt des Risikobewertungsprozesses liegt der Fokus darauf, wie die Verpflichtung zur Risikominderung praktisch umgesetzt werden kann, unter der Berücksichtigung des Grundsatzes der Verhältnismäßigkeit sowie der anschließenden Effektivitätsbewertung. Gemäß Art. 35 DSA sind Anbieter*innen dazu verpflichtet, eigenständig Maßnahmen zu ergreifen, um die in Art. 34 DSA beschriebenen und durch den Bewertungsprozess ermittelten spezifischen systemischen Risiken wirksam zu reduzieren. Der Hintergrund dieser gesteuerten Selbstregulierung liegt in der Informationsasymmetrie zwischen den Regulierungsbehörden und den Online-Plattformen. Diese führt dazu, dass zunächst ausschließlich die Plattformanbieter*innen Zugriff auf die Informationen haben, die erforderlich sind, um geeignete Risikominderungsmaßnahmen auszuwählen. Die Entscheidung darüber, welche Maßnahmen ergriffen werden, um die identifizierten systemischen Risiken zu mindern, obliegt somit den VLOPs und VLOSEs. Gleichwohl stellen der Art. 35 DSA sowie die Erwägungsgründe 86ff DSA einen Katalog mit einer Reihe von Beispielen für potenzielle Risikominderungsmaßnahmen bereit.

Neben dem DSA und den Erwägungsgründen gibt es derzeit weitere offizielle Dokumente, die die im DSA genannten Risikominderungsmaßnahmen bei spezifischen Risiken konkretisieren. Das ist zum einen der nur bedingt rechtlich bindende Verhaltenskodex zur Bekämpfung von rechtswidriger Hassrede im Internet, der 2016 unterzeichnet wurde und Risikominderungsmaßnahmen zur Minderung der Verbreitung rechtswidriger Inhalte (Art. 34 Abs. 1a) DSA) enthält. Zum anderen wurde im Vorlauf der Wahl zum Europäischen Parlament mit den Leitlinien für die Minderung systemischer Risiken für Wahlen weitere Empfehlungen für Maßnahmen zur Minderung tatsächlicher oder absehbarer nachteiliger Auswirkungen auf Wahlprozesse (Art. 34 Abs. 1c) DSA) ausgesprochen.

Diese vier Dokumente – die Erwägungsgründe, der DSA, der Verhaltenskodex und die Leitlinien – sind Grundlage für die im Rahmen dieser Studie erstellte Kategorisierung von Risikominderungs-maßnahmen. Zudem stärkt der überarbeitete Verhaltenskodex zur Bekämpfung von Desinformation die Selbstregulierungsstandards der Online-Plattformen (Europäische Kommission, 2022). Da dieser Kodex jedoch noch nicht ko-regulatorisch mit dem DSA verknüpft sind, wurden die darin enthaltenen Risikominderungsmaßnahmen nicht in die hier erstellte Übersicht einbezogen. Darüber hinaus erarbeitet die Europäische Kommission derzeit eine Leitlinie zum Kinder- und Jugendschutz im Rahmen des DSA, die Online-Plattformen als Orientierung dienen soll, um diesen Schutz mit Hilfe von spezifischen Risikominderungsmaßnahmen effektiver umzusetzen. Die hierin enthaltenen Empfehlungen könnten zukünftig ebenfalls in die Kategorisierung von Risikominderungsmaßnahmen aufgenommen werden.

Um sicherzustellen, dass VLOSPs und VLOSEs systemische Risiken wie vorgeschrieben mindern, sind sie durch Art. 41 DSA dazu verpflichtet, eine von den operativen Einheiten unabhängige Compliance-Abteilung einzurichten. Diese soll gewährleisten, dass »alle in Art. 34 DSA genannten Risiken ermittelt und ordnungsgemäß gemeldet werden, und dass angemessene, verhältnismäßige und wirksame Risikominderungsmaßnahmen gemäß Art. 35 DSA ergriffen werden« (Art. 41 Abs. 3b) DSA). Auch sind die Risikominderungsmaßnahmen, wie auch die Risikobewertungsmaßnahmen, Gegenstand der jährlichen unabhängigen Prüfung durch unabhängige Anbieter*innen (Art. 37 DSA, Erwägungsgrund 92 DSA).

Erwägungsgrund 93 DSA beinhaltet, dass diese Überprüfung in Form eines Berichts dem DSC am Niederlassungsort sowie der Europäischen Kommission und dem Europäischen Gremium für digitale Dienste übermittelt werden. In diesem Bericht sollten die Online-Plattformen darlegen, wie sie systemische Risiken ermittelt und welche Risikominderungsmaßnahmen sie umgesetzt haben. Die zuständigen Behörden übernehmen in dem Bewertungsprozess eine Kontrollfunktion. Das Gremium und die Europäische Kommission veröffentlichen jährlich einen Bericht, der neben systemischen Risiken auch sogenannte Gute Praktiken zur Minderung derselben enthält. Auch können Europäische Kommission und nationale DSCs Leitlinien veröffentlichen, um bewährte Verfahren vorzustellen und mögliche Maßnahmen zu empfehlen (Art. 35 Abs. 3 DSA). Darüber hinaus haben die Europäische Kommission und der zuständige DSC einen Zugang zu allen notwendigen Daten (Art. 40 DSA, Erwägungsgrund 96 DSA), um die Risikobewertung inklusive der Risikominderungsmaßnahmen zu überprüfen.

Erwähnenswert ist, dass der DSA mit dem Prinzip der Risikominderungsmaßnahmen davon ausgeht, dass durch die Nutzung von Technologien immer auch Restrisiken bestehen bleiben, die nicht vollständig verhindert wer-

den können (Mantelero, 2022). Es kann also vorkommen, dass Dienste deren Verwendung ein hohes systemisches Risiko darstellen, zulässig sind, sofern Risikominderungsmaßnahmen implementiert wurden.

5.1 Kategorien der Minderungsmaßnahmen

In den offiziellen Dokumenten, die im Zusammenhang mit dem DSA veröffentlicht wurden, werden Maßnahmen zur Minderung systemischer Risiken vorgestellt, die im Rahmen dieser Studie einer umfassenden Analyse unterzogen wurden. Die in den relevanten Gesetzestexten und Unterlagen identifizierten Maßnahmen sind dabei unterschiedlich detailliert und spezifisch ausgearbeitet. Diese Unterschiede wurden auf verschiedenen Ebenen betrachtet: Zum einen wurden allgemeine Risikominderungen erfasst, zum anderen spezifische, detaillierte Maßnahmen hervorgehoben. Die für diese Studie verwendete Kategorisierung der Minderungsmaßnahmen basiert auf den in Art. 35 Abs. 1 DSA genannten übergeordneten Maßnahmen.

Die im Rahmen dieser Studie vorgenommene Kategorisierung der Maßnahmen ist deshalb sinnvoll, um in der Vielzahl an Maßnahmen, die in den offiziellen Dokumenten aufgeführt werden, eine klare Struktur zu gewährleisten. Die Kategorien umfassen technische, organisatorische und inhaltliche Dimensionen, die für den effektiven Umgang mit systemischen Risiken von Bedeutung sind. Die festgelegten Ziele der jeweiligen Maßnahmen bilden die Grundlage für ihre Zuordnung zu spezifischen systemischen Risiken. Dabei zeigt sich, dass sich die Ziele und Einsatzbereiche vieler Maßnahmen teilweise überschneiden. Oft greifen die Maßnahmen ineinander, sodass sich einige Risikominderungsmaßnahmen in anderen Maßnahmen manifestieren und gegenseitig verstärken. Im Verlauf der folgenden Kategorienvorstellung (s. Anhang 1 für eine weiterführende Literatur-Datenbank zur Risikominderung und regulatorischen Leitlinien) wird diese Überlappung und Verflechtung deutlich werden.

Moderation

Art. 3 DSA definiert die Moderation von Inhalten als »die automatisierten oder nicht automatisierten – Tätigkeiten der Anbieter von Vermittlungsdiensten, mit denen insbesondere rechtswidrige Inhalte oder Informationen, die von Nutzern bereitgestellt werden und mit den allgemeinen Geschäftsbedingungen des Anbieters unvereinbar sind, erkannt, festgestellt und bekämpft werden sollen [...]«.

Moderative Ansätze verfolgen eine Vielzahl von Zielen, die im Rahmen der Risikobewertung identifizierte, systemische Risiken adressieren. Zu diesen Zielen zählen der Schutz von Kindern und Jugendlichen, der Schutz der Integrität demokratischer Wahlprozesse und die Sicherstellung der Erkennbarkeit und Unterscheidbarkeit von generativen KI-Inhalten sowie von anderen Formen synthetischer und manipulierter Inhalte. Ein weiteres Ziel ist die Transparenz hinsichtlich der Herkunft und Authentizität von Inhalten, um den Nutzer*innen die Bewertung der Vertrauenswürdigkeit von Informationsquellen zu erleichtern.

Inhaltsmoderation kann dabei auf verschiedenen Ebenen erfolgen (Zannettou, 2021). Auf der Post-Ebene beinhaltet sie Maßnahmen wie das Entfernen schädlicher Inhalte. Auf der Benutzerebene umfasst sie Eingriffe wie Accountsperren oder die Einschränkung der Sichtbarkeit von Beiträgen, auch bekannt als Shadow Banning (Leerssen, 2023). Zusätzlich können Online-Plattformen auf Gemeinschaftsebene tätig werden, indem sie spezifische Untergemeinschaften moderieren, beispielsweise durch das Sperren von Facebook-Gruppen oder Subreddits.

Grundsätzlich lassen sich Moderationsmaßnahmen in zwei Kategorien einteilen: harte und weiche Interventionen. Harte Maßnahmen beinhalten das vollständige Entfernen von Inhalten, Nutzer*innen oder Gemeinschaften. Sowohl in Art. 35 Abs. 1g) des DSA, in Erwägungsgrund 87 DSA als auch im Verhaltenskodex zur Bekämpfung rechtswidriger Hassrede im Internet findet das Entfernen von Inhalten oder Accounts als Minderungsmaßnahme Erwähnung. Online-Plattformen verpflichten sich in dem Verhaltenskodex, die Mehrheit der gültigen Meldungen zur Entfernung rechtswidriger Hassrede innerhalb von weniger als 24 Stunden zu überprüfen und, falls erforderlich, den Zugang zu solchen Inhalten zu sperren oder entfernen. An dieser Stelle wird darauf hingewiesen, stets den Grundsatz der Meinungsfreiheit zu wahren.

Weiche Maßnahmen hingegen zielen darauf ab, Inhalte einzuordnen, ihre Sichtbarkeit einzuschränken oder Interaktionen zu begrenzen, ohne sie vollständig zu entfernen. Zu diesen Maßnahmen gehören unter anderem die Verifizierung von Inhalten und Nutzerkonten, das

Anbringen von Wasserzeichen und Labels, die Verwendung von Metadaten, kryptografischen Methoden und digitalen Fingerabdrücken sowie die Protokollierung von Inhalten.

Während diese Kategorien die Grundprinzipien der Moderation umreißen, spielt die Art und Weise, wie diese Maßnahmen umgesetzt werden, eine entscheidende Rolle. Insbesondere die zunehmende Automatisierung von Moderationsprozessen hat die Diskussion um Chancen und Herausforderungen solcher Systeme intensiviert. Automatisierte Moderation kann für Online-Plattformen eine Entlastung darstellen, birgt jedoch auch Risiken. Studien zeigen, dass automatisierte Systeme aufgrund unzureichenden Trainings Bilder aus der realen Welt oft nicht korrekt erkennen und daher falsch einstufen (Taş et al., 2023). Dies kann bestehende Verzerrungen verstärken und eine faire Moderation beeinträchtigen. Auch menschliche Moderator*innen sind nicht vor Fehlern gefreit und können daher ebenfalls für Fehlentscheidungen verantwortlich sein. Deshalb ist es wichtig, dass Online-Plattformen ihre Moderationsprozesse transparent gestalten.

Design/Nutzererfahrung

Die Kategorie Design/Nutzererfahrung umfasst Maßnahmen zur Minderung systemischer Risiken aus der Perspektive der Plattformnutzer*innen und fokussiert sich auf sichtbare, für die Nutzer*innen wahrnehmbare Maßnahmen. Dazu zählt beispielsweise die Bereitstellung einer Funktion, die es den Nutzer*innen des Dienstes ermöglicht anzugeben, dass Inhalte von generativer KI stammen oder es sich um andere Arten von synthetischen und manipulierten Inhalten handelt. Mit dem Ziel, vor Desinformation und deren Verbreitung zu schützen, ermöglichen diese Maßnahmen den Online-Plattformen sicherzustellen, dass die Nutzer*innen sinnvolle Wahlmöglichkeiten und Kontrolle über die ihnen angezeigten Inhalte haben. Auch einfache und altersgruppengerechte Meldefunktionen von rechtswidrigen Inhalten zählen zu dieser Kategorie.

Bildung/Informationen

Im Kontext des DSA zielen Bildung und die Bereitstellung von Informationen als Risikominderungsmaßnahmen darauf ab, Nutzer*innen sowohl mit den nötigen Informationen zu versorgen als auch ihre Medienkompetenz so zu stärken, dass sie zwischen Desinformation und verlässlichen Informationen unterscheiden können. Das ist besonders im Wahlkontext entscheidend, um die Integrität demokratischer Prozesse zu schützen.

Hierzu können Online-Plattformen Informationen in einen passenden Kontext einbetten, beispielsweise durch die Kennzeichnung offizieller Accounts oder die Bereitstellung ergänzender Informationen über die Herkunft oder das Sponsoring von Inhalten. Entsprechend schlägt die Europäische Kommission in den Leitlinien zur Minderung systemischer Risiken in Wahlprozessen vor, Banner, Links und Pop-ups auf den Online-Plattformen zu integrieren, die kontextualisierte Informationen zur Wahl bereitstellen. Zudem können Anreize für Nutzer*innen geschaffen werden, Inhalte gründlich zu lesen und deren Richtigkeit sowie Quellen zu prüfen, bevor sie diese weiterverbreiten.

Die Leitlinien der Europäischen Kommission zur Minderung systemischer Risiken in Wahlprozessen schlagen zusätzlich vor, In-App-Funktionen zur Förderung der Medienkompetenz einzuführen, um Nutzer*innen gegenüber erwarteten Desinformationsnarrativen und Manipulationstechniken widerstandsfähiger zu machen. Die Leitlinien empfehlen spezifische Maßnahmen wie spielerische Interventionen und Videos, um kritisches Denken der Nutzer*innen zu fördern.

Obwohl die Einführung von Medienkompetenz-Initiativen grundsätzlich sinnvoll erscheint, zeigen Studien jedoch, dass deren Effekt auf die Erkennung von Desinformation oft moderat ausfällt und sie tendenziell auch allgemeine Skepsis gegenüber sämtlichen Informationen fördern (Guess et al., 2020). Um die Wirkung solcher Initiativen zu verbessern, schlägt eine Studie vor, Medienkompetenzansätze mit Maßnahmen zur Steigerung der Selbstwirksamkeit zu kombinieren (Ferrucci & Hopp, 2023). Diese Kombination könnte einen stärkeren Effekt auf die Erkennung und Eindämmung von Desinformation erzielen. Die Studie legt dar, dass Social-Media-Plattformen wie Facebook gezielte Selbstwirksamkeitsinterventionen implementieren könnten, um Nutzer*innen aktiv bei der Erkennung von Desinformation zu unterstützen. Es wäre vergleichsweise einfach umsetzbar, so die Autor*innen der Studie, wenn eine Plattform wie Facebook allen Nutzer*innen beim ersten Login eine Selbstwirksamkeitsintervention bereitstellen würde. Diese könnte darauf hinweisen, dass Fehlinformationen auf der Plattform weit verbreitet sind, und

den Nutzer*innen gleichzeitig das Vertrauen vermitteln, solche Inhalte sicher zu identifizieren.

Jugendschutz

Kinder und Jugendliche sind auf Online-Plattformen besonders gefährdet, weshalb sie besonderen Schutz genießen müssen. Ebenso wie der Jugendschutz insgesamt darauf abzielt, das Wohl von Kindern und Jugendlichen zu fördern und möglichen Schäden vorzubeugen (Schneider & Toyka-Seid, 2024), verfolgen auch alle Maßnahmen in dieser Kategorie das Ziel, eine gesunde Entwicklung junger Nutzer*innen zu gewährleisten.

Im Zentrum dieser Maßnahmen steht der Schutz von Kindern und Jugendlichen vor körperlicher, sexueller und seelischer Gewalt auf Online-Plattformen (BMFSFJ, 2024). Dazu gehört unter anderem eine kindgerechte Gestaltung der Benutzeroberflächen, die es ermöglicht, Missbrauch einfach zu melden und im Bedarfsfall Unterstützung zu erhalten. Der DSA sieht darüber hinaus vor, dass Altersverifikations- und Elternkontroll-Tools den Jugendschutz ergänzen und so zur sicheren Nutzung der Online-Plattformen beitragen. Zudem entwickelt die Europäische Kommission die bereits erwähnten Leitlinien zum Schutz von Minderjährigen im Internet unter dem DSA (Europäische Kommission, 2024d). Diese sollen im Sommer 2025 in Kraft treten und festlegen, wie Online-Plattformen Datenschutz, Sicherheit und Schutz für Minderjährige im Internet umsetzen müssen. Hier werden vermutlich weitere Risikominderungsmaßnahmen Erwähnung finden, die spezifisch auf den Kinder- und Jugendschutz ausgelegt sind.

Geschäftsbedingungen

Allgemeine Geschäftsbedingungen sind vorformulierte Vertragsklauseln, die für eine Vielzahl von Verträgen gelten. Im Fall von VLOPs und VLOSEs gelten diese also zwischen Anbieter*innen und Nutzer*innen. Die Geschäftsbedingungen von VLOPs und VLOSEs dienen dazu, Transparenz über den angebotenen Service zu gewährleisten und klare Regeln festzulegen. Diese Bedingungen müssen verständlich darlegen, welche Regeln auf der Plattform gelten und welche Maßnahmen zu deren Durchsetzung eingesetzt werden. Aus diesem Grund sind viele der zuvor genannten Risikominderungsmaßnahmen ausdrücklich in den Geschäftsbedingungen verankert.

Art. 35 DSA nennt die Anpassung und Durchsetzung der allgemeinen Geschäftsbedingungen als Risikominderungsmaßnahme, um systemischen Risiken bestmöglich entgegenzuwirken. Beispielsweise kann die Erstellung und Verbreitung rechtswidriger Inhalte eingedämmt werden, indem klare Informationen zu den spezifischen Unternehmensregeln sowie zu den Gemeinschaftsrichtlinien zur Steuerung von Verhalten und Interaktion auf den Online-Plattformen bereitgestellt werden. So können verständlich erklärte Melde- und Benachrichtigungsprozesse eine einfache Meldung und Kennzeichnung von Inhalten ermöglichen, die zu Gewalt aufrufen oder hasserfülltes Verhalten fördern.

Interne Prozesse

In die Kategorie »interne Prozesse« fallen alle Risikominderungsmaßnahmen, die innerhalb der Organisation der Anbieter*innen implementiert werden können. Sie zielen darauf ab, dass Plattformbetreiber*innen durch optimierte Abläufe und klare Verantwortlichkeiten systemische Risiken frühzeitig erkennen, kontrollieren und effektiv abmildern können. Hierfür bietet es sich an, dass Online-Plattformen sich an bewährten Verfahren orientieren. Risikominderungsmaßnahmen dieser Kategorie umfassen beispielsweise regelmäßige Trainings für Mitarbeiter*innen der Online-Plattformen, inklusive der höheren Führungsebene, um diese bestmöglich auf systemische Risiken und deren Erkennung und Eindämmung vorzubereiten. Auch das Erstellen einer Übersicht der an der Arbeit zu systemischen Risiken beteiligten Stakeholder*innen kann als Minderungsmaßnahme in Betracht gezogen werden. Eine weitere Maßnahme ist das Entwickeln, Abstimmen und Testen interner Prozesse, unter anderem durch sogenannte »Red Teaming« Übungen[1].

1 »Red Teaming« ist ein Verfahren zur Überprüfung der Wirksamkeit von Cybersicherheitsmaßnahmen, bei dem ethische Hacker einen simulierten, nicht-destruktiven Cyberangriff durchführen (IBM, 2024). Diese simulierte Attacke spiegelt der Organisation, in diesem Fall der Plattform, vorhandene Schwachstellen im System und in der Infrastruktur wider, sodass gezielte Verbesserungen in den Sicherheitsabläufen vorgenommen werden können.

Algorithmische Systeme

Ziel der Anpassung algorithmischer Systeme ist es, deren Transparenz und Fairness zu gewährleisten. Konkret bedeutet dies, dass Nutzer*innen Kontrolle über die ihnen angezeigten Inhalte erhalten und die Verbreitung von Desinformationen sowie gefälschter oder radikalisierender Inhalte algorithmisch eingedämmt werden. Auf diese Weise werden Medienpluralismus und Inhaltsvielfalt sowie die Selbstbestimmung der Nutzer*innen über die ausgespielten Inhalte gefördert. Unter Minderungsmaßnahmen dieser Kategorie fällt beispielsweise Transparenz in Bezug auf die Gestaltung und Funktionsweise von Empfehlungssystemen, insbesondere hinsichtlich der Daten und Informationen, die beim Systementwurf verwendet werden. So wird eine externe Überprüfung und Forschung durch Dritte ermöglicht. Auch Werbesysteme können so angepasst werden, dass risikobehaftete Inhalte, insbesondere in der politischen Werbung, nicht gefördert werden. Auf diese Weise soll verhindert werden, dass Entscheidungen der Nutzer*innen, vor allem im Hinblick auf Wahlen, durch algorithmische Systeme beeinflusst werden können[2].

Kooperation mit anderen Online-Plattformen, Behörden und zivilgesellschaftlichen Organisationen

Der DSA sowie Erwägungsgrund 90 DSA nennen zudem die Kooperation mit anderen Online-Plattformen, Behörden und zivilgesellschaftlichen Organisationen als eine Maßnahme, um systemische Risiken zu mindern. Diese Zusammenarbeit kann durch den Austausch von Guten Praktiken und gemeinsamen Initiativen geprägt sein. So könnte beispielsweise eine Kooperation mit lokalen Faktenprüfer*innen während Wahlperioden dazu beitragen, dass Faktencheck-Labels in der gesamten EU und in allen Sprachen verfügbar gemacht werden. Auch die enge Zusammenarbeit zwischen Online-Plattformen und vertrauenswürdigen Hinweisgeber*innen (engl. Trusted Flagger) fällt in diese Kategorie. Ebenso ist das zuvor erwähnte »Red Teaming« nicht nur ein Bestandteil der Kategorie algorithmischer Systeme, sondern auch der Kooperationskategorie, da diese Testung meist von externen Expert*innen, sogenannten ethischen Hacker*innen, durchgeführt wird und hierfür ein enger Austausch zwischen beiden Parteien erforderlich ist.

5.2 Wahl und Bewertung der Minderungsmaßnahmen

Art. 35 DSA spezifiziert nicht nur, welche Maßnahmen zur Risikominderung ergriffen werden sollen, sondern auch, wie diese Maßnahmen im Verhältnis zu systemischen Risiken stehen sollten: »Die Anbieter sehr großer Online-Plattformen und sehr großer Online-Suchmaschinen ergreifen angemessene, verhältnismäßige und wirksame Risikominderungsmaßnahmen, die auf die gemäß Art. 34 ermittelten besonderen systemischen Risiken zugeschnitten sind, wobei die Auswirkungen solcher Maßnahmen auf die Grundrechte besonders zu berücksichtigen sind.«

Die Forderung nach angemessenen und verhältnismäßigen Risikominderungsmaßnahmen orientiert sich an dem in den EU-Verträgen verankerten Grundsatz der Verhältnismäßigkeit (Art. 5 Abs. 4 Vertrag über die Europäische Union), der Unionsorganen vorschreibt, dass ergriffene Maßnahmen drei Anforderungen erfüllen müssen:

- Sie müssen **geeignet** sein, um das angestrebte Ziel zu erreichen;
- Sie müssen **erforderlich** sein, das angestrebte Ziel zu erreichen;
- Sie dürfen Einzelpersonen im Verhältnis zum angestrebten Ziel nicht übermäßig belasten (**Verhältnismäßigkeit im engeren Sinne**).

Im Kontext des DSA ergreifen zwar nicht Unionsorgane, sondern private Akteur*innen (VLOPS & VLOSEs) Maßnahmen zur Risikominderung, gleichwohl sollen gemäß Art. 35 DSA auch diese Maßnahmen dem Grundsatz der

2 2024 wurden zudem neue Vorschriften für politische Werbung beschlossen wurden: Im Februar 2024 wurde die Verordnung über die Transparenz und das Targeting politischer Werbung (2024/900) verabschiedet. Die Verordnung beinhaltet nicht nur eine einheitliche Definition über politische Werbung, sondern schreibt auch vor, dass politische Werbung klar gekennzeichnet werden muss und Nutzer*innen nur dann gezielt angesprochen werden dürfen, wenn sie ihre Zustimmung zur Erhebung ihrer personenbezogenen Daten gegeben haben. Auch verbietet die Verordnung das Sponsoring von Wahlen von außerhalb der EU in den drei Monaten vor der Wahl.

Verhältnismäßigkeit entsprechen. Konkret bedeutet dies, dass Anbieter*innen die Verhältnismäßigkeit der ergriffenen Risikominderungsmaßnahmen hinsichtlich ihres angestrebten Ziels, der Reduktion des identifizierten systemischen Risikos, prüfen sollen. Zunächst ist zu prüfen, ob die Maßnahmen zur Minderung des systemischen Risikos geeignet sind. Geeignet bedeutet hier, dass die Risikominderungsmaßnahmen das Ziel tatsächlich fördern oder erreichen können.

Neben der Eignung ist nach dem Grundsatz der Verhältnismäßigkeit auch zu prüfen, ob die Maßnahme erforderlich ist. Erforderlichkeit bedeutet, dass unter den geeigneten Maßnahmen zur Risikominderung diejenige ausgewählt werden muss, die das übergeordnete Ziel erreicht und dabei die geringstmögliche Beeinträchtigung anderer schützenswerter Interessen verursacht. In diesem Kontext heißt dies, dass, wenn eine alternative Maßnahme existiert, die das Risiko mindert und gleichzeitig weniger stark in andere schützenswerte Interessen (z.B. Datenschutz oder Privatsphäre) eingreift, diese Alternative bevorzugt werden sollte.

Zudem wird im Rahmen der Verhältnismäßigkeit im engeren Sinne geprüft, ob das Ziel der Maßnahme die möglichen negativen Auswirkungen auf andere schützenswerte Interessen, insbesondere auf die Grundrechte, rechtfertigt. Dies bedeutet: Je größer der Eingriff der Risikominderungsmaßnahme in ein anderes, schützenswertes Interesse ist, desto wichtiger muss der Nutzen für das übergeordnete Ziel sein, um diesen Eingriff zu rechtfertigen. Dieses dritte Element des Verhältnismäßigkeitsgrundsatzes, wird durch die zusätzliche Nennung von »Angemessenheit« und dem Zusatz, dass »die Auswirkungen solcher Maßnahmen auf die Grundrechte besonders zu berücksichtigen sind« besonders hervorgehoben. Der DSA gibt der Meinungsfreiheit eine herausgehobene Bedeutung (Erwägungsgründe 86, 90 DSA).

Bei der Einführung von Risikominderungsmaßnahmen sollte somit abgewogen werden, inwiefern gegebenenfalls auch die Einführung von Maßnahmen selbst ein »Zweitrisiko« darstellen könnte. Darauf wurde auch in mehreren Expertengesprächen hingewiesen. Sie verweisen beispielsweise darauf, dass die Einführung einer Verifizierung des Alters der Nutzer*innen als Minderungsmaßnahme von systemischen Risiken hinsichtlich des Schutzes öffentlicher Gesundheit im Konflikt mit dem Schutz persönlicher Daten stehen könnte. Eine Maßnahme wird daher nur dann als verhältnismäßig betrachtet, wenn es keine alternative Maßnahme gibt, die eine vergleichbare oder höhere positive Gesamtwirkung erzielt, ohne dabei andere wichtige Grundrechte und Prinzipien unverhältnismäßig zu beeinträchtigen.

Zusätzlich zum Gebot der Verhältnismäßigkeit sollen die von Anbieter*innen ergriffenen Risikominderungsmaßnahmen gemäß Art. 35 DSA »wirksam« sein. Wann genau eine Maßnahme als wirksam eingestuft werden kann, ist im DSA nicht weiter definiert und kann, anders als Erforderlichkeit und Geeignetheit, erst nach der Einführung der Maßnahme entschieden werden. Innerhalb dieser Studie wird eine Maßnahme dann als wirksam bewertet, wenn zwischen Zeitpunkt ***t0*** und Zeitpunkt ***t1*** eine Minderung des systemischen Risikos festgestellt werden kann.

Die empirische Bewertung der Wirksamkeit von Risikominderungsmaßnahmen birgt Herausforderungen. Erstens können Maßnahmen sich in ihrer Wirkung überschneiden, mehrere Risiken gleichzeitig adressieren oder sogar neue Risiken hervorrufen. Zusätzlich sind einige Risiken, wie negative Auswirkungen auf die physische und mentale Gesundheit der Nutzer*innen, nicht nur vom Gebrauch einer Plattform abhängig, sondern auch von externen Faktoren, die in der Bewertung der Maßnahme nicht berücksichtigt werden können. Schließlich kann eine effektive Maßnahme ein Risiko mindern, bevor es zu einem Schadenfall kommt, was die Zuordnung der Wirksamkeit erschwert. In solchen Fällen bleibt unklar, ob das Ausbleiben des Schadens auf die Maßnahme zurückzuführen ist oder ob der Schadenfall unabhängig davon nicht eingetreten wäre.

Zweitens besteht aufgrund der Komplexität des Untersuchungsgegenstandes sowie des begrenzten Datenzugangs zu Forschungsdaten auch in der wissenschaftlichen Literatur bei einigen Risikominderungsmaßnahmen kein Konsens darüber, inwiefern eine Maßnahme als wirksam eingestuft werden kann. Welche Kriterien erfüllt sein müssen, um wirksam Risiken zu mindern und gleichzeitig nicht zu stark in andere, schützenswerte Rechtsgüter einzugreifen, wird daher auch in den kommenden Jahren ein zentraler Gegenstand der Forschung durch Wissenschaftler*innen und Zivilgesellschaft sein.

Drittens birgt die Einführung von Risikominderungsmaßnahmen Zweitrisiken für andere schützenswerte Rechtsgüter. So sind Maßnahmen wie Deplatforming – also das Entfernen oder Sperren von Inhalten, Accounts oder Gruppen von Online-Plattformen, um deren Reichweite und Einfluss nach wiederholtem Regelverstoß zu begrenzen zwar ein zentrales Mittel, um etwa die Verbreitung rechtswidriger Inhalte zu mindern. Sie können gleichzeitig aber auch einen Einschnitt in das Recht auf freie Meinungsäußerung darstellen. Auch die politische und soziale Dimension, die in diesem Zusammenhang berücksichtigt werden muss, kann eine Rolle dabei spielen, wie eine Maßnahme hinsichtlich ihrer Wirksamkeit bewertet werden kann. Zivilgesellschaft und Wissenschaft werden in Zukunft weiterhin eine zentrale Rolle dabei spielen, diese Herausforderungen zu adressieren und Metriken zur Bewertung herauszuarbeiten.

In Bezug auf das Vorherige ist jedoch zu sagen, dass für eine rechtliche Beurteilung allein entscheidend ist, ob eine Plattform ihrer Sorgfaltspflicht nachgekommen ist. In diesem Kontext spielt eine mögliche Überkompensation wie etwa die Durchführung von Risikominderungsmaßnahmen, obwohl kein Schadenfall eintritt – keine Rolle. Maßgeblich ist, ob die Plattform alle angemessenen und verhältnismäßigen Maßnahmen ergriffen hat, um Risiken zu mindern und ihrer Verantwortung gerecht zu werden.

Vor dem Hintergrund der schwierigen Herleitung von Wirksamkeit sind die folgenden Fallbeispiele als mögliche Lösungswege zu verstehen, die einen Einblick geben, wie die Wirksamkeit von Risikominderungsmaßnahmen auf Online-Plattformen bewertet werden könnte. Dabei gehen wir davon aus, dass der Prozess der Risikoidentifizierung und Zuordnung der Risikominderungsmaßnahme nach dem Grundsatz der Verhältnismäßigkeit bereits erfolgt ist und Anbieter*innen eine verhältnismäßige und erforderliche Risikominderungsmaßnahme implementiert haben.

Um zu veranschaulichen, wie die Wirksamkeit von Maßnahmen bewertet werden kann, stellen wir exemplarisch in drei Fallstudien unterschiedliche Risikominderungsmaßnahmen und die Bewertung ihrer jeweiligen Wirksamkeit vor. Die Fallstudien orientieren sich an den drei Risikodimensionen der Verhaltens-, Inhalts- und Nutzungsebene.

Fallbeispiel zur Minderung von nutzungsbezogenen Risiken: Live-Banner zur Nutzungsdauer und Inhalten

Potenzielle Online-Schäden, die von der technischen Infrastruktur der Plattform ausgehen, fallen nach den hier definierten Risikoebenen in die Kategorie der nutzungsbezogenen Risiken. Diese Kategorie umfasst u.a. alle tatsächlichen oder absehbaren nachteiligen Auswirkungen auf den Schutz der öffentlichen Gesundheit in Bezug auf die Gestaltung, die Funktionsweise oder die Nutzung von VLOPs und VLOSEs. So könnte das Design von Empfehlungsalgorithmen zu einem Abhängigkeitsverhalten bei den Nutzer*innen führen (Bhargava & Velasquez, 2021).

Wie schon in der Minderungskategorie Algorithmische Systeme beschrieben, zielen Anpassungen algorithmischer Systeme darauf ab, Transparenz und Fairness zu gewährleisten, indem sie Nutzer*innen die Kontrolle über die ihnen angezeigten Inhalte übertragen und die Verbreitung von schädigenden Inhalten eindämmen. Gleichzeitig sollen sie Pluralismus und Inhaltsvielfalt fördern. Die Messung der Wirksamkeit von implementierten Risikominderungsmaßnahmen auf der Nutzungsebene ist insofern herausfordernd, als dass diese ein Plattformdesign erfordern, das Anpassungen der algorithmischen Systeme unter Zustimmung der Nutzer*innen zulässt und gegebenenfalls für die Testung eine Nachmodellierung der Funktionsweise der Plattform erfordert (Metzler & Garcia, 2024), was wiederum methodisch limitiert ist. So konnten Studien zum sogenannten »Amplification Paradox« nachweisen, dass nachgebildete Accounts sich radikalisieren, reale Nutzer*innen jedoch weniger über Empfehlungsalgorithmen radikalere Inhalte konsumieren, sondern eher über andere Online-Plattformen oder Webseiten zu diesen Inhalten gelangen (Ribeiro et al., 2023b). Dennoch bietet die Nachmodellierung der Nutzungsoberfläche von Online-Plattformen methodisch Möglichkeiten, um die Wirksamkeit von Minderungsmaßnahmen auf der Nutzungsebene zu beleuchten, was im Folgenden beschrieben wird.

Online-Plattformen bieten Nutzer*innen die Möglichkeit, sich zwischen zwei Formaten der Inhalteausgabe zu entscheiden. Sie können in der Regel auswählen,

ob sie empfohlene Inhalte, ausgespielt bekommen möchten, die algorithmisch kuratiert werden, oder nur Inhalte von Konten, denen die/der Nutzer*in folgt. Beide Formate der Inhalteausgabe bergen das Risiko für exzessives Scrolling. Für die zweite Möglichkeit der Inhalteausgabe haben Forscher*innen daher als Risikominderungsmaßnahme ein Banner entwickelt, das Nutzer*innen signalisiert, dass sie alle neuen Inhalte bereits konsumiert haben (Zhang et al., 2022). Das soll verhindern, dass Nutzer*innen in ein exzessives Scrolling verfallen und ihnen bereits bekannte Inhalte mehrfach lesen.

Für die Messung der Wirksamkeit einer solchen Funktion haben die Autor*innen der Studie eine Software für X-nutzer*innen entwickelt, die die X-Umgebung nachmodelliert. In einer ersten Umfrage haben die Wissenschaftler*innen erfasst, wie Nutzer*innen der X-App ihr Handlungsempfinden bei der Nutzung der App wahrnehmen. Im Anschluss daran entwickelten die Forscher*innen auf Grundlage eines Design-Workshops zusammen mit Nutzer*innen Funktionen, die darauf abzielen, die Nutzungserfahrung zu verbessern. Unter diesen Funktionen war auch die Erinnerung, dass Nutzer*innen Inhalte bereits konsumiert haben. Die nachgebildete Nutzungsoberfläche wurde anschließend von Nutzer*innen über einen Zeitraum von vier Wochen getestet und die Erfahrung der Teilnehmenden mit einer abschließenden Umfrage ausgewertet.

Die Ergebnisse zeigten auf, dass die Nutzungszeit mit der Einführung der neuen Funktion zwar nicht verringert wurde. Dennoch haben Teilnehmer*innen das Gefühl vermittelt bekommen haben, mehr Kontrolle über die Zeit zu haben, die sie auf der nachgebildeten Nutzungsoberfläche verbringen im Vergleich zur X-Oberfläche. Insbesondere führte das Banner dazu, dass die Teilnehmer*innen reflektierten, wie viel Zeit sie bereits auf der Plattform verbracht haben. In der Auswertung der Umfragedaten kamen die Wissenschaftler*innen zu dem Schluss, dass diese Maßnahme wirksames Instrument sein kann, um das Risiko der Abhängigkeit zu mindern, das sich beispielsweise durch exzessives Scrolling äußert.

Fallbeispiel zur Minderung von inhaltsbezogenen Risiken: Die Wirkung von Deplatforming

Inhaltsbezogene Risiken ergeben sich aus der Erstellung, Verbreitung oder Verstärkung rechtswidriger Inhalte, die über Online-Plattformen hinaus zu gesellschaftlichen Schäden oder Rechtsverletzungen führen können. Ein systemisches Risiko dieser Kategorie ist die Verbreitung rechtswidriger Hassrede. Eine Maßnahme zur Minderung dieses Risikos ist das Aussetzen der Erbringung der Dienste nach vorheriger Warnung für einen angemessenen Zeitraum für Nutzer*innen, »die häufig und offensichtlich rechtswidrige Inhalte bereitstellen« (Art. 23, Abs. 1 DSA). Diese Risikominderungsmaßnahme wird Deplatforming genannt und zählt zur Kategorie der (harten) Moderation. Die Maßnahme wirkt insbesondere systemischen Risiken entgegen, die mit rechtswidrigen Inhalten auf Online-Plattformen einhergehen, indem sie gezielt Nutzer*innen oder Inhalte von der Plattform ausschließt, die nachweislich und wiederholt rechtswidrige Inhalte verbreiten. Die Wirksamkeit von Deplatforming zeigt sich darin, dass die Präsenz rechtswidriger Inhalte auf Online-Plattformen verringert wurde.

Jhaver et al. untersuchen in ihrer Studie (2021) die Wirksamkeit von Deplatforming anhand der Exklusion von drei prominenten Accounts von der damaligen Plattform Twitter. Hierbei handelt es sich um Influencer, die regelmäßig rechtswidrige Inhalte verbreiteten. Methodisch untersuchten die Autor*innen Tweets mit Nennung der Influencer in den sechs Monaten vor und den sechs Monaten nach deren Deplatforming. Der untersuchte Datensatz enthielt insgesamt 49 Millionen Tweets. In ihrer Analyse wurden folgende Metriken zur Messung der Wirksamkeit von Deplatforming verglichen: Zur Messung der Konversation über die Influencer wurden das Aktivitätslevel beim Posting pro Influencer erfasst, die Anzahl von einmaligen Nutzer*innen, die über die Influencer schreiben, und die Anzahl neuer Nutzer*innen, die zu den Influencern posten. Zur Messung der Verbreitung der Ideen, die mit den Influencer*innen in Verbindung gebracht werden, wurde die Anzahl der Tweets erfasst, die diese Ideen nennen, sowie die Anzahl einmaliger Nutzer*innen, die Bezug auf diese Ideen nehmen. Darüber hinaus wurde zudem die Aktivität

der Nutzer*innen anhand ihrer veröffentlichten Posts registriert.

Durch einen Vergleich der erhobenen Daten vor und nach dem Deplatforming konnten die Autor*innen nachweisen, dass das Deplatforming von Influencern, die regelmäßig rechtswidrige Inhalte verbreiten, nicht nur zu einer Reduktion der rechtswidrigen Inhalte dieser Influencer auf der Plattform selbst führte, sondern auch dazu beitrug, dass die Beiträge ihrer Anhängerschaft sowohl im Umfang als auch in ihrer Toxizität zurückgehen. Die Autor*innen schlussfolgern, dass die Popularität der Ideen, die hinter den rechtswidrigen Inhalten stehen, durch Deplatforming abklingen. Zu einem ähnlichen Schluss kommen auch Rauchfleisch und Kaiser (2021) in ihrer Untersuchung zu Deplatforming von Rechten auf YouTube und BitChute.

Fallbeispiel zur Minderung von verhaltensbezogenen Risiken: Labelling von Desinformation

Verhaltensbedingte Risiken können entstehen, wenn Akteur*innen die Schwachstellen oder Nutzungsbedingungen von Online-Plattformen ausnutzen und dadurch Risiken verursachen, indem sie rechtswidrige oder schädliche Aktivitäten ausüben. Ein solches Risiko stellte beispielsweise während der Corona-Pandemie die systematische Verbreitung von Desinformationen auf Online-Plattformen dar (Tasnim et al., 2020). Diese schürten Skepsis gegenüber wissenschaftlich fundierten Maßnahmen wie Impfungen, wurden teils millionenfach geteilt und beeinflussten so die gesellschaftliche Debatte erheblich. In unserem Fallbeispiel stellt dies insofern ein systemisches Risiko dar, als dass es schwerwiegende Auswirkungen auf die physische und psychische Gesundheit von Nutzer*innen haben kann. Zur Minderung dieses Risikos hat die Plattform unter dem Grundsatz der Verhältnismäßigkeit das Labeling von Inhalten als geeignete Minderungsmaßnahme ausgewählt. Labeling, also die Kennzeichnung von Inhalten mit Warnhinweisen oder zusätzlichen Informationen, dient dazu, Nutzer*innen auf potenziell schädliche oder irreführende Inhalte aufmerksam zu machen.

Labeling als Instrument wird vor allem in den Leitlinien für die Minderung systemischer Risiken für Wahlen als wichtige Minderungsmaßnahme genannt und fällt in die Kategorie der Moderation (weiche Moderation). Das übergeordnete Ziel von Labeling als Minderungsmaßnahme ist es, wie auch bei anderen Maßnahmen in dieser Kategorie, die Verbreitung schädlicher Inhalte oder falscher Informationen effektiv einzudämmen und den Zugang zu zuverlässigen Informationen zu fördern.

Die Herausforderung in der Bewertung der Wirksamkeit auf der Verhaltensebene besteht darin, nachzuweisen, dass Maßnahmen auf systemischer Ebene zur Reduktion von bestimmten Inhalten führen. Studien zeigen, dass die Effektivität von Labels u.a. von dem Vertrauen in die Faktenprüfer*innen sowie individuellen Eigenschaften wie der politischen Ausrichtung der Nutzer*innen abhängig ist (X. Liu et al., 2023). Dies verdeutlicht die Schwierigkeit, Minderungsmethoden isoliert hinsichtlich ihrer Wirksamkeit zu bewerten. Gleichzeitig kann die Wirksamkeit durch das Heranziehen von leicht zugänglichen Metriken wie Sichtbarkeit von Inhalten gut erfasst werden. Die Herangehensweise an die Messung der Effektivität von Labelling wird im Folgenden im Kontext von Desinformation während der COVID-19-Pandemie untersucht.

Leicht et al. beleuchten in ihrer Forschung (2022) die Wirksamkeit der Labels, die Twitter und Facebook während der Pandemie eingeführt haben. Zur Bewertung dieser Wirksamkeit ziehen sie unter anderem einen Vergleich der Verbreitung falscher Informationen auf Facebook vor und nach der Einführung der Maßnahme heran. Ihre Ergebnisse zeigen, dass bereits der einfache Hinweis, einen Beitrag als Fehlinformation zu kennzeichnen, die Weiterverbreitung solcher Inhalte spürbar reduzieren kann. Ein Vergleich zwischen Facebook und X verdeutlicht außerdem, dass Facebook bei der Kennzeichnung von COVID-19-Fehlinformationen auf seiner Plattform effektiver agierte als Twitter. Das machen die Autor*innen daran fest, dass Twitter Inhalte mit falschen Informationen, die von Facebook als schädlich gekennzeichnet werden, nicht kennzeichnete.

Einen alternativen Ansatz zur Bewertung der Wirksamkeit stellt der Mixed-Methods-Ansatz dar, den Geeng et al. (2020) verwenden. Sie untersuchen die Effektivität von Warnlabels im Zusammenhang mit

COVID-19, indem sie Nutzer*innen direkt nach ihrer Einschätzung zur Wirksamkeit dieser Maßnahmen befragen. So fanden sie heraus, dass Teilnehmer*innen zwar positiv gegenüber den Labels eingestellt sind, ihre Informationen jedoch nicht über die Plattform, sondern bspw. über eine Websuche korrigieren. Die Autor*innen argumentieren, dass Online-Plattformen noch mehr tun könnten, um die Verbreitung von COVID-19-Fehlinformationen einzudämmen. So gaben die Teilnehmer*innen der Studie beispielsweise an, dass sie häufig Fehlinformationen auf sozialen Online-Plattformen sehen, die von den Online-Plattformen nicht gekennzeichnet werden. Diese Erkenntnis unterstreicht die dringende Notwendigkeit, dass Social-Media-Plattformen sowohl die Anzahl als auch die Häufigkeit der Kennzeichnung von Fehlinformationen deutlich erhöhen.

Daraus lässt sich ableiten, dass auch die Implementierung von Minderungsmaßnahmen selbst kontinuierlich überprüft werden sollte, um deren Effektivität bestmöglich zu fördern. So zeigt eine Analyse von Ling, Gummadi und Zannettou (2024), dass TikTok im Kontext von COVID-19 fälschlicherweise Labels auf Videos anbrachte, die inhaltlich keinen Bezug zur Pandemie hatten, während gleichzeitig schädliche oder falsche Inhalte oft unmarkiert blieben. Diese Erkenntnisse wurden durch das Scraping von Videos und die Analyse einer randomisierten Stichprobe gewonnen. Die Fehler lassen sich vermutlich darauf zurückführen, dass Online-Plattformen automatisierte Verfahren einsetzen, um Inhalte als falsch zu labeln. Ein solcher Umstand kann zu neuen systemischen Risiken führen, was verdeutlicht, wie entscheidend eine regelmäßige Überprüfung der Risikominderungsmaßnahmen ist, um die Genauigkeit und Wirksamkeit der Implementierung von Minderungsmaßnahmen zu gewährleisten und kontinuierlich zu verbessern.

Fallstudie: Online-Bots

Aus dem zuvor durchgeführten Bewertungsprozess für diese Fallstudie ist bekannt, dass es sich bei Bot-Netzen um ein verhaltensbezogenes Risiko handelt. Im Risikoprofil wurde das potenzielle Risiko, dass von Online-Bots ausgeht als schwerwiegend eingestuft. Nach Art. 35 DSA sind die Dienste somit dazu verpflichtet, Risikominderungsmaßnahmen zu ergreifen. Sie sollten nach Art. 35 DSA verhältnismäßig und wirksam sein. Um die Verhältnismäßigkeit zu prüfen, gilt es, in einem ersten Schritt geeignete Maßnahmen zu identifizieren, mit dem Ziel, Bot-Netzwerke eindämmen zu können. Die von ISD erstellte Tabelle in Anhang 3 kann hilfreich sein, um eine passende Maßnahme zu identifizieren. Das zuvor identifizierte systemische Risiko der Bedrohungen im Zusammenhang mit Ausländischer Informationsmanipulation und Einflussnahme (FIMI) sowie dem umfassenderen Phänomen der Desinformation ist in der Tabelle unter Zeile 25 aufgeführt. Daher wird im Folgenden auf Risiko 25 verwiesen. Auf Grundlage der offiziellen Dokumente ergeben sich für Risiko 25 mehrere Risikominderungsmaßnahmen (s. Anhang 3), die im Folgenden ausgewertet werden.

Die erste Gruppe von Risikominderungsmaßnahmen, die anhand der Tabelle identifiziert werden konnten, fällt in die Kategorie von Bildung/Information. X könnte beispielsweise durch gamifizierte Interventionen und Videos, die auf den deutschen Kontext angepasst sind, Nutzer*innen über Inhalte aufklären, die gemäß den Gemeinschaftsrichtlinien nicht erlaubt sind. Dazu gehören auch Online-Bots. Auch könnten Medienkompetenzkampagnen kritisches Denken fördern und Nutzer*innen dabei unterstützen, Inhalte zu hinterfragen und ggf. nicht weiter zu teilen. In diesem Fall würden sie dabei unterstützt werden, Bots und die von Bots geteilten Inhalte anhand typischer Merkmale wie Accountname, Anzahl der Follower*innen oder Accountaktivität zu erkennen.
Zudem könnte X Maßnahmen ergreifen, die sowohl in die Kategorie der Moderation als auch in die der Kooperation mit relevanten Akteur*innen fallen. So könnten Nutzer*innen etwa unterstützt werden, die Vertrauenswürdigkeit von Inhalten durch Informationen zur Bearbeitungshistorie und Faktenprüfung zu bewerten, die auf Grundlage von einer Kooperation mit Faktenprüfungsorganisationen aus dem deutschsprachigen Raum entstanden sind. Faktenprüfungskennzeichen wiederum könnten Nutzende davon abhalten, von Online-Bots veröffentlichten Desinformationen weiter zu teilen und somit deren Verbreitung und Einfluss verringern.

Weitere Minderungsmaßnahmen fallen in die Kate-

gorie der internen Prozesse. X hätte beispielsweise die Möglichkeit, spezifische Regeln aufzustellen oder anzupassen, die sich gegen die Erstellung von unauthentischen Konten oder Bot-Netzen sowie gegen die irreführende Nutzung des Dienstes richten. Zudem könnten effektive interne Maßnahmen geschaffen werden, um den Missbrauch von Faktenprüfungskennzeichen zu verhindern, insbesondere zur Verifizierung von gekennzeichneten Konten und Inhalten. Schließlich könnte X auch eine Anpassung der algorithmischen Systeme vornehmen und die Verbreitung von Desinformationen durch Online-Bots einschränken.

X hat somit im vorliegenden Fall eine breite Auswahl an Risikominderungsmaßnahmen, die ergriffen werden könnten. Im Rahmen der Verhältnismäßigkeitsprüfung sollte X in einem zweiten Schritt prüfen, inwiefern die Maßnahmen erforderlich wären und damit das übergeordnete Ziel – die Minderung der mit Online-Bots einhergehenden Risiken – erreicht wird. Gleichzeitig ist sicherzustellen, dass andere schützenswerte Rechte und Interessen so wenig wie möglich beeinträchtigt werden. Von den zur Auswahl stehenden Maßnahmen beeinträchtigen jene, die auf die Stärkung der Medienkompetenz der Nutzer*innen abzielen, andere schützenswerten Interessen am wenigsten.

Gleichwohl könnten Kampagnen zur Medienkompetenz angesichts des Risikoprofils der Bot-Netzwerke in dem vorliegenden Fall nicht ausreichen, um das Risiko einer negativen Beeinflussung der gesellschaftlichen Debatte einzuschränken. X sollte daher in diesem Fall auch überprüfen, inwiefern die Anpassung interner Prozesse und algorithmischer Systeme zur Prävention von der Erstellung unauthentischer Konten beitragen kann. Zudem gilt es zu prüfen, ob auch die Einschränkung der Verbreitung von irreführenden Inhalten eine erforderliche Maßnahme darstellt, um negative Auswirkungen der Bot-Netzwerke zu verhindern – vorausgesetzt, automatisierte Erkennungssysteme blockieren fälschlicherweise Inhalte, die nicht von einem Online-Bot, sondern von einer authentischen Person stammen, könnten dadurch schützenswerte Interessen beeinträchtigt werden. Ein solches Interesse wäre in diesem Fall etwa das Recht auf Meinungsfreiheit.

In einem letzten Schritt der Verhältnismäßigkeitsprüfung sollte daher die Verhältnismäßigkeit im engeren Sinne geprüft werden, um zu entscheiden, inwiefern eine Anpassung der algorithmischen Systeme angesichts der potenziellen Beeinträchtigung der Meinungsfreiheit gerechtfertigt ist. Sobald feststeht, welche Risikominderungsmaßnahmen von X ergriffen werden, sollte daher geprüft werden, inwiefern die ergriffenen Maßnahmen auch wirksam sind. In dem vorliegenden Fall eines nutzungsbezogenen Risikos ist die Messung der Wirksamkeit dahingehend herausfordernd, dass die Wirksamkeit von Medienkompetenzkampagnen vor allem langfristig und anhand von Umfrageforschung erfasst werden müsste. Die Wirksamkeit einer Anpassung interner technischer Prozesse könnte wiederum simulierte Tests mit Bot-Netzen erfordern. Abschließend müsste auch getestet werden, inwiefern die Maßnahmen einen kausal nachweisbaren Wirkungseffekt im Hinblick auf die Reduzierung des identifizierten Risikos im zeitlichen Verlauf zeigen. Eine solche Berechnung, die beispielsweise anhand von Kontrollgruppen durchgeführt werden könnte, ist anspruchsvoll.

6. Überlegungen zum Aufbau eines Risikofrühwarnsystems

6.1 Vorüberlegungen

In den vorangegangenen Kapiteln lag der Fokus darauf, Risiken innerhalb unseres theoretischen Rahmens von nutzungsbezogenen, inhaltsbezogenen und verhaltensbezogenen Risiken zu kategorisieren und zu zeigen, wie diese auf einer Risikomatrix positioniert werden können, um den Bedarf an Minderungsmaßnahmen zu bewerten. Dabei wurde insbesondere die Frage behandelt, wie systemische Risiken identifiziert und bewertet werden können – also solche Risiken, die aufgrund ihrer weitreichenden Auswirkungen umfassendere und koordinierte Interventionen erfordern. In diesem Kapitel wird der Schwerpunkt auf die Identifikation spezifischer Risiken gelegt, durch die Entwicklung eines Frühwarnsystems.

Das Frühwarnsystem, das hier entwickelt wird, konzentriert sich auf die Erkennung von Risiken, bevor eine Aussage über ihr Ausmaß gemacht wird. Im Gegensatz zur Frage, wann beispielsweise algorithmisch bedingte Abhängigkeit – ausgelöst durch eine bestimmte Plattformfunktion wie endloses Scrollen oder kurze Videos (als nutzungsbezogenes Risiko) – ein Maß an Wirkung (hinsichtlich des Umfangs und der Reichweite) erreicht, dass eine VLOP zur Minderung dieses Risikos verpflichtet, besteht oft die Herausforderung darin, überhaupt das Vorhandensein von Suchtverhalten zu erkennen. Obwohl der Fokus in diesem Kapitel auf der Identifikation spezifischer Risiken liegt, könnte ein solches Frühwarnsystem jedoch auch so erweitert werden, dass es das Potenzial dieser spezifischen Risiken zur Entwicklung systemischer Ausmaße bewertet. Im Folgenden wird erläutert, wie das System angepasst werden müsste, um nicht nur einzelne Risiken zu erkennen, sondern auch in der Lage zu sein, abzuschätzen, wann sich diese Risiken zu systemischen Bedrohungen entwickeln könnten, die breitere, umfassendere Eingriffe erfordern.

Für die theoretische Konzeption eines Frühwarnsystems, das spezifische Risiken identifiziert, werden konkrete Fallbeispiele vorgestellt. Die Einführung konkreter Fallbeispiele wie algorithmisch bedingte Abhängigkeit, terroristische Inhalte und koordiniertes unauthentisches Verhalten ermöglicht eine intensive Analyse spezifischer Indikatoren und Messwerkzeuge. Es erlaubt auch zu zeigen, wie diese in ein System integriert werden können.

Fortschritte im Bereich der Künstlichen Intelligenz (KI), des Natural Language Processing (NLP) und großer Sprachmodelle (engl. Large Language Models, LLM) bieten leistungsstarke Werkzeuge zur Erkennung spezifischer Risiken wie diejenigen, die in Teil 3.2 aufgeführt wurden. Dies macht die Umsetzung eines Frühwarnsystems realistisch. Wohingegen sich das Frühwarnsystem stark auf den Einsatz von Künstlicher Intelligenz (KI) stützt, liegt der Fokus nicht auf kommerziellen Werkzeugen zur Erkennung von Teilrisiken, wie etwa TrueMedia oder Sensity AI, die speziell KI-generierte Inhalte oder Deepfakes identifizieren. Vielmehr wird ein Konzept entwickelt, das sowohl theoretische als auch praktische Überlegungen zur Schaffung eines skalierbaren Systems zur Risikoerkennung umfasst.

Das System nutzt KI zur proaktiven Identifizierung spezifischer Risiken und bietet einen flexiblen Leitfaden für den Aufbau eines umfassenden Monitoring-Systems. Dabei werden auch die Herausforderungen thematisiert, die im Bereich der ethischen Aufsicht der Datenzugangsbeschränkungen und des Modelltrainings bestehen, um Verzerrungen zu vermeiden, die Privatsphäre der Nutzer*innen zu schützen und die Resilienz des Systems in einer dynamischen digitalen Umgebung zu gewährleisten.

Das Konzept eines Frühwarnsystems ist vor allem für Plattformbetreiber*innen von Interesse, die eine rechtliche Verantwortung zur Risikoerkennung haben. Darüber hinaus könnte es aber auch für verschiedene andere Akteur*innen von Bedeutung sein, wie etwa Regulierungsbehörden, Expert*innen aus Wissenschaft und Zivilgesellschaft. Für diese Gruppen stellt das System eine wertvolle Ressource dar, um Risiken frühzeitig zu erkennen, deren potenzielle systemische Auswirkungen zu bewerten und entsprechend zu reagieren.

6.2 Integration von Indikatoren, Datenquellen und Analysewerkzeugen

Um ein wirksames Frühwarnsystem für spezifische Risiken zu entwickeln, müssen Indikatoren, Datenquellen und Analysetools in einem kohärenten Rahmen integriert werden. Dieser Ansatz erfordert die Auswahl zuverlässiger Indikatoren, die das Auftreten spezifischer Risiken frühzeitig signalisieren können, die Erfassung relevanter Daten und den Einsatz geeigneter Analyseinstrumente zur Verarbeitung und Interpretation dieser Informationen. Im Folgenden skizzieren wir diesen Ansatz anhand von drei zentralen Beispielen: Algorithmisch

bedingte Abhängigkeit als nutzungsbezogenes Risiko, terrorismusbezogene Inhalte als inhaltsbezogenes Risiko und koordiniertes unauthentisches Verhalten als verhaltensbezogenes Risiko.

Die folgenden Abschnitte zeigen, wie zum Beispiel maschinelles Lernen zur Erkennung von Suchtmustern eingesetzt werden kann, die mit spezifischen Plattformfunktionen wie Autoplay und Endlos-Scrolling verknüpft sind. Indikatoren wie Interaktionshäufigkeit und wiederholte Nutzung dieser Funktionen geben erste Hinweise auf potenzielles Suchtverhalten. Maschinelle Lernmodelle und benutzerdefinierte Überwachungstools können auf solche Verhaltensmuster trainiert werden, um anhaltende, hochfrequente Nutzung zu erkennen und plötzliche Veränderungen im Engagement als Frühwarnsignale zu identifizieren. Diese Kombination aus Datenerfassungs- und Analysewerkzeugen ermöglicht eine präzise Echtzeit-Überwachung und schafft ein flexibles Instrument zur Früherkennung süchtig machender Nutzungsmuster. Datenrechtliche Bedenken eines solchen Systems werden in den weiteren Abschnitten behandelt.

Durch die Abstimmung von Indikatoren, Datenquellen und Analysemethoden schafft dieser Rahmen eine robuste Grundlage für ein Frühwarnsystem, das auf die proaktive Identifizierung spezifischer Risiken ausgerichtet ist. Dieses System könnte jedoch auch dahingehend erweitert werden, dass es bewertet, ob ein bestimmtes Risiko systemische Dimensionen erreicht, indem zusätzliche Wirkungsindikatoren integriert werden, insbesondere solche, die das Ausmaß und die Tragweite des Risikos beurteilen. In den folgenden Abschnitten werden diese Risikokategorien detaillierter beschrieben und es wird untersucht, wie jede Kategorie innerhalb eines kohärenten Überwachungssystems wirksam kontrolliert werden kann.

6.3 Indikatoren und Datenquellen für die Automatisierung

Um ein wirksames Frühwarnsystem zu schaffen, müssen die Indikatoren flexibel sein und Verhaltensweisen in verschiedenen Risikokategorien abbilden können. Unsere Fallbeispiele verdeutlichen, wie diese Indikatoren das automatisierte Monitoring risikorelevanter Inhalte und Aktivitäten auf verschiedenen Online-Plattformen unterstützen können.

Zu den zentralen Indikatoren für algorithmisch bedingte Abhängigkeit gehören zum Beispiel verlängerte Sitzungszeiten bei Funktionen wie Autoplay und Endlos-Scrolling, die zwanghafte Nutzungsmuster anzeigen können. Eine erhöhte Interaktionshäufigkeit, wie häufige Likes und Kommentare, kann auf eine soziale Abhängigkeit hinweisen – ein Merkmal von Suchtverhalten. Diesen Daten stammen aus sozialen Interaktionsprotokollen und ermöglichen es dem System, Nutzer*innen hervorzuheben, die sich übermäßig an Verhaltensmustern sozialer Bestätigung beteiligen (Suma et al., 2021). Auch das Muster des Medienkonsums gibt Aufschluss, da Nutzer*innen, die Inhalte mit hohem Aufmerksamkeitswert bevorzugen, wie Autoplay-Videos, bei längerer Nutzung Anzeichen von Abhängigkeit entwickeln können (Suma et al., 2021). Obwohl Online-Plattformen rechtlich nicht verpflichtet sind, hohes Engagement aktiv zu begrenzen, wird suchtähnliches Verhalten explizit als systemisches Risiko im DSA anerkannt, was deren Verantwortung bei der Risikominderung unterstreicht.

Im Zusammenhang mit terrorismusbezogenen Inhalten konzentrieren sich die Indikatoren auf die Identifizierung von rechtswidrigem oder schädlichem Material. Extremistische Schlüsselwörter und Symbole sind primäre Indikatoren, wobei NLP-Modelle diese durch die Erkennung von Schlüsselwörtern und kontextbezogene Analysen identifizieren, um extremistische Rhetorik und Aufrufe zur Gewalt zu erfassen (Nouh et al., 2019). Eine Netzwerkanalyse unterstützt die Erkennung weiter, indem sie die Beziehungen zwischen den Nutzer*innen untersucht, um Cluster aufzudecken, die extremistische Ideologien fördern, was auf ein organisiertes Verhalten hindeuten kann (Hohenwalde, 2023). Darüber hinaus ermöglicht die Analyse von Multimedia-Inhalten dem System, extremistische Symbole, Flaggen und Hymnen in Bildern, Videos und Audioinhalten zu erkennen, die in terroristischer Propaganda häufig vorkommen (Altitude, 2024). Zeitliche Aktivitätsmuster tragen ebenfalls dazu bei, Mobilisierungsbemühungen aufzudecken, wobei Methoden zur Erkennung von Veränderungspunkten einen signifikanten Anstieg der Aktivitäten im Zusammenhang mit extremistischen Kampagnen aufzeigen (Theodosiadou et al., 2021). Geolokalisierungsdaten, sofern verfügbar, geben zusätzliche geografische Einblicke, indem sie Konzentrationen von Aktivitäten in Hochrisikogebieten aufzeigen (TCAP, 2024).

Ein zentraler Indikator für koordiniertes unauthentisches Verhalten sind synchronisierte Postings, bei denen mehrere Konten innerhalb kurzer Zeit identische oder ähnliche Inhalte verbreiten (Gregoire, 2021). Dieses Muster kann anhand von Netzwerkdaten einschließlich Zeitstempeln identifiziert werden, die auf koordinierte Posting-Aktivitäten hindeuten. Eine einheitliche Nachrichtenübermittlung – bei den mehrere Konten identische Phrasen oder Hashtags verwenden – stellt einen weiteren Indikator dar, da sie gezielt bestimmte Erzählungen verstärkt; diese können durch Textanalyse erkannt werden (Gregoire, 2021). Auch anomale Kontoeröffnungen können auf CIB hinweisen, insbesondere wenn ein plötzlicher Anstieg neuer Konten verzeichnet wird, die ein ähnliches Verhalten aufweisen (Gregoire, 2021). Die Beobachtung von Registrierungsdaten hilft, solche Anomalien zu identifizieren. Darüber hinaus können Netzwerkmuster wie hochfrequente Interaktionen innerhalb geclusterter Konten koordinierte Netzwerke anhand von Follower- und Interaktionsdaten aufdecken (Gregoire, 2021).

Die folgende Tabelle gibt einen Überblick über die Übereinstimmung zwischen unseren Fallbeispielen von Risiken, ihren jeweiligen Indikatoren und Datenquellen:

Risikotyp	Beispiel für ein spezifisches Risiko	Indikator	Beispiel/Details	Datenquelle
Nutzungsbezogene Risiken	Algorithmische Sucht	Verlängerte Sitzungsdauer	Sitzungsdauer bei Autoplay, unendlichem Scrollen	Analysen auf Geräteebene, benutzerdefinierte Überwachungstools
		Erhöhte Interaktionshäufigkeit	Häufigkeit von Likes, Kommentaren, die auf soziale Abhängigkeit hinweisen	Daten zur sozialen Interaktion (Likes, Kommentare)
		Muster des Inhaltskonsums	Vorliebe für bestimmte Arten von Inhalten, die zu längerer Nutzung anregen	Analyse des Engagements für Inhalte, Daten zur Nutzung der Plattform
Inhaltsbezogene Risiken	Terroristischer Inhalt	Extremistische Schlüsselwörter, Symbole	Schlüsselwörter, Symbole, die mit extremistischen Gruppen in Verbindung gebracht werden	Textdaten, Metadaten (Zeitstempel)
		Sentiment-Analyse	Erkennung von extremistischer Rhetorik und Hassreden	Textuelle Analyse, NLP-Modelle
		Multimedia-Inhaltsanalyse	Identifizierung von Symbolen, Flaggen und Hymnen in Bildern, Videos und Audiodateien	Beiträge mit Bild und Ton
		Standortbezogene Indikatoren	Geolokalisierungsdaten in Verbindung mit Konfliktregionen	Metadaten, Beiträge mit Geotags, Standortdaten

Risikotyp	Beispiel für ein spezifisches Risiko	Indikator	Beispiel/Details	Datenquelle
Verhaltensbezogene Risiken	Koordiniertes unauthentisches Verhalten	Synchronisierte Buchungen	Identische Inhalte von mehreren Konten innerhalb eines kurzen Zeitrahmens	Netzdaten einschließlich Zeitstempel
		Einheitliches Messaging	Konten, die identische Phrasen oder Hashtags in ihren Beiträgen verwenden	Textuelle Analyse des Beitragsinhalts (semantische Ähnlichkeit)
		Anomale Kontoerstellung	Anstieg neuer Konten, die ähnliche Verhaltensmuster aufweisen	Kontoregistrierungsdaten, Musteranalyse
		Netzwerk-Muster	Hochfrequente Interaktionen zwischen gebündelten Konten	Follower- und Interaktionsnetzwerkdaten

Tabelle 9: Risikotypen mit Beispielen, Indikatoren, Details und Datenquellen

Dieser Beispielsatz von Indikatoren unterstreicht die Notwendigkeit von Flexibilität innerhalb eines automatisierten Monitoring-Systems mit anpassungsfähigen Metriken zur Erfassung der spezifischen Merkmale jeder Risikokategorie. Sobald die Indikatoren und Datenquellen definiert sind, können wir uns den Analysetools zuwenden, die zur Messung der Risiken verwendet werden können.

6.4 Auswahl der geeigneten Analysewerkzeuge

Die Wirksamkeit eines Frühwarnsystems für das Monitoring spezifischer Risiken hängt von der Auswahl von Analysewerkzeugen ab, die sich eng an den einzigartigen Merkmalen der einzelnen Risikotypen orientieren. Dieser Abschnitt enthält Beispiele dafür, wie maschinelles Lernen, die Verarbeitung natürlicher Sprache (z.B. NLP) und Netzwerkanalysetools flexibel auf verschiedene Risiken angewendet werden können. Diese Beispiele sind eher illustrativ als erschöpfend und sollen anpassungsfähige Methoden aufzeigen, die für verschiedene systemische Risiken geeignet sind.

Bei dem Monitoring nutzungsbezogener Risiken wie Sucht sind Modelle des maschinellen Lernens nützlich, um Muster im Nutzerverhalten zu erkennen, die auf ein übermäßiges oder zwanghaftes Engagement hindeuten. Diese Modelle analysieren Daten über Nutzerinteraktionen, wie z.B. die Häufigkeit und Dauer von Aktivitäten bei Funktionen wie Autoplay-Videos und endlosem Scrollen, um Anzeichen für potenziell süchtiges Verhalten zu erkennen. Durch die Erkennung eines plötzlichen Anstiegs der Nutzung oder eines ungewöhnlich hohen Maßes an Interaktion können diese Tools frühe Anzeichen für eine Sucht erkennen. Studien haben gezeigt, dass Algorithmen des maschinellen Lernens Nutzer*innen mit hohem Risiko für süchtiges Verhalten auf der Grundlage dieser Metriken effektiv klassifizieren können (Kuo, 2024). Solche Modelle helfen dabei, einen proaktiven Rahmen für die Erkennung von Suchtmustern zu schaffen, bevor sie eskalieren.

Bei inhaltsbezogenen Risiken, wie der Erkennung extremistischer oder terroristischer Inhalte, spielen NLP-Werkzeuge eine zentrale Rolle in der Analyse und Verarbeitung von Textdaten. NLP-Modelle, einschließlich Sprachmodelle (z.B. Transformer-Modelle), werden darauf trainiert, sprachliche Muster zu erkennen, etwa aggressive Tonlagen oder extremistische Symbole, indem der Kontext und die Verwendung spezifischer Begriffe untersucht werden. Durch die Erkennung von Schlüsselwörtern, Symbolen oder Phrasen, die mit schädlichen

Inhalten in Verbindung stehen, können diese Modelle sprachliche Veränderungen identifizieren, die auf neuartige Methoden extremistischer Kommunikation hinweisen. Darüber hinaus können Beiträge oder Kommentare mit bestimmten Schlüsselwörtern oder geografischen Metadaten, die auf konfliktbetroffene Regionen hinweisen, vorrangig einer Überprüfung unterzogen werden. NLP-Tools ermöglichen die effiziente Verarbeitung großer Mengen an Inhalten, sodass Systeme potenzielle Risiken schnell und präzise erfassen können (Gongane et al., 2022).

Verhaltensrisiken wie koordiniertes unauthentisches Verhalten beinhalten organisierte Bemühungen von Gruppen von Accounts, die öffentliche Meinung zu manipulieren. Dazu können Aktivitäten gehören, bei denen mehrere Konten gleichzeitig ähnliche Inhalte posten, um bestimmte Botschaften oder Narrative zu verstärken. Die Erkennung dieser Verhaltensweisen erfordert Netzwerkanalysetools, die Beziehungen und Interaktionen zwischen Konten abbilden. Graphenbasierte Modelle, wie z.B. Graph Neural Networks (GNNs), wurden entwickelt, um Netzwerkstrukturen zu analysieren und dabei zu helfen, Cluster von Accounts zu identifizieren, die ein ungewöhnlich koordiniertes Verhalten aufweisen (Kuo, 2024). Clustering-Algorithmen, die Konten auf der Grundlage gemeinsamer Merkmale gruppieren, sind ebenfalls nützlich, um Muster zu erkennen, die organische Interaktionen von koordinierten Bemühungen unterscheiden, z.B. Gruppen von Konten, die innerhalb kurzer Zeit identische Inhalte posten. Diese Werkzeuge sind wirksam bei der Aufdeckung von unauthentischen Aktivitäten, die sonst unbemerkt bleiben könnten (Smith et al. 2024).

Durch die Ausrichtung der einzelnen Instrumente auf die spezifischen Bedürfnisse der verschiedenen Risikokategorien unterstützt dieser Ansatz ein flexibles und effektives Frühwarnsystem. Die beschriebenen Instrumente und Methoden bilden die Grundlage für eine reaktionsschnelle, anpassungsfähige Überwachung von Nutzungs-, Inhalts- und Verhaltensrisiken und ermöglichen eine proaktive Erkennung und Intervention.

In der folgenden Tabelle sind die von uns ausgewählten Risiko-Fallbeispiele mit geeigneten Analyseinstrumenten aufgeführt:

Risikotyp	Beispiel für ein spezifisches Risiko	Werkzeug/Modell	Zweck
Nutzungsbedingte Risiken	Algorithmische Sucht	Modelle des maschinellen Lernens (z.B. SVMs, Random Forests)	Klassifizierung von Nutzern mit hohem Risiko auf der Grundlage von Engagement-Mustern wie Sitzungsdauer und Interaktionshäufigkeit
		Modelle zur Erkennung von Anomalien	Identifiziert plötzliche Spitzen in der Benutzeraktivität, um potenzielles Suchtverhalten zu erkennen
Inhaltsbezogene Risiken	Terroristischer Inhalt	NLP-Modelle (z.B. Transformator-basiert wie BERT)	Erkennt extremistische Sprache, Symbole und Geotags in Verbindung mit gefährdeten Gebieten
Verhaltensbedingte Risiken	Koordiniertes unauthentisches Verhalten	Graphische neuronale Netze (GNNs)	Analysiert Netzwerkstrukturen, um dichte Cluster von koordinierten Konten zu identifizieren
		Clustering-Algorithmen	Gruppiert Konten mit ähnlichen Mustern (z.B. synchronisierte Buchungen), um organisierte Manipulationen zu erkennen

Tabelle 10: Risikotypen mit Beispielen, Werkzeugen/Modellen und Zweck

Diese Tools bieten eine flexible Grundlage für ein automatisiertes Monitoring-System, bei dem jeder analytische Ansatz an die unterschiedlichen Anforderungen der verschiedenen Risikotypen angepasst werden kann. Durch die Nutzung der Stärken von maschinellem Lernen, NLP und Netzwerkanalyse kann dieses System ein umfassendes und präzises Risiko-Monitoring unterstützen und eine reaktionsschnelle Erkennung über verschiedene Kategorien hinweg gewährleisten.

6.5 Umwandlung in ein Frühwarnsystem

Der Aufbau eines wirksamen Frühwarnsystems für spezifische Risiken beginnt mit der sorgfältigen Auswahl und Identifizierung geeigneter Indikatoren. Diese Indikatoren bilden die Grundlage für die Bestimmung relevanter Datenquellen, die zur proaktiven Risikodetektion genutzt werden. Abhängig vom Fokusrisiko sind dies Datenquellen wie Plattform-APIs, Benutzerprotokolle, Metadaten oder andere Datenquellen, die beispielsweise in Tabelle 8 aufgeführt sind. Erst durch die gezielte Verknüpfung dieser Elemente gelangen die relevanten Daten zu den Analysetools und werden in einem koordinierten Rahmen für eine umfassende Analyse integriert.

Nach der Datenerfassung interpretieren Analysetools wie maschinelles Lernen, NLP und Netzwerk-Analysen die Informationen, um Muster oder Anomalien zu identifizieren, die auf entstehende Risiken hinweisen. Ein längerer Anstieg der Sitzungsdauer kann so als Hinweis auf potenziell süchtig machende Nutzungsgewohnheiten erkannt werden, oder synchronisierte Posting-Muster als Zeichen für koordiniertes unauthentisches Verhalten.

Bei der Identifikation eines Risikoereignisses priorisiert das System die Risiken nach Dringlichkeit, wobei Faktoren wie der Umfang (z.B. Anzahl der betroffenen Nutzer*innen oder Beiträge) und die Reichweite (z.B. Einfluss oder Verbreitung des Ereignisses) in die Bewertung einfließen. Diese Priorisierung minimiert Fehlalarme und stellt sicher, dass priorisierte Warnungen sofortige Überprüfungen oder Interventionen auslösen, was eine reaktionsfähige Feedbackschleife schafft.

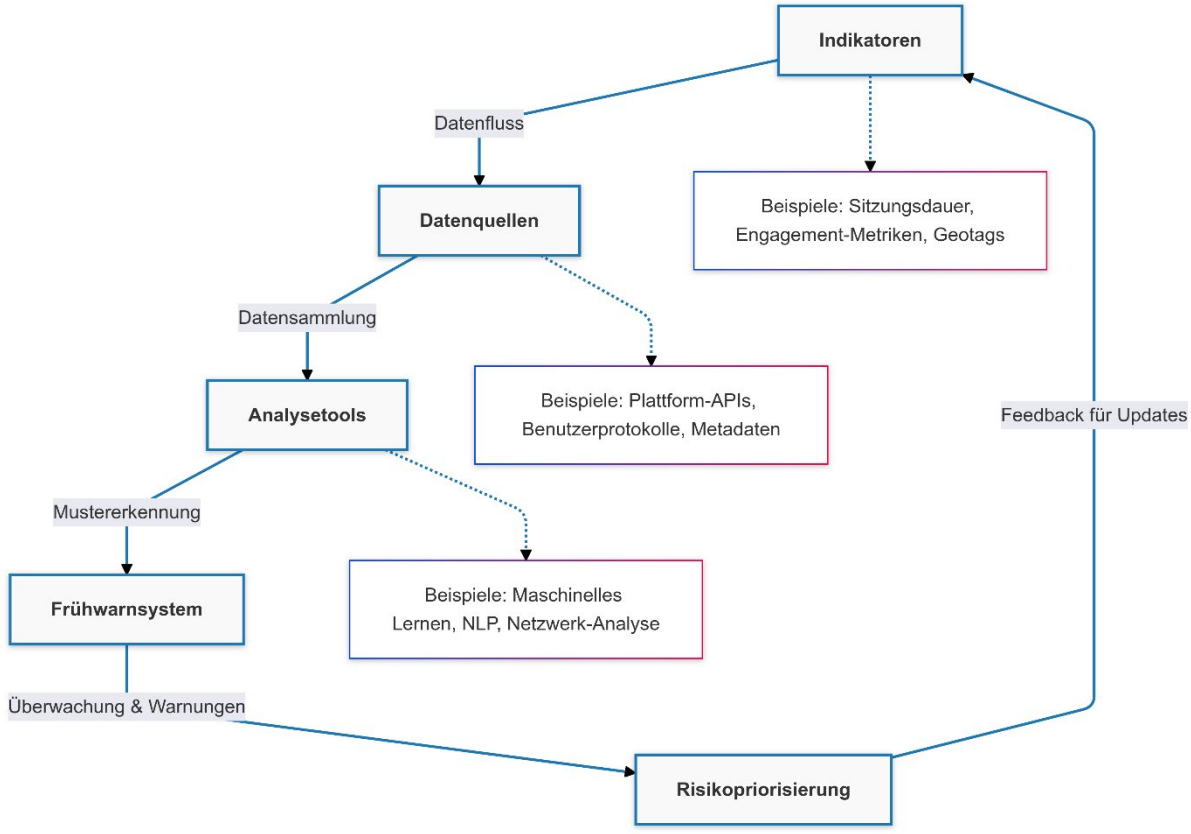

Abbildung 5: Stilisierte Version des Prozesses des Frühwarnsystems

Da sich das Nutzerverhalten und die Inhalte kontinuierlich weiterentwickeln, werden die Analysemodelle regelmäßig aktualisiert, um die Erkennungsgenauigkeit zu bewahren. Diese Anpassungsfähigkeit stellt sicher, dass das System im Laufe der Zeit neue und subtile Risikosignale erfasst und seine Widerstandsfähigkeit angesichts einer sich verändernden Risikolandschaft aufrechterhält. Dieser Ansatz schafft ein dynamisches Frühwarnsystem, das Indikatoren, Daten und Analysetools zu einer kohärenten, proaktiven Überwachungslösung verbindet.

6.6 Herausforderungen, Beschränkungen und Risiken

Die Entwicklung eines automatisierten Frühwarnsystems zur Erkennung spezifischer Risiken, das auch das Potenzial zur Bewertung systemischer Risiken durch zusätzliche Wirkungsmessungen bietet, bringt sowohl technische als auch ethische Herausforderungen mit sich. Eine zentrale technische Hürde besteht in der Auswahl zuverlässiger Indikatoren, die auf entstehende Risiken hinweisen, ohne eine übermäßige Anzahl von Fehlalarmen auszulösen. Viele Indikatoren erfordern eine sorgfältige Kalibrierung, um sicherzustellen, dass harmlose Aktivitäten nicht übererfasst und tatsächliche Risiken nicht unterschätzt werden. Der eingeschränkte Zugang zu Plattformdaten erschwert dies zusätzlich, da Forscher*innen oft begrenzte Echtzeit-Einblicke in Nutzerinteraktionen und Inhalte haben. Darüber hinaus müssen maschinelle Lernmodelle zur Risikoerkennung regelmäßig aktualisiert werden, um Verhaltensänderungen – etwa in der extremistischen Sprache oder den Taktiken koordinierter Aktionen – abzubilden. Diese Updates sind ressourcenintensiv und können die Reaktionsfähigkeit des Systems beeinflussen.

Auch ethische Überlegungen spielen eine wesentliche Rolle. Das Monitoring von Inhalten und Nutzerverhalten wirft Fragen des Datenschutzes auf, insbesondere wenn die Datenerfassung sensible oder persönliche Informationen betrifft. Zudem besteht das Risiko algorithmischer Verzerrungen, bei denen Erkennungsmodelle unbeabsichtigt bestimmte Muster bevorzugen oder Inhalte falsch klassifizieren, was zu einer potenziell unfairen Ausrichtung führen kann. Um diese Verzerrungen zu mindern, sind strenge Aufsicht, Transparenz und Schutzmaßnahmen zur Wahrung der Nutzerrechte erforderlich. Darüber hinaus sind die betrieblichen Anforderungen an die Echtzeitüberwachung und Verarbeitungsleistung oft hoch, was die Skalierbarkeit herausfordernd macht. Da sich digitale Umgebungen ständig weiterentwickeln, müssen die Indikatoren, Datenquellen und Modelle des Systems anpassungsfähig bleiben; andernfalls droht das System schnell veraltet zu sein. Die Entwicklung eines resilienten, ethisch verantwortungsvollen Frühwarnsystems für spezifische Risiken, das bei Bedarf auf die Erkennung systemischer Risiken skaliert werden kann, erfordert die Bewältigung dieser Herausforderungen, um eine effektive und verantwortungsvolle Überwachung sicherzustellen.

7. Fallstudie: Telegram

7.1 Ausgangslage: Telegram als Nicht-VLOP

Die zusätzlichen Verpflichtungen in Bezug auf den Umgang mit systemischen Risiken gelten nach Art. 33 Abs. 1 DSA nur für Online-Plattformen und -Suchmaschinen, die durchschnittlich mindestens 45 Millionen monatlich aktiven Nutzer*innen in der EU (ca. 10 % der Bevölkerung in der EU) haben und als VLOPs oder VLOSEs benannt sind. Hierzu erläutert Erwägungsgrund 76 DSA, dass VLOPs und VLOSEs »gesellschaftliche Risiken bewirken [können], die sich hinsichtlich Umfang und Auswirkungen von denen kleinerer Online-Plattformen unterscheiden.« Daher sollten sie »höchsten Sorgfaltspflichten unterliegen, die in einem angemessenen Verhältnis zu ihren gesellschaftlichen Auswirkungen stehen.« Die Anbieter*innen von Online-Plattformen und -Suchmaschinen müssen nach Art. 24 Abs. 2 DSA mindestens alle sechs Monate Informationen über die durchschnittliche monatliche Zahl ihrer aktiven Nutzer*innen veröffentlichen. Zur Festlegung der Methode zur Bestimmung der aktiven Nutzer*innen hat die Europäische Kommission ein Q&A-Dokument veröffentlicht (Europäische Kommission, 2023).

Eine risikobasierte Abstufung auf Grundlage der Nutzerzahlen erlaubt zwar einerseits – zumindest theoretisch – eine klare Eingrenzung des Anwendungsbereichs der zusätzlichen Verpflichtungen zum Umgang mit systemischen Risiken. Andererseits nimmt der nutzerzahlbasierte Ansatz aber auch solche Dienste von den zusätzlichen Verpflichtungen aus, die durchschnittlich weniger als 45 Millionen monatlich aktiven Nutzer*innen in der EU haben, gleichzeitig aber aufgrund anderer Risikofaktoren gesellschaftliche Risiken bewirken können, die sogar größer als die einiger VLOPs und VLOSEs sein könnten. In anderen Worten: Die Verpflichtungen solcher Nicht-VLOPs stehen nicht in einem angemessenen Verhältnis zu ihren gesellschaftlichen Auswirkungen.

Ein solches Beispiel könnte Telegram sein, das nach eigenen Angaben schätzungsweise durchschnittlich 41 Millionen monatlich aktive Nutzer*innen in der EU zwischen September 2023 und Februar 2024 aufwies und damit knapp unter dem DSA-Schwellenwert lag (Telegram, 2024). Während Telegram damit zum Zeitpunkt der Studienveröffentlichung nicht dazu verpflichtet ist, Risikobewertungen durchzuführen, könnte eine erste externe vereinfachte Risikoermittlung jedoch Aufschluss darüber geben, welche Kernrisiken von Telegram und seiner Nutzung ausgehen und worauf sich eine mögliche Risikobewertung und -minderung fokussieren sollte.

7.2 Vereinfachte Risikoermittlung bei Telegram

Bei der Risikoermittlung geht es im Kern darum, ein solides Verständnis der Kernrisiken zu erhalten. Dafür wird im Folgenden ein spezifisches vorläufiges Risikoprofil erarbeitet, welches das Verständnis potenzieller Schäden auf der Grundlage zentraler Merkmale von Telegram zusammenfasst, sogenannter Risikofaktoren. Aufgrund des Umfangs dieser Studie kann dabei nur auf einige ausgewählte Merkmale des Dienstes eingegangen werden. Zudem basiert die durchgeführte Risikoermittlung auf einer Literaturrecherche. Für die Erarbeitung eines vollständigen Risikoprofils von Telegram wäre nicht nur die Berücksichtigung weiterer Risikofaktoren, sondern auch die zusätzliche Konsultation interner Telegram-Daten, Nutzerfeedback und die Einbeziehung externer Stakeholder*innen erforderlich. Dabei sollte berücksichtigt werden, dass eine externe Risikobewertung nur eingeschränkt möglich ist, solange Telegram nicht als VLOP eingestuft ist und damit Forschenden keinen Zugang zu öffentlichen oder nicht-öffentlichen Daten unter bestimmten Voraussetzungen bereitstellen muss. Ohne einen verpflichteten Zugang unterliegt die externe Erforschung von Telegram fragmentierungsbedingen Hindernissen (u.a. Vielzahl an privaten Räumen) und forschungsethischen Bedenken (u.a. Verschleierung der eigenen Identität, um ethnografische Forschung durchzuführen) (Tuck et al., 2023).

Den ersten Schritt der Risikoermittlung stellt die Identifizierung von Risikofaktoren dar (s. a)). Grundlage hierfür ist der in dieser Studie entwickelte Fragenkatalog. Der zweite Schritt besteht aus der konkreten Erstellung eines Risikoprofils (s. b)). Im Fokus steht hierbei die Zuordnung der im ersten Schritt identifizierten Risikofaktoren zu konkreten systemischen Risiken von Telegram. Diese Zuordnung erfolgt dabei auf Grundlage vorliegender Forschungsbefunde zu relevanten Risiken.

a) Identifizierung von Risikofaktoren bei Telegram

Bei der Identifizierung von Risikofaktoren sind die Risikofaktoren »Durchsetzung der Allgemeinen Geschäftsbedingungen«, »Systeme zur Auswahl und Anzeige von Werbung« und »Datenbezogene Praktiken« im Folgen-

den bewusst außen vor geblieben, da der Fokus zunächst auf grundlegenden Risikofaktoren lag, wie der »Art der Dienste«. Die weiteren Faktoren sollten jedoch in einer umfassenderen Risikoermittlung einbezogen werden.

Art der Dienste: Telegram ist ein hybrider Instant-Messaging-Dienst, der Funktionen einer Social-Media-Plattform aufweist. Zwar fallen interpersonelle Kommunikationsdienste laut Erwägungsgrund 14 DSA nicht in den Anwendungsbereich der Begriffsbestimmung für Online-Plattformen, da sie »für die interpersonelle Kommunikation zwischen einer endlichen Zahl von Personen verwendet werden, die vom Absender der Kommunikation bestimmt wird«. Trotzdem sollen die einschlägigen DSA-Vorgaben laut Erwägungsgrund 15 DSA auch für diejenigen Dienste gelten, die in deren Anwendungsbereich fallen. Funktionen einer Online-Plattform stellen bei Telegram dabei vor allem Gruppen und Kanäle ohne Zugangsbeschränkungen dar, sogenannte öffentliche Gruppen und Kanäle (s. Tabelle 11). Davon abzugrenzen sind Einzelchats oder kleinere geschlossene Gruppen und Kanäle. Geschlossene Gruppen oder Kanäle mit sehr vielen Nutzer*innen fallen wiederum in eine Grauzone (s. 7.3).

Eher öffentlich	Öffentliche Gruppen und Kanäle
	Große private Gruppen, deren Links in öffentlichen Kanälen und auf anderen Social-Media-Plattformen veröffentlicht werden.
	Kleinere private Gruppen, deren Links nur in anderen privaten Gruppen für eine begrenzte Zeit veröffentlicht werden; für den Beitritt ist gegebenenfalls eine Genehmigung des Admins erforderlich.
	Private Gruppen, deren Links nicht auf einer Social-Media-Plattform gepostet werden, sondern in privaten Nachrichten zwischen einzelnen Nutzer*innen ausgetauscht werden; Administrator*innen können festlegen, dass nur Mitglieder aus ihren vertrauenswürdigen Kontakten ausgewählt werden.
Eher privat	Private Nachrichten zwischen zwei Nutzer*innen

Tabelle 11: Maß an Öffentlichkeit für Gruppen- und Kanalarten auf Telegram

Größe und Nutzerbasis: Telegram wies nach eigenen Angaben schätzungsweise durchschnittlich 41 Millionen monatlich aktive Nutzer*innen in der EU zwischen September 2023 und Februar 2024 auf. Weltweit sollen es jedoch 950 Millionen monatlich aktive Nutzer*innen sein. Während die Analysefirma Similarweb von knapp 51 Millionen monatlich aktive Nutzer*innen allein in sieben EU-Mitgliedsstaaten zwischen Februar und Juli 2024 ausgeht, verweist Telegram selbst darauf, dass die »Zahl der monatlich aktiven Nutzer, die für die Berechnung dieses Schwellenwerts relevant ist, sogar noch niedriger sein [dürfte als die bereits angegebenen 41 Millionen Nutzer*innen], da nur einige der Funktionen von Telegram als Online-Plattformen im Sinne des Gesetzes über digitale Dienste gelten können.« Ob Telegram bislang Nutzer*innen von mitgliedsstarken Kanälen und Gruppen mit Zugangsbeschränkungen einberechnet hat oder nicht, ist aus externer Perspektive kaum überprüfbar. Der belgische DSC, der für Telegram als Nicht-VLOP aufgrund der gesetzlichen Vertretung in Belgien zuständig ist, geht nicht davon aus, dass der VLOP-Schwellenwert überschritten ist. Die Europäische Kommission überprüft wiederum die angewandte Einstufungsmethode des Unternehmens. Im Hinblick auf die Nutzerbasis in Deutschland ist der Anteil derjenigen Nutzer*innen von Instant-Messaging-Diensten, die Telegram nutzen, über alle Altersgruppen hinweg vom 04.10.2023 bis 26.09.2024 vergleichsweise konstant. Am höchsten lag der Nutzeranteil in den Altersgruppen der 25- bis 34-Jährigen sowie der 35- bis 44-Jährigen mit 20 bzw. 21 Prozent. Im Hinblick auf Geschlechterverhältnis waren im Jahr 2022 knapp zwei Drittel der Nutzer*innen männlich und nur etwas mehr als ein Drittel weiblich. Laut Allgemeinen Geschäftsbedingungen (AGB) ist die Nutzung des Dienstes erst ab einem Alter von 18 Jahren im Europäischen Wirtschaftsraum erlaubt.

Geschäftsmodell: Insgesamt basiert das Geschäftsmodell auf einem Freemium-Ansatz, kombiniert mit dezenten Werbe- und innovativen Blockchain-Strategien (Kollmann, 2024). Telegram wurde ursprünglich kostenlos und werbefrei durch Eigenfinanzierung von Gründer Pawel Durow betrieben. Im Laufe der Zeit hat Telegram jedoch mehrere Einnahmequellen eingeführt, um sich zu finanzieren. Eine zentrale Säule des Geschäftsmodells ist Telegram Premium, ein kostenpflichtiges Abonnement, das exklusive Funktionen wie größere Upload-Limits, schnellere Downloads und werbefreie Nutzung bietet. Zusätzlich generiert Telegram Einnahmen durch dezente Werbung in öffentlichen Kanälen, die themenspezifisch ist und ohne Nutzertracking abläuft. Blockchain-Technologie spielt ebenfalls eine Rolle: Über das TON-Netzwerk verkauft Telegram beispielsweise Benutzernamen. Ergänzend hat Telegram über Anleihen Kapital aufgenommen, um weiteres Wachstum zu finanzieren. Im Jahr 2023 lagen die Einnahmen von Telegram bei 342 Millionen US-Dollar (36 Cent/Nutzer*in bei 950 Millionen Nutzer*innen) (R. Smith & Murphy, 2024). Gleichzeitig betragen die Kosten per Nutzer*in nach Angaben des Telegram-Gründers Durov circa 70 Cent (Murphy, 2024). Damit existiert ein hoher Kostendruck im Unternehmen.

Gestaltung von Empfehlungssystemen und anderen relevanten algorithmischen Systemen: Telegram verwendet keine algorithmischen Empfehlungssysteme. Trotzdem können Inhalte in anderen Kanälen geteilt und in Gruppen diskutiert werden. Wird ein Inhalt geteilt, so nimmt damit auch die Zahl der Aufrufe des ursprünglichen Inhalts zu. Wird ein Kanal gelöscht, können daher immer noch viele Gruppen existieren, auf denen die Inhalte eines Nachfolgekanals geteilt werden – eine Möglichkeit zur Schaffung dezentraler Netzwerke (CeMAS, 2024). Kanäle und Gruppen ohne Zugangsbeschränkungen können außerdem über die Funktion »Globale Suche« von Telegram gefunden werden (Telegram, 2024), wobei Kanäle und Gruppen nach der Anzahl der Abonnent*innen und verifizierte Konten an erster Stelle gelistet und Suchergebnisse aus dem Land der Nutzer*in bevorzugt angezeigt werden. Die Inhalte der Kanäle sind auch sichtbar, wenn man den Kanälen nicht beigetreten ist (Tuck et al., 2023). Für den Beitritt zu Gruppen oder Kanälen mit Zugangsbeschränkung ist dagegen eine Einladung über Link erforderlich, der von Administrator*innen bereitgestellt wird. Dabei existiert die Möglichkeit zur Erstellung von Einladungslinks mit unterschiedlichen Bedingungen (z.B. zusätzliche Freigabe der Beitrittsanfrage, Zeitlimit), weswegen auch die Zahl der Dritten, denen Informationen bereitgestellt werden, stark variieren kann.

Systeme zur Moderation von Inhalten: Telegram kann bestimmte Funktionen eines Kontos vorübergehend oder dauerhaft einschränken (Telegram, 2024). Dazu gehören beispielsweise das Kommunizieren mit Personen, die vermutlich niemanden kennen, oder das Er-

stellen sowie Teilnehmen an Kanälen und Gruppen ohne Zugangsbeschränkungen. Konten, Bots, Gruppen und Kanäle, die sich falsch ausgeben oder versuchen, andere Nutzer*innen zu täuschen, können mit einer Kennzeichnung »FÄLSCHUNG« oder »BETRUG« im öffentlichen Profil gekennzeichnet werden. Bei schwerwiegenden Verstößen haben die Moderator*innen von Telegram die Möglichkeit, Nutzer*innen, Bots, Beiträge, Kanäle oder Gruppen zu sperren oder vollständig zu entfernen.

Allgemeine Geschäftsbedingungen: Die Nutzungsbedingungen von Telegram, die nur für Kanäle und Gruppen ohne Zugangsbeschränkungen gelten (»öffentliche Plattform«), untersagen folgende Inhalte und Aktivitäten: Spam (unaufgeforderte oder unerwünschte Werbung sowie jede Form von kommerzieller Belästigung); Förderung von Gewalt (Aufrufe zu Gewalt, Terrorismus, das Sammeln von Geldern für terroristische Organisationen usw.); rechtswidrige sexuelle Inhalte (Material über sexuellen Kindesmissbrauch, Bestiality und die nicht-einvernehmliche Veröffentlichung von sexuellem Material); Aktivitäten, die in den meisten Ländern als rechtswidrig anerkannt sind (Kindesmissbrauch, der Verkauf oder das Anbieten von rechtswidrigen Waren und Dienstleistungen wie Drogen, Schusswaffen und gefälschte Dokumente), die Weitergabe von persönlichen Daten anderer, um sie einzuschüchtern oder zu schikanieren (Doxing) (Telegram, 2024).

b) Erstellung eines vorläufigen Risikoprofils von Telegram

Im Rahmen einer vereinfachten Risikoermittlung fokussiert sich die Profilerstellung im Folgenden einzig auf die Zuordnung des Risikofaktors »Art der Dienste« zu systemischen Risiken. Die weiteren Risikofaktoren sollten allerdings bei der Erstellung eines vollständigen Risikoprofils von Telegram mit einbezogen werden, um den spezifischen Charakteristika von Telegram gerecht zu werden. Ungeachtet dessen liefert der Risikofaktor »Art der Dienste« bereits erste Indizien für Kernrisiken von Telegram.

Bei der Analyse des Risikofaktors »Art der Dienste« ist im ersten Schritt festgestellt worden, dass Telegram Funktionen einer Social-Media-Plattform umfasst. Für Social-Media-Plattformen sieht das im Rahmen dieser Studie entwickelte allgemeine Risikoprofi wiederum gemeinhin folgende potenzielle Online-Schäden vor: 1) Nutzungsbedingte Risiken: Suchtverhalten und Aussetzung gegenüber schädlichen Interaktionen, die die psychische Gesundheit beeinträchtigen; 2) inhaltsbezogene Risiken: Verstärkung rechtswidriger Inhalte wie Hassrede, Fehlinformationen und CSAM – Material über sexuellen Kindesmissbrauch – sowie 3) verhaltensbedingte Risiken: Ausnutzung der Plattform durch Akteur*innen zur Belästigung, für Desinformationskampagnen oder die gezielte Angriffe auf gefährdete Nutzer*innen. Wird nach diesen Risiken auf Grundlage frei verfügbarer Quellen recherchiert, so finden sich relevante Forschungsbefunde insbesondere zu den Kategorien 2 und 3.

Inhaltsbezogene Risiken: Dem Center für Monitoring, Analyse und Strategie (CeMAS) zufolge gilt Telegram mittlerweile als die »wichtigste Plattform für Verschwörungsideolog*innen und Rechtsextreme im deutschsprachigen Raum« (CeMAS, 2024). Laut CeMAS wurden unter anderem Putsch- und Terrorpläne öffentlich auf Telegram diskutiert und antisemitische Inhalte massenhaft verbreitet. Während Extremist*innen auf den etablierten größeren Online-Plattformen ihre »weicheren« Inhalte platzieren, um nicht gelöscht zu werden, werden ihre Follower*innen für radikalere bzw. extremere Inhalte auf Telegram verwiesen. Darüber hinaus wird Telegram vom Rechtsaußen-Online-Milieu in Deutschland zusätzlich als Instrument zur externen Vermittlung von Inhalten verwendet. In anderen Worten: Beiträge werden auf Telegram in einem gemäßigteren Tonfall formuliert, um außerhalb des Rechtsaußen-Milieus Zuspruch zu finden (Tuck et al., 2023). Während der COVID-19-Pandemie ist Telegram zudem dazu genutzt worden, Verschwörungserzählungen über die Impfung zu verbreiten. So nahm die Leserschaft verschiedener Telegram-Kanäle der impfskeptischen Szene in Deutschland im Zeitraum vom 21. Dezember 2020 bis zum 5. April 2021 um ca. 471 Prozent zu. Viele dieser Kanäle verbreiteten gefährliche Gesundheitsfehlinformationen. Im Fall des russischen Angriffs auf die Ukraine förderten rechtsradikale und rechtsextreme Kanäle auf Telegram zudem die Verbreitung russischer Kriegspropaganda, indem sie Inhalte von sanktionierten russischen Staatsmedien posteten. Neben diesen inhaltsbezogenen Risiken steht Telegram auch wegen der Verbreitung von CSAM in der Kritik. Zuletzt sind in Frankreich Ermittlungen gegen Gründer Durov eingeleitet worden wegen des Verdachts, er habe sich durch fehlendes Eingreifen und unzureichende Kooperation des Drogenhandels, Betrugs und Vergehen im Zusammenhang mit Kindesmissbrauch mitschuldig

gemacht.

Verhaltensbedingte Risiken: Rechtsextreme und Verschwörungsideolog*innen haben in den letzten Jahren ihre Infrastruktur auf Telegram erheblich ausgebaut (Gerster et al., 2021). Dabei existieren auch einzelne sogenannte Poweruser*innen aus diesen Milieus, die eine zentrale Rolle beim Verlinken externer Inhalte spielen. Manche bauen sich komplexe Netzwerke zwischen verschiedenen Kanälen und Gruppen auf, um ihre Reichweite zu vergrößern und gegen eine mögliche Löschung resilient zu sein. Denn kommt es zur Löschung eines Kanals, so können Gruppen weiter existieren, auf denen dann wiederum die Inhalte eines Nachfolgekanals geteilt werden. Insbesondere Akteur*innen der genannten Milieus haben diese Strategie der dezentralen Netzwerkbildung angewandt. Während der COVID-19-Pandemie hat die russische Staatspropaganda Konten von Rechtsextremen und Verschwörungsideolog*innen auf Telegram als nützliche Multiplikator*innen entdeckt und sie gezielt mit vertrauensbildenden Inhalten, welche die Narrative der Community widerspiegeln, versorgt, um diese Konten zu einem späteren Zeitpunkt wiederum bei der Verbreitung von Kriegspropaganda im Ukraine-Krieg einzubinden (Smirnova & Winter, 2021).

Seit Beginn des russischen Angriffskrieges gegen die Ukraine hat sich die Anzahl der Abonnierenden von pro-Kreml-Kanälen auf Telegram verdreifacht (Generaldirektion CNECT, 2023). Dokumente der vom russischen Staat beauftragten Social Design Agency (SDA) belegen weiter, dass auf Telegram gezielt eigene Kanäle erstellt und Kommentare unter Beiträge authentischer Kanäle als Teil einer Desinformationskampagne verfasst wurden (Smirnova, 2024). Im Rahmen dieser Kampagne definierte die SDA sogar konkrete Zielvorgaben und Kennzahlen für Deutschland. So sollte die Anzahl der Menschen, die in Umfragen angeben, Angst vor der Zukunft zu haben, auf mindestens 50 Prozent steigen. 55 Prozent der Deutschen sollten außerdem in Folge der Kampagne der Behauptung zustimmen, sie würden ihren Wohlstand nicht für den Sieg über Russland aufopfern wollen. Diese Befunde zeigen, dass Telegram vorsätzlich von Dritten für unterschiedliche Ziele ausgenutzt wurde, wie etwa zur gesellschaftlichen Destabilisierung in Deutschland.

7.3 Anpassungen des Rechtsrahmens

Zwar konnten im Rahmen dieser Studie nur ein vorläufiges Risikoprofil erarbeitet und keine nachgelagerte Risikobewertung und Risikominderung im Sinne des vorgeschlagenen Bewertungsrahmen durchgeführt werden. Dennoch hat die vereinfachte Risikoermittlung aufgezeigt, dass Telegrams Funktionen einer Social-Media-Plattform inhalts- und verhaltensbezogene Kernrisiken bewirken können. Dazu zählen beispielsweise die Verbreitung von extremistischen Inhalten, von CSAM oder von pro-russischen Desinformationskampagnen. Dennoch muss Telegram – genauso wie andere Nicht-VLOPs, von denen gesellschaftliche Risiken ausgehen – keine zusätzlichen Verpflichtungen in Bezug auf den Umgang mit systemischen Risiken erfüllen. Vor diesem Hintergrund sollten Möglichkeiten zur Anpassung des Anwendungsbereichs solcher Verpflichtungen, wie beispielsweise die Aufstellung restriktiver Kriterien für geschlossene Informationsräume (s. a)) oder das Hinzuziehen einer einmaligen Risikobewertung als VLOP-Einstufungskriterium (s. b)), diskutiert werden.

a) Restriktive Kriterien für geschlossene Informationsräume

Wird der Ansatz einer VLOP-Einstufung auf Grundlage der Nutzerzahl grundsätzlich beibehalten, so könnte geprüft werden, unter welchen Voraussetzungen sehr mitgliedsstarke geschlossene Informationsräume, wie beispielsweise Telegram-Gruppen mit Zugangsbeschränkungen, als Online-Plattform eingestuft werden können. Telegram selbst definiert diese bislang als nicht-öffentlich und nimmt sie von den Nutzungsbedingungen aus (Tuck et al., 2023). Nach Art. 3 i) DSA meint eine Online-Plattform »einen Hostingdienst, der im Auftrag eines Nutzers Informationen speichert und öffentlich verbreitet, sofern es sich bei dieser Tätigkeit nicht nur um eine unbedeutende und reine Nebenfunktion eines anderen Dienstes oder um eine unbedeutende Funktion des Hauptdienstes handelt […].« Weiter wird definiert, dass unter der öffentlichen Verbreitung die »Bereitstellung von Informationen für eine potenziell unbegrenzte Zahl von Dritten im Auftrag des Nutzers, der die Informationen bereitgestellt hat« (Art. 3 k) DSA) zu verstehen ist.

Im Hinblick auf die Klassifikation von sehr mitgliedsstarken Telegram-Gruppen mit Zugangsbeschränkungen ergeben sich daraus zwei Problemstellungen: Zum einen stellt sich die Frage, ob diese Gruppen eine reine Nebenfunktion eines anderen Dienstes oder eine unbedeutende Funktion des Hauptdienstes darstellen. Angesichts der Relevanz von Gruppen mit Zugangsbeschränkungen im Geschäftsmodell von Telegram ist jedoch davon auszugehen, dass es sich nicht um eine Nebenfunktion oder unbedeutende Funktion des Hauptdienstes handelt. Zum anderen muss beurteilt werden, ob sich die Gruppen an eine potenziell unbegrenzte Zahl von Dritten richten. Ist der Einladungslink zu einer Gruppe öffentlich zugänglich und keine weitere menschliche Freigabe erforderlich, so könnte es sich bereits um öffentliche Kommunikation handeln. So konkretisiert Erwägungsgrund 14 DSA, dass »in Fällen, in denen eine Registrierung oder die Aufnahme in eine Nutzergruppe erforderlich ist, um Zugang zu Informationen zu erlangen, nur dann von einer öffentlichen Verbreitung der Informationen ausgegangen werden [soll], wenn die Nutzer, die auf die Informationen zugreifen möchten, automatisch registriert oder aufgenommen werden, ohne eine menschliche Entscheidung oder Auswahl, wem Zugang gewährt wird.« Zusätzlich sollte ein hoher Schwellenwert (z.B. 10.000 Mitglieder) festgelegt werden (Panahi et al., 2024).

Sollten entsprechende Kriterien entwickelt und verabschiedet werden (z.B. über einen delegierten Rechtsakt nach Art. 33 Abs. 3 DSA), so müssen in jedem Fall die geltenden Rechte von Nutzer*innen und Anbieter*innen gewahrt bleiben. Alternativ wäre es auch denkbar, alle registrierten Nutzer*innen eines hybriden Instant-Messaging-Dienstes oder eines anderen hybriden Dienstes, der Funktionen einer Online-Plattform umfasst, in die Berechnung mit einzubeziehen (Panahi et al., 2024). Denn auch sie haben in der Regel Zugriff auf die Telegram-Funktionen einer Social-Media-Plattform, indem sie beispielsweise Kanälen ohne Zugangsbeschränkungen folgen können.

b) Einmalige Risikobewertung als Einstufungskriterium

Ein anderer Ansatz könnte darin bestehen, die VLOP-Einstufung auf Grundlage der Nutzerzahl zu verwerfen oder um weitere alternative Einstufungskriterien zu erweitern. Seit der Einstufung der ersten VLOPs und VLOSEs hat es immer wieder Kritik an den Einstufungskriterien gegeben. Beispielsweise reichte Zalando, das als einziger in Deutschland ansässiger Dienst als VLOP eingestuft wurde, Klage gegen die Einstufung durch die Europäische Kommission ein. Konkret argumentiert das Unternehmen, dass seine Nutzerzahl falsch interpretiert und das hauptsächlich auf den Einzelhandel ausgerichtete Geschäftsmodell nicht anerkannt worden sei (Killeen, 2023).

Ungeachtet der Einbeziehung bestimmter Funktionen des Dienstes wird im Falle von Zalando jedoch auch ein anderer Aspekt deutlich. Im Falle von Zalando hat es bislang kaum Befunde über die Verbreitung rechtswidriger Inhalte gegeben, einschließlich Informationen, die durch ihre Bezugnahme auf den Verkauf von Produkten nicht im Einklang mit dem Unionsrecht oder dem Recht eines Mitgliedsstaats stehen, bzw. nachteiliger Auswirkungen auf die Grundrechte oder die gesellschaftliche Debatte hatte. Telegram hingegen ist nicht als VLOP eingestuft worden – trotz belastbarer Erkenntnisse für systemische Risiken, die vom Dienst und seiner Nutzung ausgehen. Vor diesem Hintergrund scheint das 45-Millionen-Nutzer-Kriterium folglich keine sinnvolle Unterscheidung zwischen vertretbaren und nichtvertretbaren gesellschaftlichen Auswirkungen darzustellen.

Vor diesem Hintergrund sollte eine Einstufung auf Grundlage der Nutzerzahl kritisch hinterfragt werden. Zum einen könnten andere relevante Risikofaktoren (z.B. Nutzerbasis, Empfehlungssysteme) und damit einhergehende quantitative und qualitative Schwellenwerte als alternative Einstufungskriterien herangezogen werden. Zum anderen könnte eine Möglichkeit darin bestehen, eine einmalige Risikobewertung eines Nicht-VLOPs wie Telegram mit einer durchschnittlichen monatlichen Zahl von weniger als 45 Millionen aktiven Nutzer*innen auf Grundlage externer Erkenntnisse zu systemischen Risiken in der EU anzufordern, sofern hinreichende externe Evidenzen für potenzielle systemische Risiken vorliegen. Diese einmalige Bewertung könnte wiederum herangezogen werden, um zu einem abschließenden Ergebnis einer Einstufung als VLOP zu gelangen. Eine solch grundlegende Änderung am bisherigen Einstufungsansatz sollte jedoch nur im Rahmen einer umfassenden Novellierung der Verordnung vorgenommen werden. Eine Evaluation des DSA soll bis Anfang 2027 erfolgen.

8. Schlussfolgerungen

8.1 Veröffentliche Risikobewertungen

Ein entscheidendes Element im Umgang mit den mit digitalen Plattformen verbundenen systemischen Risiken ist die Veröffentlichung von Risikobewertungen der VLOPs und VLOSEs im Rahmen des DSA. Zwar schreibt der DSA nicht ausdrücklich die vollständige Veröffentlichung von Risikobewertungen vor, doch wird in Erwägungsgrund 100 die Bedeutung einer umfassenden Berichterstattung betont. Eine solche Transparenz ist unerlässlich, um externen Forschenden die Möglichkeit zu geben, ihre Identifizierung und ihr Verständnis von systemischen Risiken zu verfeinern. Erwägungsgrund 90 DSA plädiert ferner für die Einbeziehung von Dienstleistungsempfänger*innen, betroffenen Gruppen, unabhängigen Expert*innen und Organisationen der Zivilgesellschaft in der Ausarbeitung ihrer Methoden zur Bewertung von Risiken und der Gestaltung von Risikominderungsmaßnahmen. Unter in Art. 42 Absatz 5 des DSA definierten Umständen, sind Anbieter*innen dazu berechtigt, vertrauliche Informationen der Plattformen oder von Nutzer*innen aus den öffentlich zugänglichen Berichten zu entfernen. Die erste Runde von Risikobewertungen wurde bis Ende November 2024 von den meisten VLOPs und VLOSEs veröffentlicht. Im Folgenden werden grundsätzliche Herausforderungen der Risikobewertungen skizziert, welche sich aus der ersten Runde von der veröffentlichten Risikoberichten ergeben. Dies kann als Grundlage für eine tiefergehende Analyse dienen, um die Methodiken und Ansätze zu bewerten und darauf aufbauend gute Praktiken für potenzielle Leitlinien künftiger Risikobewertungen zu entwickeln.

Methodische Ansätze zur Risikobewertung und -ermittlung

In den Berichten wurden von vielen Anbieter*innen, darunter AliExpress, Bing, Google und Meta, Risikobewertungsmethoden eingeführt, die sich an den Grundprinzipien strukturierter Rahmenwerke, wie dem Ansatz des ISD, orientieren. Diese Methoden beinhalten im Allgemeinen die Definition eines Risikoprofils, die Identifizierung von Risikofaktoren, die über die vier systemischen Risiken gemäß Art. 34 Absatz 1 des DSA hinausgehen, und die Bewertung von Risiken auf der Grundlage der Wahrscheinlichkeit und Schwere ihrer Auswirkungen. Dabei wurde unter anderem auf etablierte Rahmenwerke wie die Sorgfaltspflicht der Vereinten Nationen im Bereich der Menschenrechte (DGCN, 2014) oder den ISO 31000-Standards (ISO, 2018) Bezug genommen. Die Risikoberichte, in denen strukturierte Methoden zur Anwendung kamen, erweiterten oft die Anforderungen des DSA und bezogen zusätzliche Risikofaktoren ein, was eine positive Entwicklung ist. Andere Unternehmen wie TikTok, Amazon und Zalando boten jedoch nur vage Beschreibungen ihrer Risikobewertungsprozesse an und ließen eine klare Struktur vermissen, was die Transparenz und Vergleichbarkeit erschwert und erhebliche Lücken in Bezug auf Klarheit und Rechenschaftspflicht hinterlässt.

Herausforderungen bei der Entwicklung von Metriken

Eine zentrale Herausforderung für alle Plattformen ist das Fehlen spezifischer Indikatoren für die Bewertung der Wahrscheinlichkeit und der Auswirkungen systemischer Risiken. Ohne klare Metriken ist es schwierig, Risiken plattformübergreifend oder sogar innerhalb derselben Plattform im Zeitverlauf zu messen oder zu vergleichen. Die Verschiedenartigkeit der Risiken und ihre kontextabhängige Dynamik erschweren die Entwicklung von Indikatoren, die sowohl umsetzbar als auch universell anwendbar sind. Während einige Plattformen beispielsweise den Schweregrad der Auswirkungen als gewichtigeres Kriterium betrachten (z. B. gewichtet Booking.com den Schweregrad dreimal so hoch wie die Wahrscheinlichkeit), verwenden andere ein ausgewogeneres Gewichtungssystem, was zu Unstimmigkeiten bei der Priorisierung von Risiken führt.

Zusätzlich zu dieser Komplexität unterscheiden sich die Plattformen erheblich in der Art und Weise, wie sie das Risikoniveau nach der Analyse von Schweregrad und Wahrscheinlichkeit bewerten. Dabei werden neben unterschiedlichen Metriken auch unterschiedliche Skalen verwendet. So verwendet AliExpress beispielsweise eine Risikomatrix zur Bewertung und Kategorisierung von Risiken, während Meta eine fünfstufige Schweregrad-Rubrik verwendet. In Metas Modell steht Stufe 1 für Risiken mit extrem niedrigem Schweregrad und geringer Wahrscheinlichkeit, während Stufe 5 Risiken mit extrem hohem Schweregrad kennzeichnet. TikTok hingegen verwendet ein einfacheres dreistufiges System, bei dem Risiken der Stufe 1 für eine sofortige Abschwächung priorisiert werden und Risiken der Stufe 3 eine niedrigere

Priorität zugewiesen wird. Diese unterschiedlichen Ansätze unterstreichen das Fehlen eines einheitlichen Rahmens und erschweren vergleichende Analysen zwischen den Plattformen (Broughton Micova, 2024). Einige Plattformen unterschieden in der Analyse zwischen einem inhärenten Risiko ohne Einbeziehung der Minderungsmaßnahmen und einem verbleibendem Risiko, welches jene Maßnahmen bereits in die Berechnung einbezieht (Broughton Micova, 2024).

Empirische Grundlagen in öffentlichen Berichten

Eine grundsätzliche Herausforderung im Umgang mit den Risikobewertungen besteht darin, dass die Plattformen zwar die Ergebnisse, und teils auch ihre Methodik zur Einstufung der Risiken darlegen, jedoch die konkreten nicht Teil der veröffentlichten Berichte sind. Viele der VLOPs und VLOSEs haben dabei von ihrem Recht unter Art. 42 Absatz 5 Gebrauch gemacht und entweder eine nicht vertrauliche Version des Risikobewertungsberichts veröffentlicht (z.B. Aliexpress und Apple) oder bestimmte Informationen aus dem Bericht entfernt (z.B. Amazon, Meta und X). Dies erschwert einerseits eine externe Überprüfung, kann aber anderseits als Chance für Wissenschaftler*innen gesehen werden, gezielten Datenzugang nach Art. 40 zu beantragen (Broughton Micova, 2024). Die Risikobewertungen selbst geben jedoch wenig Einblick in die empirischen Grundlagen, auf denen die konkrete Risikoeinstufung beruht.

Variabilität der Risikokategorien und Auslegungen

Die vier in Artikel 34 Absatz 1 des DSA genannten Bereiche des systemischen Risikos werden von den einzelnen Plattformen ebenfalls sehr unterschiedlich interpretiert, was den Vergleich zusätzlich erschwert. X und TikTok gehen beispielsweise explizit auf »Risiken terroristischer Inhalte« ein, während Meta weiter gefasste Kategorien wie »Koordinierung von Schaden und Förderung von Verbrechen« und »gefährliche Organisationen und Personen« umfasst. Diese Diskrepanzen erschweren es Forschenden, zu beurteilen, wie Plattformen mit ähnlichen Risiken umgehen oder die Wirksamkeit ihrer Maßnahmen zur Risikominderung zu bewerten. Eine solche Fragmentierung in der Kategorisierung behindert nicht nur Meta-Analysen, sondern wirft auch Fragen nach der Konsistenz und Übereinstimmung mit den Zielen des DSA auf (Broughton Micova, 2024).

Risikobewertungen als »Tick-Box Exercise«?

Plattformen bewerten zumeist Risiken, für welche sie bereits Maßnahmen getroffen haben, und beziehen sich dabei allen hauptsächlich auf die eigenen Nutzungsrichtlinien. Tieferliegende Risiken welche sich aus dem Plattformdesign ergeben, wie suchtbezogene Nutzungsrisiken, oder Risiken von auf »Engagement« aufgebauten Empfehlungssystemen werden im Großen und Ganzen nicht erfasst (Bernard, 2024). Zivilgesellschaftliche Organisationen hatten bereits im Vorlauf der Veröffentlichung der Risikoberichte gewarnt, dass diese mehr sein sollten als lediglich eine »Tick-Box« Exercise (Center for Democracy and Technology, 2024). Diese Befürchtungen scheinen sich jedoch weitgehend bestätigt zu haben. Beobachter wiesen darauf hin, dass die Risikobewertungen großteils lediglich einen Einblick in bereits existierende Inhaltsrichtlinien und Nutzungsbedingungen bieten, jedoch keine tiefergehenden Informationen über die Funktionsweise der Plattformen beinhalten (Moraht, 2024; Scott, 2024). Statt einen ganzheitlichen Ansatz zu verfolgen, fokussieren sich die Plattformen stark auf die eigenen Nutzungsrichtlinien und die Durchsetzung von Inhaltsmoderationsmaßnahmen.

Risikobewertungen ohne klare Definitionen und messbare Metriken können sich nachteilig auf andere Grundrechte auswirken (Pielemeier et al., 2024). Dies spiegelt sich in den Risikobewertungen wider, welche negative Auswirkungen von Minderungsmaßnahmen auf andere Grundrechte nur oberflächlich berücksichtigen. Zur Bewältigung von systemischen Risiken wird dabei oft auf Inhaltemoderation wie das Entfernen von Inhalten oder Nutzerkonten oder Einschränken dieser zurückgegriffen, ohne dabei das Plattformdesign oder algorithmische Systeme in einem ganzheitlichen Ansatz zu analysieren. Während die Plattformen zumeist versichern, die Angemessenheit und Verhältnismäßigkeit in ihren Maßnahmen zu berücksichtigen, mangelt es an Details wie dies bewertet wurde und welche anderen Maßnahmen möglicherweise zur Verfügung standen, die Prüfung jedoch nicht bestanden haben.

Einbindung von Wissenschaftler*innen, Zivilgesellschaft und betroffenen Gruppen

Nach Erwägungsgrund 90 DSA, sollten von den Risiken besonders betroffene Gruppen sowie unabhängige

Sachverständige und zivilgesellschaftliche Organisationen in den Prozess der Risikobewertungen und der Gestaltung der Minderungsmaßnahmen einbezogen werden. Zivilgesellschaftliche Gruppen weisen darauf hin, dass dies aus den aktuellen Berichten nicht ausreichend deutlich wird (AlgorithmWatch, 2024) oder dass sie nicht in den Prozess involviert wurden (People vs Big Tech, 2024). Für künftige Risikobewertungen könnte eine Institutionalisierung iterativer und integrativer Prozesse zur Einbindung relevanter Gruppen sicherstellen, dass Risikominderungsmaßnahmen akkurat, effektiv und zuverlässig sind (GNI & DTSP, 2023).

8.2 Fazit und Limitationen

Obgleich Risikobewertungen bereits in anderen Sektoren angewandt werden und der DSA einige richtungsweisende Hinweise beinhaltet, fehlt es an Klarheit in Bezug auf die Definition bestimmter Rechtsbegriffe wie »Systemische Risiken« sowie auf zugehörige Indikatoren, Messmethoden und Gute Praktiken der Risikobewertung und -minderung.

Der im Kontext dieser Studie entwickelte Bewertungsrahmen ist ein Beitrag zur Schließung dieser Lücke, indem er hilfreiche Elemente bisheriger Risikomanagementansätze adaptiert und einen allgemein übertragbaren und praktisch anwendbaren Ansatz vorlegt. Basierend auf einer Spezifikation der Risikokategorien im DSA und einer Arbeitsdefinition von systemischen Risiken wurde ein dreistufiger Bewertungsprozess (Risikoidentifikation – Risikobewertung – Risikominderung) entwickelt. Dessen Kernelement stellt die Bewertung systemischer Risiken anhand der Bewertungsmaßstäbe »Auswirkungen« und »Wahrscheinlichkeit« über die ISD-Risikoebenen (nutzungsbezogene Risiken – inhaltsbezogene Risiken – verhaltensbezogene Risiken) hinweg dar. Die klare Trennung in diese drei Ebenen ermöglicht es, sowohl die zugrunde liegenden Ursachen als auch die spezifischen Dynamiken systemischer Risiken präzise zu analysieren. Dadurch wird eine zielgerichtete Entwicklung von Indikatoren und Maßnahmen ermöglicht, die sich passgenau auf die jeweilige Risikoart anwenden lassen. Für die Ermittlung von Kernrisiken zu Beginn des Bewertungsprozesses wurde zudem ein allgemeines Risikoprofil einschließlich eines Katalogs mit Leitfragen zur Bestimmung von spezifischen Risikofaktoren erarbeitet. Trotz dieser umfassenden konzeptionellen Arbeit unterliegt der Bewertungsrahmen jedoch auch einigen Limitationen, die bei nachfolgenden Forschungsarbeiten oder der praktischen Anwendung berücksichtigt werden sollten.

Erstens stellt der Bewertungsrahmen nur eine allgemeine Hilfestellung für Bewertungsprozesse in der Praxis dar. Die Feststellung spezifischer systemischer Risiken ist dienst- und kontextgebunden. Dabei können die im Rahmen der Studie identifizierten Indikatoren je nach Anwendungsfall auch variieren. Beispielsweise existieren Dienste, die keine algorithmischen Empfehlungssysteme einsetzen. Stattdessen können sie aber andere wichtige Plattformmerkmale aufweisen, die systemische Risiken beeinflussen können. Folglich sollte die bereitgestellte Indikatoren-Basis durch weitere sinnvolle Indikatoren kontinuierlich ergänzt werden. Außerdem können die Indikatoren auf unterschiedlichen Wegen gemessen werden. Diese Herausforderung zeigt sich bereits an unterschiedlichen Methoden zur Ermittlung der Nutzerzahl. Unterschiedliche Funktionen und Strukturen digitaler Dienste erfordern außerdem eine jeweils an den Dienst angepasste Indikatoren-Messung, was wiederum die Vergleichbarkeit von systemischen Risiken auf unterschiedlichen VLOPs und VLOSEs beeinträchtigen kann.

Zweitens enthält der Bewertungsrahmen keine Grenzwerte für die identifizierten Indikatoren, wie beispielsweise eine spezifische Zahl, die die Prävalenz von rechtswidrigen Inhalten ausdrückt. Dies liegt unter anderem daran, dass es oft keinen wissenschaftlichen Konsens darüber gibt, wie die spezifischen Phänomene (z.B. virale Desinformation) gemessen werden sollten und welche Werte als nicht akzeptabel gelten. Die Festlegung von Grenzwerten erfordert jedoch eindeutig definierte Indikatoren, standardisierte Messmethoden und einen klar definierten Toleranzrahmen. Zum anderen verändern sich Dienste im Zuge neuer technologischer Innovationen und des globalen Wettbewerbs fortlaufend. Feste Grenzwerte würden daher der Dynamik des Online-Umfelds nicht gerecht werden, selbst wenn sie eine große Relevanz für die Regulierungspraxis darstellen. Zugleich könnte eine übermäßige Quantifizierung von Indikatoren auch dazu führen, dass sich Risikobewertungen vor allem auf leicht messbare oder vorteilhafte Grenzwerte fokussieren. Qualitative Indikatoren könnten dann möglicherweise nicht mehr die notwendige Beachtung im Risikomanagement erhalten.

Eine dritte Limitation ergibt sich aus der erschwerten Messbarkeit der Wirksamkeit von Risikominderungsmaßnahmen. Zum einen können diese nicht isoliert von anderen Einflussfaktoren betrachtet werden. Es bleibt häufig unklar, ob eine Maßnahme tatsächlich für das Ausbleiben eines Schadens verantwortlich ist, wodurch eine präzise Evaluation ihrer Wirksamkeit kaum möglich ist. Zum anderen wird die Einschätzung zusätzlich durch den begrenzten Zugang zu Informationen über bereits ergriffene Maßnahmen behindert. Hinzu kommt, dass bei der Auswahl von Maßnahmen stets die Gefahr besteht, in schützenswerte Rechtsgüter einzugreifen. Die Entscheidung über verhältnismäßige und notwendige Maßnahmen ist daher häufig von individuellen Abwägungen abhängig, was universelle Lösungen oder standardisierte Ansätze erschwert. Aus diesem Grund können Gute Praktiken nur begrenzt abgeleitet werden. Die vorliegende Studie verweist daher auf bewährte Praktiken, die als flexible Ansätze dienen und kontinuierlich weiterentwickelt werden können. Diese Herangehensweise erscheint sinnvoll, da sie die Dynamik von Online-Plattformen, den jeweiligen Kontext und den stetigen Wandel der funktionalen Rahmenbedingungen berücksichtigt.

Viertens macht diese Studie keine konkreten Aussagen zur prozessualen Einrichtung und Verankerung der Risikobewertung und -minderung nach dem DSA im Unternehmen, zu den Grundsätzen der Durchführung von Risikobewertungen oder zur Art und Weise, wie externe Gruppen identifiziert und in den Bewertungsprozess einbezogen werden sollten. Alle Aspekte sind jedoch zentral, damit die zusätzlichen Sorgfaltspflichten in die Aktivitäten und Organisationsstrukturen der Anbieter*innen von VLOPs und VLOSEs angemessen integriert werden sowie effizient und auf Grundlage der besten verfügbaren Informationen ablaufen. In diesem Zusammenhang liefern bisherige Risikomanagementansätze in Organisationen oder etablierte Prinzipien menschenrechtsbasierter Folgeabschätzungen bereits Anknüpfungspunkte. Eine Standardisierung im DSA-Kontext ist jedoch empfehlenswert.

8.3 Ausblick

Ende November 2024 veröffentlichten die designierten Anbieter*innen von VLOPs und VLOSEs zum ersten Mal ihre Berichte über die Ergebnisse der Risikobewertungen (Art. 42 Abs. 4 a) DSA). Obwohl sie dabei vertrauliche, sicherheitsrelevante oder für Nutzer*innen schädliche Informationen aus den Berichten entfernen durften (Art. 42 Abs. 5 DSA), waren sie dazu angehalten, über die vorgenommenen Risikobewertungen umfassend Bericht erstatten (Erwägungsgrund 100 DSA). Eine umfassende Berichterstattung ist vor allem deshalb wichtig, damit auch externe Forschende ihre Beiträge zur Aufspürung, zur Ermittlung und zum Verständnis systemischer Risiken fokussieren und weiterentwickeln können. Dadurch können ihre Beiträge noch relevanter für die Durchführung von Risikobewertungen werden, in die sie künftig noch enger im Sinne von Erwägungsgrund 90 DSA einbezogen werden sollten.

Damit Risikobewertung und -minderung im Rahmen einer Multi-Stakeholder-Durchsetzungsstruktur überprüft und kontinuierlich verbessert werden können, ist es unerlässlich, dass die Anbieter*innen von VLOPs und VLOSEs die Konsultation von betroffenen Gruppen, unabhängigen Sachverständigen und zivilgesellschaftlichen Organisationen in ihre Methoden der Bewertung und Minderung von Risiken integrieren. Auf der anderen Seite können Forschende selbst von der Konsultation der Anbieter*innen, unabhängigen Prüfer*innen oder Regulierungsbehörden in Forschungsprojekten profitieren. Denn nach wie vor sind interne Einblicke und Produktkenntnisse unentbehrlich, um den Einfluss von Diensten und ihrer Nutzung auf systemische Risiken beurteilen zu können. Zwar umfasst der DSA neue Transparenzpflichten, wie beispielsweise die Darlegung der wichtigsten Parameter von Empfehlungssystemen (Art. 27 Abs. 1 DSA). Dennoch erfordert gerade die Bewertung und Minderung von spezifischen systemischen Risiken im Anwendungsfall einen vertieften und konstruktiven Austausch.

Die durch solche Austauschformate gewonnenen Erkenntnisse können wiederum von der Europäischen Kommission aufgegriffen und in Zusammenarbeit mit den nationalen DSCs in Leitlinien für die Risikominderung in Bezug auf besondere Risiken überführt und regelmäßig aktualisiert werden (Art. 35 Abs. 3 DSA). Ein jährlicher Bericht des Europäischen Gremiums für digitale Dienste in Zusammenarbeit mit der Europäischen Kommission gibt zudem Auskunft über die auffälligsten wiederkehrenden systemischen Risiken und bewährte Verfahren zur Minderung dieser Risiken (Art. 35 Abs. 2 DSA). Darüber hinaus sieht der DSA vor, dass die Europä-

ische Kommission und das Gremium die Ausarbeitung von freiwilligen Verhaltenskodizes auf Unionsebene als Beitrag zur DSA-Durchsetzung fördern und erleichtern (Art. 45 Abs. 1 DSA). Im Sinne einer Ko-Regulierung können solche Kodizes als geeignete Risikominderungsmaßnahme angesehen werden. Im Gegenzug kann auch eine Nicht-Beteiligung bei der Feststellung möglicher Zuwiderhandlungen berücksichtigt werden (Erwägungsgrund 104 DSA). Dennoch besteht keine Verpflichtung für die Anbieter*innen von VLOPs und VLOSEs zur Umsetzung von Leitlinien oder Verhaltenskodizes.

Schlussfolgernd unterliegt die Auslegung bestimmter Rechtsbegriffe wie »Systemische Risiken« oder zugehöriger Indikatoren, Messmethoden und Gute Praktiken der Risikobewertung und -minderung trotz einschlägiger aktueller und künftiger Leitlinien und Verhaltenskodizes vor allem einem gemeinsamen Lernprozess, in dem unterschiedliche Interessen ausbalanciert werden müssen. In diesem Sinne ist auch der im Zusammenhang mit dieser Studie entwickelte Bewertungsrahmen oder die Arbeitsdefinition von systemischen Risiken im DSA-Kontext ein Diskussionsbeitrag.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Aalbers, G., McNally, R. J., Heeren, A., de Wit, S., & Fried, E. I. (2019). Social media and depression symptoms: A network perspective. *Journal of Experimental Psychology: General, 148*(8), 1454–1462. https://doi.org/10.1037/xge0000528

AccessNow, & ECNL. (2023, September 18). *Towards Meaningful Fundamental Rights Impact Assessments under the DSA*. https://www.accessnow.org/fundamental-rights-impact-assessments-for-dsa-enforcement/

Alberto, S., Dakanalis, A., Mura, M., Colmegna, F., & Clerici, M. (2022). Instagram Use and Mental Well-Being: The Mediating Role of Social Comparison. *The Journal of Nervous and Mental Disease, 210*(12). https://doi.org/10.1097/NMD.0000000000001577

AlgorithmWatch. (2024). *DSA: Erste Risikobewertungsberichte über systemische Risiken von großen Online-Plattformen lassen viele Fragen offen*. https://algorithmwatch.org/de/dsa-risikobewertungsberichte/

Allem, J.-P., & Ferrara, E. (2018). Could Social Bots Pose a Threat to Public Health? *American Journal of Public Health, 108*(8), 1005–1006. https://doi.org/10.2105/AJPH.2018.304512

Allen, A. (2022). An Intersectional Lens on Online Gender Based Violence and the Digital Services Act. *Verfassungsblog: On Matters Constitutional*. https://doi.org/10.17176/20221101-215626-0

Altitude. (2024). Altitude. https://altitude.google.com/

Arendt, F., Scherr, S., & Romer, D. (2019). Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults. *New Media & Society, 21*(11–12), 2422–2442. https://doi.org/10.1177/1461444819850106

Ashraf, A. R., Mackey, T. K., & Fittler, A. (2024). Search Engines and Generative Artificial Intelligence Integration: Public Health Risks and Recommendations to Safeguard Consumers Online. *JMIR Public Health and Surveillance, 10*. https://doi.org/10.2196/53086

Ashurst, L., & McAlinden, A.-M. (2015). Young people, peer-to-peer grooming and sexual offending: Understanding and responding to harmful sexual behaviour within a social media society. *Probation Journal, 62*(4), 374–388. https://doi.org/10.1177/0264550515619572

Barthorpe, A., Winstone, L., Mars, B., & Moran, P. (2020). Is social media screen time really associated with poor adolescent mental health? A time use diary study. *Journal of Affective Disorders, 274*, 864–870. https://doi.org/10.1016/j.jad.2020.05.106

Bengani, P., Thorburn, L., & Stray, J. (2022, August 8). A Menu of Recommender Transparency Options. *Tech Policy Press*. https://techpolicy.press/a-menu-of-recommender-transparency-options

Bernstein, S., Warburton, W., Bussey, K., & Sweller, N. (2023). Mind the Gap: Internet Pornography Exposure, Influence and Problematic Viewing Amongst Emerging Adults. *Sexuality Research and Social Policy, 20*(2), 599–613. https://doi.org/10.1007/s13178-022-00698-8

BEUC. (2024). *Taming Temu: Why the fast-growing online marketplace fails to comply with the EU Digital Services Act*. https://www.beuc.eu/reports/taming-temu-why-fast-growing-online-marketplace-fails-comply-eu-digital-services-act

Beyens, I., Pouwels, J. L., van Driel, I. I., Keijsers, L., & Valkenburg, P. M. (2020). The effect of social media on well-being differs from adolescent to adolescent. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-020-67727-7

Bhargava, V. R., & Velasquez, M. (2021). Ethics of the Attention Economy: The Problem of Social Media Addiction. *Business Ethics Quarterly, 31*(3), 321–359. https://doi.org/10.1017/beq.2020.32

Bloch Veiberg, C. (2023). *Key Principles for Human Rights Impact Assessment of Digital Business Activities*. The Danish Institute for Human Rights. https://www.humanrights.dk/files/media/document/Key%20Principles%20for%20HRIA%20of%20digital%20business%20activities.pdf

BMFSFJ. (2024). *Kinder- und Jugendschutz*. BMFSFJ. https://www.bmfsfj.de/bmfsfj/themen/kinder-und-jugend/kinder-und-jugendschutz

Boer, M. (2022). *#ConnectedTeens: Social media use and adolescent wellbeing* [Doctoral thesis 1 (Research UU / Graduation UU), Utrecht University]. https://doi.org/10.33540/1272

Borgogna, N. C., Duncan, J., & McDermott, R. C. (2018). Is scrupulosity behind the relationship between problematic pornography viewing and depression, anxiety, and stress? *Sexual Addiction & Compulsivity, 25*(4), 293–318. https://doi.org/10.1080/10720162.2019.1567410

Brailovskaia, J., & Margraf, J. (2023). Less sense of control, more anxiety, and addictive social media use: Cohort trends in German university freshmen between 2019 and 2021. *Current Research in Behavioral Sciences, 4*. https://doi.org/10.1016/j.crbeha.2022.100088

Broughton Micova, S. (2024, Dezember 6). *Evaluating systemic risk management under the DSA*. CERRE. https://cerre.eu/news/evaluating-systemic-risk-management-under-the-dsa/

Broughton Micova, S., & Calef, A. (2023, Juli). *Elements for Effective Systemic Risk Assessment Under the DSA*. Centre on Regulation in Europe (CERRE). https://www.ssrn.com/abstract=4512640

Brown, M., Bisbee, J., Lai, A., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4114905

Brühwiler, B., & Romeike, F. (2010). *Praxisleitfaden Risikomanagement: ISO 31000 und ONR 49000 sicher anwenden*. Erich Schmidt Verlag.

Bundesamt für Sicherheit in der Informationstechnik. (2024). *Exkurs: Social Bots und Chatbots*. https://www.bsi.bund.de/DE/Themen/Verbraucherinnen-und-Verbraucher/Informationen-und-Empfehlungen/Onlinekommunikation/Soziale-Netzwerke/Sichere-Verwendung/Exkurs-bots/social-bots.html?nn=896986

Bundeskriminalamt. (2024). *BKA - Cybergrooming*. https://www.bka.de/DE/UnsereAufgaben/Aufgabenbereiche/Zentralstellen/Kinderpornografie/Cybergrooming/Cybergrooming_node.html

Burke, M., & Kraut, R. E. (2016). The Relationship Between Facebook Use and Well-Being Depends on Communication Type and Tie Strength. *Journal of Computer-Mediated Communication, 21*(4), 265–281. https://doi.org/10.1111/jcc4.12162

Burnell, K., George, M. J., Vollet, J. W., Ehrenreich, S. E., & Underwood, M. K. (2019). Passive social networking site use and well-being: The mediating roles of social comparison and the fear of missing out. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 13*(3). https://doi.org/10.5817/CP2019-3-5

Calabrese, S., & Reich, O. (2024, Januar). *Identifying, Analysing, Assessing and Mitigating Potential Negative Effects on Civic Discourse and Electoral Processes. A Minimum Menu of Risks Very Large Online Platforms Should Take Head Of*. Civil Liberties Union for Europe. https://dq4n3btxmr8c9.cloudfront.net/files/mpdgy5/DSA_Risk_Analysis_LibertiesxEPDfin.pdf

Camilleri, C., Perry, J. T., & Sammut, S. (2021). Compulsive Internet Pornography Use and Mental Health: A Cross-Sectional Study in a Sample of University Students in the United States. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.613244

Cano, A. E., Fernandez, M., & Alani, H. (2014). Detecting Child Grooming Behaviour Patterns on Social Media. In L. M. Aiello & D. McFarland (Hrsg.), *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings* (S. 412–427). Springer International Publishing. https://doi.org/10.1007/978-3-319-13734-6_30

CeMAS. (2024). *Telegram: Chronologie einer Radikalisierung*. Telegram: Chronologie einer Radikalisierung. https://report.cemas.io/telegram/

Center for Democracy and Technology. (2024, November 8). *Joint Civil Society Statement on Meaningful Transparency of Risk Assessments under the Digital Services Act*. https://cdt.org/insights/joint-civil-society-statement-on-meaningful-transparency-of-risk-assessments-under-the-digital-services-act/

Chen, A. Y., Nyhan, B., Reifler, J., Robertson, R. E., & Wilson, C. (2023). *Subscriptions and external links help drive resentful users to alternative and extremist YouTube videos*. https://doi.org/10.48550/arXiv.2204.10921

Children's Commissioner. (2023). *'A lot of it is actually just abuse'- Young people and pornography*. https://www.childrenscommissioner.gov.uk/resource/a-lot-of-it-is-actually-just-abuse-young-people-and-pornography/

Council for Media Services & Trust Lab. (2023). *The prevalence of harmful or potentially illegal content on digital platforms following the Bratislava terrorist attack*. European Platform of Regulatory Authorities. https://www.epra.org/news_items/the-role-of-online-platforms-in-harmful-content-slovak-regulator-investigates-user-s-report-mechanisms

Coyne, S. M., Rogers, A. A., Zurcher, J. D., Stockdale, L., & Booth, M. (2020). Does time spent using social media impact mental health? An eight year longitudinal study. *Computers in Human Behavior, 104*. https://www.sciencedirect.com/science/article/pii/S0747563219303723

Das, S., Lavoie, A., & Magdon-Ismail, M. (2016). Manipulation among the Arbiters of Collective Intelligence: How Wikipedia Administrators Mold Public Opinion. *ACM Trans. Web, 10*(4), 24:1-24:25. https://doi.org/10.1145/3001937

Denning, P., Horning, J., Parnas, D., & Weinstein, L. (2005). Wikipedia risks. *Commun. ACM, 48*(12), 152. https://doi.org/10.1145/1101779.1101804

Deutscher Bundestag. (2022). *„Echokammern" und „Filterblasen" in digitalen Medien*. https://www.bundestag.de/resource/blob/898208/396d70db93fbc68bca40726b4d5308db/WD-10-007-22-pdf-data.pdf

DGCN. (2014, Juni). *Leitprinzipien für Wirtschaft und Menschenrechte. Umsetzung des Rahmens der Vereinten Nationen „Schutz, Achtung und Abhilfe"*. Deutsches Global Compact Netzwerk (DGCN), Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ). https://www.auswaertiges-amt.de/blob/266624/b51c16faf1b3424d7efa060e8aaa8130/un-leitprinzipien-de-data.pdf

Din Media. (2024). *Risiken identifizieren und vermeiden – mit DIN ISO 31000*. https://www.dinmedia.de/de/themenseiten/resilienz-in-unternehmen/risikomanagement

Djeffal, C. (2022, November 1). Is the DSA Revolutionizing Algorithmic Risk Governance? *Heinrich Böll Stiftung*. https://il.boell.org/en/2022/11/01/dsa-revolutionizing-algorithmic-risk-governance

Dreier, M., Wölfling, K., Duven, E., Giralt, S., Beutel, M. E., & Müller, K. W. (2017). Free-to-play: About addicted Whales, at risk Dolphins and healthy Minnows. Monetarization design and Internet Gaming Disorder. *Addictive Behaviors, 64*, 328–333. https://doi.org/10.1016/j.addbeh.2016.03.008

Ebert, I., Wildhaber, I., Shakil, M., & De Oliveira, A. (2023). The Business & Human Rights Dimension of the Digital Services Act. *Gesellschaft Für Freiheitsrechte*. https://freiheitsrechte.org/uploads/publications/Digital/Grundrechte-im-Digitalen/The-Business-Human-Rights-Dimension-of-the-Digital-Services-Act.pdf

Edwards, M., & Hollely, N. M. (2023). Online sextortion: Characteristics of offences from a decade of community reporting. *Journal of Economic Criminology, 2*. https://doi.org/10.1016/j.jeconc.2023.100038

Engeln, R., Loach, R., Imundo, M. N., & Zola, A. (2020). Compared to Facebook, Instagram use causes more appearance comparison and lower body satisfaction in college women. *Body Image, 34*, 38–45. https://doi.org/10.1016/j.bodyim.2020.04.007

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management, 55*, 62–73. https://doi.org/10.1016/j.tourman.2016.01.013

Europäische Kommission. (2020, Dezember 3). *Europäischer Aktionsplan für Demokratie*. https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:52020DC0790

Europäische Kommission. (2022). *The 2022 Code of Practice on Disinformation*. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

Europäische Kommission. (2023, Januar 31). *DSA: Guidance on the requirement to publish user numbers | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/library/dsa-guidance-requirement-publish-user-numbers

Europäische Kommission. (2024a). *Deep Learning*. https://open-research-europe.ec.europa.eu/collections/deep-learning/about

Europäische Kommission. (2024b). *EU Code of Practice on Disinformation*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/new-push-european-democracy/protecting-democracy/strengthened-eu-code-practice-disinformation_en

Europäische Kommission. (2024c). *Leitlinien der Kommission für Anbieter sehr großer Online-Plattformen und sehr großer Online-Suchmaschinen zur Minderung systemischer Risiken in Wahlprozessen gemäß Artikel 35 Absatz 3 der Verordnung (EU) 2022/2065*. https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:C_202403014

Europäische Kommission. (2024d). *Schutz Minderjähriger im Internet: EU-Kommission startet Sondierung zu DSA-Leitlinien*. https://germany.representation.ec.europa.eu/news/schutz-minderjahriger-im-internet-eu-kommission-startet-sondierung-zu-dsa-leitlinien-2024-07-31_de

Europäische Kommission. (2024e, Oktober). *Kommission leitet förmliches Verfahren gegen Temu nach dem Gesetz über digitale Dienste ein*. https://digital-strategy.ec.europa.eu/de/news/commission-opens-formal-proceedings-against-temu-under-digital-services-act

European Data Protection Board. (2022, März 15). *EDSA nimmt Leitlinien zu Artikel 60 DSGVO und zu Dark Patterns auf der Benutzeroberfläche von Social-Media-Plattformen sowie eine „Toolbox" für wesentliche Datenschutzgarantien bei der Zusammenarbeit zwischen dem EWR und Aufsichtsbehörden von Drittländern*. https://www.edpb.europa.eu/news/news/2022/edpb-adopts-guidelines-art-60-gdpr-guidelines-dark-patterns-social-media-platform_de

European External Action Service. (2023, Februar 7). *1st EEAS Report on Foreign Information Manipulation and Interference Threats*. https://www.eeas.europa.eu/eeas/1st-eeas-report-foreign-information-manipulation-and-interference-threats_en

European Institute for Gender Equality. (2024a). *EU gender-based violence survey: Key results*. https://eige.europa.eu/publications-resources/publications/eu-gender-based-violence-survey-key-results?language_content_entity=en

European Institute for Gender Equality. (2024b, August). *Revenge porn*. https://eige.europa.eu/publications-resources/thesaurus/terms/1459?language_content_entity=en

European Parliament. (2021). *Briefing: Vulnerable consumers*. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/690619/EPRS_BRI(2021)690619_EN.pdf

Faelens, L., Hoorelbeke, K., Fried, E., De Raedt, R., & Koster, E. H. W. (2019). Negative influences of Facebook use through the lens of network analysis. *Computers in Human Behavior, 96*, 13–22. https://doi.org/10.1016/j.chb.2019.02.002

Fardouly, J., Magson, N. R., Rapee, R. M., Johnco, C. J., & Oar, E. L. (2020). The use of social media by Australian preadolescents and its links with mental health. *Journal of Clinical Psychology, 76*(7), 1304–1326. https://doi.org/10.1002/jclp.22936

Ferrucci, P., & Hopp, T. (2023). Let's intervene: How platforms can combine media literacy and self-efficacy to fight fake news. *Communication and the Public, 8*(4), 367–389. https://doi.org/10.1177/20570473231203081

Geeng, C., Francisco, T., West, J., & Roesner, F. (2020). *Social Media COVID-19 Misinformation Interventions Viewed Positively, But Have Limited Impact*. https://doi.org/10.48550/arXiv.2012.11055

Generaldirektion CNECT. (2023). *Digital Services Act: Application of the risk management framework to Russian disinformation campaigns*. Publications Office of the European Union. https://data.europa.eu/doi/10.2759/764631

Gerster, L., Kuchta, R., Hammer, D., & Schwieter, C. (2021, Dezember 17). *Stützpfeiler Telegram: Wie Rechtsextreme und Verschwörungsideolog:innen auf Telegram ihre Infrastruktur ausbauen*. ISD. https://isdgermany.org/stuetzpfeiler-telegram-wie-rechtsextreme-und-verschwoerungsideologinnen-auf-telegram-ihre-infrastruktur-ausbauen/

Gibson, C., Olszewski, D., Brigham, N. G., Crowder, A., Butler, K. R. B., Traynor, P., Redmiles, E. M., & Kohno, T. (2024, November). *Analyzing the AI Nudification Application Ecosystem*. https://arxiv.org/html/2411.09751v1

Global Online Safety Regulators Network. (2024). *Regulatory Index: Comparing International Approaches and Perspectives to Online Safety Regulation*. https://www.esafety.gov.au/sites/default/files/2024-10/GOSRN-Regulatory-Index-2024-final.pdf

GNI, & DTSP. (2023). *Implementing risk assessments under the Digital Services Act*. https://globalnetworkinitiative.org/wp-content/uploads/2023/06/Discussion-summary-%E2%80%93-GNI-and-DTSP-workshops-on-implementing-risk-assessments-under-the-DSA-June-2023.pdf

Goldman, E. (2021). *Content Moderation Remedies*. Social Science Research Network. https://doi.org/10.2139/ssrn.3810580

Gomez, M., Klare, D., Ceballos, N., Dailey, S., Kaiser, S., & Howard, K. (2022). Do You Dare to Compare?: The Key Characteristics of Social Media Users Who Frequently Make Online Upward Social Comparisons. *International Journal of Human–Computer Interaction, 38*(10), 938–948. https://doi.org/10.1080/10447318.2021.1976510

Gongane, V. U., Munot, M. V., & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: Current status and future directions. *Social Network Analysis and Mining, 12*(1), 129. https://doi.org/10.1007/s13278-022-00951-3

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), 1–15. https://doi.org/10.1177/2053951719897945

Gov.uk. (2019). *State of the nation 2019: Children and young people's wellbeing*. https://www.gov.uk/government/publications/state-of-the-nation-2019-children-and-young-peoples-wellbeing

Grabowski, J., & Klein, S. (2023). Wikipedia's Intentional Distortion of the History of the Holocaust. *The Journal of Holocaust Research, 37*(2), 133–190. https://doi.org/10.1080/25785648.2023.2168939

Gregoire, A. (2021, April). CIB Detection Tree: 1st Branch. *EU DisinfoLab*. https://www.disinfo.eu/publications/cib-detection-tree1/

Griffioen, N., Scholten, H., Lichtwarck-Aschoff, A., Maciejewski, D., & Granic, I. (2023). Heterogeneity in some relationships between social media use and emerging adults' affective wellbeing. *Current Psychology, 42*(34), 30277–30292. https://doi.org/10.1007/s12144-022-04035-5

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Haidt, J., Rausch, Z., & Twenge, J. (o. J.). *Social Media and Mental Health*. Abgerufen 26. November 2024, von https://docs.google.com/document/d/1w-HOfseF2wF9YIpXwUUtP65-olnkPyWcgF5BiAtBEy0/edit?usp=embed_facebook

Hanna, E., Ward, L. M., Seabrook, R. C., Jerald, M., Reed, L., Giaccardi, S., & Lippman, J. R. (2017). Contributions of Social Comparison and Self-Objectification in Mediating Associations Between Facebook Use and Emergent Adults' Psychological Well-Being. *Cyberpsychology, Behavior, and Social Networking, 20*(3), 172–179. https://doi.org/10.1089/cyber.2016.0247

Haroon, M., Wojcieszak, M., Chhabra, A., Liu, X., Mohapatra, P., & Shafiq, Z. (2023). Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences, 120*(50). https://doi.org/10.1073/pnas.2213020120

Hawes, T., Zimmer-Gembeck, M. J., & Campbell, S. M. (2020). Unique associations of social media use and online appearance preoccupation with depression, anxiety, and appearance rejection sensitivity. *Body Image, 33*, 66–76. https://doi.org/10.1016/j.bodyim.2020.02.010

Hendrix, J., & Jahangir, R. (2024, September 15). *Understanding Systemic Risks under the Digital Services Act*. Tech Policy Press. https://techpolicy.press/understanding-systemic-risks-under-the-digital-services-act

Henry, N., & Beard, G. (2024). Image-Based Sexual Abuse Perpetration: A Scoping Review. *Trauma, Violence & Abuse, 25*(5), 3981–3998. https://doi.org/10.1177/15248380241266137

Hohenwalde, C. E. (2023). *Rechtsextremismus auf Telegram: Eine Netzwerkanalyse*. Karlsruher Institut für Technologie (KIT). https://doi.org/10.5445/IR/1000174284

Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW1), 1–27. https://doi.org/10.1145/3392854

IBM. (2024, November 7). *What is Red Teaming?* https://www.ibm.com/think/topics/red-teaming

ISO. (2018, Februar). *Risk management Guidelines*. https://www.iso.org/standard/65694.html

Jarman, H. K., Marques, M. D., McLean, S. A., Slater, A., & Paxton, S. J. (2021). Social media, body satisfaction and well-being among adolescents: A mediation model of appearance-ideal internalization and comparison. *Body Image, 36*, 139–148. https://doi.org/10.1016/j.bodyim.2020.11.005

Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (2021). Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction, 5*, 1–30. https://doi.org/10.1145/3479525

Jiang, J., & Vetter, M. A. (2020). The Good, the Bot, and the Ugly: Problematic Information and Critical Media Literacy in the Postdigital Era. *Postdigital Science and Education, 2*(1), 78–94. https://doi.org/10.1007/s42438-019-00069-4

Keller, T. R., & Klinger, U. (2019). Social Bots in Election Campaigns: Theoretical, Empirical, and Methodological Implications. *Political Communication, 36*(1), 171–189. https://doi.org/10.1080/10584609.2018.1526238

Kelly, Y., Zilanawala, A., Booker, C., & Sacker, A. (2018). Social Media Use and Adolescent Mental Health: Findings From the UK Millennium Cohort Study. *eClinicalMedicine, 6*, 59–68. https://doi.org/10.1016/j.eclinm.2018.12.005

Killeen, M. (2023, Juni 27). *Digitale Dienste: Zalando klagt gegen EU-Kommission*. www.euractiv.de. https://www.euractiv.de/section/finanzen-und-wirtschaft/news/digitale-dienste-zalando-klagt-gegen-eu-kommission/

Kim, W. G., Pillai, S. G., Haldorai, K., & Ahmad, W. (2021). Dark patterns used by online travel agency websites. *Annals of Tourism Research, 88*, 103055. https://doi.org/10.1016/j.annals.2020.103055

Kleemans, M., Daalmans, S., Carbaat, I., & Anschütz, D. (2018). Picture Perfect: The Direct Effect of Manipulated Instagram Photos on Body Image in Adolescent Girls. *Media Psychology, 21*(1), 93–110. https://doi.org/10.1080/15213269.2016.1257392

Kollmann, T. (2024). *Definition: Freemium*. Gabler Wirtschaftslexikon; Springer Fachmedien Wiesbaden GmbH. https://wirtschaftslexikon.gabler.de/definition/freemium-53522

Kreski, N., Platt, J., Rutherford, C., Olfson, M., Odgers, C., Schulenberg, J., & Keyes, K. M. (2021). Social Media Use and Depressive Symptoms Among United States Adolescents. *Journal of Adolescent Health, 68*(3), 572–579. https://doi.org/10.1016/j.jadohealth.2020.07.006

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2017). Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 417–432. https://doi.org/10.1145/2998181.2998321

Kuo, F.-Y. (2024). *Online Social Network Data-Driven Early Detection on Short-Form Video Addiction*. https://doi.org/10.48550/arXiv.2407.18277

Lamis, S. F., Handayani, P. W., & Fitriani, W. R. (2022). Impulse buying during flash sales in the online marketplace. *Cogent Business & Management, 9*(1). https://doi.org/10.1080/23311975.2022.2068402

Ledwich, M., & Zaitsev, A. (2020). Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *First Monday*. https://doi.org/10.5210/fm.v25i3.10419

Lee, J. K. (2022). The effects of social comparison orientation on psychological well-being in social networking sites: Serial mediation of perceived social support and self-esteem. *Current Psychology, 41*(9), 6247–6259. https://doi.org/10.1007/s12144-020-01114-3

Leerssen, P. (2023). An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. *Computer Law & Security Review, 48*. https://doi.org/10.1016/j.clsr.2023.105790

Leicht, K. T., Yun, J., Houston, B., Auvil, L., & Bracht, E. (2022). The Presentation of Self in Virtual Life: Disinformation Warnings and the Spread of Misinformation Regarding COVID-19. *RSF: The Russell Sage Foundation Journal of the Social Sciences, 8*(8), 52–68. https://doi.org/10.7758/RSF.2022.8.8.03

Ling, C., Gummadi, K., & Zannettou, S. (2024, September 8). *Learn the Facts About COVID-19": Analyzing the Use of Warning Labels on TikTok Videos*. https://www.researchgate.net/publication/357952765_Learn_the_Facts_About_COVID-19_Analyzing_the_Use_of_Warning_Labels_on_TikTok_Videos

Liu, C. Z., Au, Y. A., & Choi, H. S. (2014). Effects of Freemium Strategy in the Mobile App Market: An Empirical Study of Google Play. *Journal of Management Information Systems, 31*(3), 326–354. https://doi.org/10.1080/07421222.2014.995564

Liu, M., Kamper-DeMarco, K. E., Zhang, J., Xiao, J., Dong, D., & Xue, P. (2022). Time Spent on Social Media and Risk of Depression in Adolescents: A Dose–Response Meta-Analysis. *International Journal of Environmental Research and Public Health, 19*(9). https://doi.org/10.3390/ijerph19095164

Liu, X., Qi, L., Wang, L., & Metzger, M. J. (2023). Checking the Fact-Checkers: The Role of Source Type, Perceived Credibility, and Individual Differences in Fact-Checking Effectiveness. *Communication Research*. https://doi.org/10.1177/00936502231206419

Loi, M. (2023, August). *Making sense of the Digital Services Act How to define platforms' systemic risks to democracy*. AlgorithmWatch. https://algorithmwatch.org/en/wp-content/uploads/2023/08/AlgorithmWatch_Risk_Assessment-DSA.pdf

Lykousas, N., & Patsakis, C. (2021). Large-scale analysis of grooming in modern social networks. *Expert Systems with Applications, 176*. https://doi.org/10.1016/j.eswa.2021.114808

Mantelero, A. (2022). Fundamental rights impact assessments in the DSA. *Verfassungsblog*. https://doi.org/10.17176/20221101-220006-0

Marsh, O. (2024). Researching Systemic Risks under the Digital Services Act. *AlgorithmWatch*. https://algorithmwatch.org/en/wp-content/uploads/2024/08/AlgorithmWatch-Researching-Systemic-Risks-under-the-DSA-240726_v2.pdf

Martens, D., & Maalej, W. (2019). Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering, 24*(6), 3316–3355. https://doi.org/10.1007/s10664-019-09706-9

Martini, M., Drews, C., Seeliger, P., & Weinzierl, Q. (2021). Dark Patterns. Phänomenologie und Antworten der Rechtsordnung. *Zeitschrift für Digitalisierung und Recht*, 47–74.

McGlynn, C. (2021). *Written evidence (OSB0014)*. Durham University. https://committees.parliament.uk/writtenevidence/39012/pdf/

McGlynn, C., Rackley, E., & Houghton, R. (2017). Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse. *Feminist Legal Studies, 25*(1), 25–46. https://doi.org/10.1007/s10691-017-9343-2

Mchangama, J., Alkiviadou, N., & Abby, F. (2023, Juli 14). Scope Creep: An Assessment of 8 Social Media Platforms' Hate Speech Policies. The Future of Free Speech. https://futurefreespeech.org/scope-creep/

McNamee, P., Mendolia, S., & Yerokhin, O. (2021). Social media use and emotional and behavioural outcomes in adolescence: Evidence from British longitudinal data. *Economics & Human Biology, 41*. https://doi.org/10.1016/j.ehb.2021.100992

Merrill, R. A., Cao, C., & Primack, B. A. (2022). Associations between social media use, personality structure, and development of depression. *Journal of Affective Disorders Reports, 10*. https://doi.org/10.1016/j.jadr.2022.100385

Meßmer, A.-K., & Degeling, M. (2023, Februar). *Auditing Recommender Systems: Putting the DSA into practice with a risk-scenario-based approach*. Stiftung Neue Verantwortung. https://www.interface-eu.org/publications/auditing-recommender-systems

Metzler, H., & Garcia, D. (2024). Social Drivers and Algorithmic Mechanisms on Digital Media. *Perspectives on Psychological Science, 19*(5), 735–748. https://doi.org/10.1177/17456916231185057

Miljeteig, K., & von Soest, T. (2022). An Experience Sampling Study on the Association Between Social Media Use and Self-Esteem. *Journal of Media Psychology, 34*(6), 373–382. https://doi.org/10.1027/1864-1105/a000333

Miller, C., Smith, M., Marsh, O., Balint, K., Inskip, C., & Visser, F. (2022). *Information Warfare and Wikipedia*. Institute for Strategic Dialogue. https://www.isdglobal.org/wp-content/uploads/2022/10/Information-Warfare-and-Wikipedia.pdf

Mirea, D.-M., Mildner, J., Kelley, S., Gillan, C., Nook, E., & Niv, Y. (2024, Juni). *Depression is associated with higher sensitivity to social media rewards*. OSF. https://doi.org/10.31234/osf.io/4ynbc

Moraht, F. (2024). DSA-Riskoberichte: „Als wäre alles in Ordnung". *Tagesspiegel Background Digitalisierung & KI*. https://background.tagesspiegel.de/digitalisierung-und-ki/briefing/dsa-riskoberichte-als-waere-alles-in-ordnung

Müller, K. W., Dreier, M., Beutel, M. E., Duven, E., Giralt, S., & Wölfling, K. (2016). A hidden type of internet addiction? Intense and addictive use of social networking sites in adolescents. *Computers in Human Behavior, 55*, 172–177. https://doi.org/10.1016/j.chb.2015.09.007

Müller-Terpitz, R., Köhler, M., & Apel, S. (2024). *Digital Services Act: Gesetz über digitale Dienste: Kommentar*. C.H. Beck.

Murphy, H. (2024, März 11). Telegram hits 900mn users and nears profitability as founder considers IPO. *Financial Times*.

Na, S. H., Cho, S., & Shin, S. (2023). Evolving Bots: The New Generation of Comment Bots and their Underlying Scam Campaigns in YouTube. *Proceedings of the 2023 ACM on Internet Measurement Conference*, 297–312. https://doi.org/10.1145/3618257.3624822

Nouh, M., Nurse, J. R. C., & Goldsmith, M. (2019). *Understanding the Radical Mind: Identifying Signals to Detect Extremist Content on Twitter* (No. arXiv:1905.08067). https://doi.org/10.48550/arXiv.1905.08067

Ofcom. (2023, November 9). *Protecting people from illegal harms online*. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/270826-consultation-protecting-people-from-illegal-content-online/associated-documents/annex-5-draft-service-risk-assessment-guidance/?v=330403

Office of the High Commissioner for Human Rights. (2021). *Human Rights Due Diligence: An Interpretive Guide*. United Nations Development Programme. https://www.undp.org/sites/g/files/zskgke326/files/2022-10/HRDD%20Interpretive%20Guide_ENG_Sep%202021.pdf

O'Malley, R. L., & Holt, K. M. (2022). Cyber Sextortion: An Exploratory Analysis of Different Perpetrators Engaging in a Similar Crime. *Journal of Interpersonal Violence, 37*(1–2), 258–283. https://doi.org/10.1177/0886260520909186

Ozimek, P., & Bierhoff, H.-W. (2020). All my online-friends are better than me – three studies about ability-based comparative social media use, self-esteem, and depressive tendencies. *Behaviour & Information Technology, 39*(10), 1110–1123. https://doi.org/10.1080/0144929X.2019.1642385

Panahi, T., Jansen, C., Ancina, A., Bader, K., Choi, J.-E., Hornung, G., Krämer, N., Rinsdorf, L., Schäfer, K., Vogel, I., Yannikos, Y., & Steinebach, M. (2024). *Desinformation in Messenger-Diensten: Aktuelle Herausforderungen & Handlungsempfehlungen für rechtliche und gesellschaftliche Maßnahmen*. Nationales Forschungszentrum für angewandte Cybersicherheit ATHENE. https://duepublico2.uni-due.de/receive/duepublico_mods_00082406

People vs Big Tech. (2024, Dezember 4). *Big Tech companies say they consult with external stakeholders to assess & mitigate risks, but it seems none of the +120 orgs in @PeoplevsBigTech incl those who published research on systemic risks re: Social media platforms were consulted. The DSA says they should. Coincidence?* X.

Pielemeier, J., Jahangir, R., & Ross, H. (2024, Februar 19). *Ensuring Digital Services Act Audits Deliver on Their Promise | TechPolicy.Press*. Tech Policy Press. https://techpolicy.press/ensuring-digital-services-act-audits-deliver-on-their-promise

Primack, B. A., Shensa, A., Escobar-Viera, C. G., Barrett, E. L., Sidani, J. E., Colditz, J. B., & James, A. E. (2017). Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among U.S. young adults. *Computers in Human Behavior, 69*, 1–9. https://doi.org/10.1016/j.chb.2016.11.013

Primack, B. A., Shensa, A., Sidani, J. E., Whaite, E. O., Lin, L. yi, Rosen, D., Colditz, J. B., Radovic, A., & Miller, E. (2017). Social Media Use and Perceived Social Isolation Among Young Adults in the U.S. *American Journal of Preventive Medicine, 53*(1), 1–8. https://doi.org/10.1016/j.amepre.2017.01.010

Rauchfleisch, A., & Kaiser, J. (2021). *Deplatforming the Far-right: An Analysis of YouTube and BitChute* (SSRN Scholarly Paper No. 3867818). Social Science Research Network. https://doi.org/10.2139/ssrn.3867818

Reed, A., Whittaker, J., Votta, F., & Looney, S. (2019). *Radical Filter Bubbles: Social Media Personalisation Algorithms and Extremist Content*. https://rusi.org/explore-our-research/publications/special-resources/radical-filter-bubbles-social-media-personalisation-algorithms-and-extremist-content

Ribeiro, M. H., Veselovsky, V., & West, R. (2023a). *The Amplification Paradox in Recommender Systems*. arXiv. https://doi.org/10.48550/ARXIV.2302.11225

Ribeiro, M. H., Veselovsky, V., & West, R. (2023b). *The Amplification Paradox in Recommender Systems*. https://doi.org/10.48550/arXiv.2302.11225

Rieder, B., & Skop, Y. (2021). The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data & Society, 8*(2). https://doi.org/10.1177/20539517211046181

Rosenbach, M., & Schult, C. (2024, Januar 26). Stimmungsmache in sozialen Medien: Baerbocks Digitaldetektive decken russische Desinformationskampagne auf. *Der Spiegel*. https://www.spiegel.de/politik/deutschland/desinformation-aus-russland-auswaertiges-amt-deckt-pro-russische-kampagne-auf-a-765bb30e-8f76-4606-b7ab-8fb9287a6948

Rozgonjuk, D., Pruunsild, P., Jürimäe, K., Schwarz, R.-J., & Aru, J. (2020). Instagram use frequency is associated with problematic smartphone use, but not with depression and anxiety symptom severity. *Mobile Media & Communication, 8*(3), 400–418. https://doi.org/10.1177/2050157920910190

Sajn, N. (2021, Mai). *Vulnerable consumers*. European Parliamentary Research Services. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/690619/EPRS_BRI(2021)690619_EN.pdf

Sampasa-Kanyinga, H., & Lewis, R. F. (2015). Frequent Use of Social Networking Sites Is Associated with Poor Psychological Functioning Among Children and Adolescents. *Cyberpsychology, Behavior, and Social Networking, 18*(7), 380–385. https://doi.org/10.1089/cyber.2015.0055

Samra, A., Warburton, W. A., & Collins, A. M. (2022). Social comparisons: A potential mechanism linking problematic social media use with depression. *Journal of Behavioral Addictions, 11*(2), 607–614. https://doi.org/10.1556/2006.2022.00023

Satici, S. A., Gocet Tekin, E., Deniz, M. E., & Satici, B. (2023). Doomscrolling Scale: Its Association with Personality Traits, Psychological Distress, Social Media Use, and Wellbeing. *Applied Research in Quality of Life, 18*(2), 833–847. https://doi.org/10.1007/s11482-022-10110-7

Schettino, G., Fabbricatore, R., & Caso, D. (2023). "To be yourself or your selfies, that is the question": The moderation role of gender, nationality, and privacy settings in the relationship between selfie-engagement and body shame. *Psychology of Popular Media, 12*(3), 268–278. https://doi.org/10.1037/ppm0000417

Schneider, G., & Toyka-Seid, C. (2024). *Jugendschutz*. Bundeszentrale für Politische Bildung. https://www.bpb.de/kurz-knapp/lexika/das-junge-politik-lexikon/320576/jugendschutz/

Scott, M. (2024, Dezember 11). *5 Things to Know about the Digital Services Act's First Risk Assessments and Audits*. Tech Policy Press. https://techpolicy.press/5-things-to-know-about-the-digital-services-acts-first-risk-assessments-and-audits

Shachaf, P., & Hara, N. (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science, 36*(3), 357–370. https://doi.org/10.1177/0165551510365390

Sharma, B., Lee, S. S., & Johnson, B. K. (2022). The dark at the end of the tunnel: Doomscrolling on social media newsfeeds. *Technology, Mind, and Behavior, 3*(1). https://doi.org/10.1037/tmb0000059

Smirnova, J. (2024, Oktober 29). *SDA-Dokumente: Einblicke in Russlands digitale Desinformationsstrategie*. CeMAS. https://cemas.io/blog/sda-dokumente-russlands-desinformationsstrategie/

Smirnova, J., & Winter, H. (2021, November 5). *Ein Virus des Misstrauens*. ISD. https://isdgermany.org/ein-virus-des-misstrauens/

Smith, D. H., Ehrett, C., & Warren, P. L. (2024). *Unsupervised detection of coordinated information operations in the wild*. https://doi.org/10.48550/arXiv.2401.06205

Smith, R., & Murphy, H. (2024, August 30). Telegram's financial future in doubt as chief faces criminal inquiry. *Financial Times*.

Spitzer, E. G., Crosby, E. S., & Witte, T. K. (2023). Looking through a filtered lens: Negative social comparison on social media and suicidal ideation among young adults. *Psychology of Popular Media, 12*(1), 69–76. https://doi.org/10.1037/ppm0000380

Steers, M.-L. N., Wickham, R. E., & Acitelli, L. K. (2014). Seeing Everyone Else's Highlight Reels: How Facebook Usage is Linked to Depressive Symptoms. *Journal of Social and Clinical Psychology, 33*(8), 701–731. https://doi.org/10.1521/jscp.2014.33.8.701

Sthapit, E., & Björk, P. (2019). Sources of distrust: Airbnb guests' perspectives. *Tourism Management Perspectives, 31*, 245–253. https://doi.org/10.1016/j.tmp.2019.05.009

Su, C., Zhou, H., Gong, L., Teng, B., Geng, F., & Hu, Y. (2021). Viewing personalized video clips recommended by TikTok activates default mode network and ventral tegmental area. *NeuroImage, 237*. https://doi.org/10.1016/j.neuroimage.2021.118136

Suleman, M., Soomro, T. R., Ghazal, T. M., & Alshurideh, M. (2021). Combating Against Potentially Harmful Mobile Apps. In A. E. Hassanien, A. Haqiq, P. J. Tonellato, L. Bellatreche, S. Goundar, A. T. Azar, E. Sabir, & D. Bouzidi (Hrsg.), *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2021)* (S. 154–173). Springer International Publishing. https://doi.org/10.1007/978-3-030-76346-6_15

Suma, S. N., Nataraja, P., & Sharma, M. K. (2021). Internet Addiction Predictor: Applying Machine Learning in Psychology. In N. N. Chiplunkar & T. Fukao (Hrsg.), *Advances in Artificial Intelligence and Data Engineering* (S. 471–481). Springer Nature. https://doi.org/10.1007/978-981-15-3514-7_36

Taş, S., Liebe, A., & Wiewiorra, L. (2023). Moderation von Inhalten auf Online-Plattformen. *Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste*. https://www.econstor.eu/bitstream/10419/280951/1/1876995785.pdf

Tasnim, S., Hossain, M. M., & Mazumder, H. (2020). Impact of Rumors and Misinformation on COVID-19 in Social Media. *Journal of Preventive Medicine and Public Health, 53*(3), 171–174. https://doi.org/10.3961/jpmph.20.094

TCAP. (2024). *TCAP Insights: Understanding Patterns of Terrorist Exploitation Online by Geographic Region*. Terrorist Content Analytics Platform. https://terrorismanalytics.org/research-news/TCAP-Insights-Terrorist-Exploitation-by-Geographic-Region

Telegram. (2024). *Nutzerleitlinie für das EU-Gesetz über digitale Dienste*. Telegram. https://telegram.org/tos/eu-dsa

Theodosiadou, O., Pantelidou, K., Bastas, N., Chatzakou, D., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2021). Change Point Detection in Terrorism-Related Online Content Using Deep Learning Derived Indicators. *Information, 12*(7), Article 7. https://doi.org/10.3390/info12070274

Thorburn, L. (2023, Oktober 25). *What's the Difference Between Search and Recommendation?* https://medium.com/understanding-recommenders/whats-the-difference-between-search-and-recommendation-c32937506a29

Trendell, S. (2024, Oktober). A Tsunami of "Nudifying" Apps Advertised on Meta Platforms. In *NCOSE*. https://endsexualexploitation.org/articles/a-tsunami-of-nudifying-apps-advertised-on-meta-platforms/

Tuck, H., Guhl, J., Smirnova, J., Gerster, L., & Marsh, O. (2023, September 20). *Erforschung des sich im Wandel begriffenen Online-Ökosystems: Telegram, Discord und Odysee*. ISD. https://www.isdglobal.org/isd-publications/erforschung-des-sich-im-wandel-begriffenen-online-okosystems-telegram-discord-und-odysee/

Vera-Gray, F., McGlynn, C., Kureshi, I., & Butterby, K. (2021). Sexual violence as a sexual script in mainstream online pornography. *The British Journal of Criminology, 61*(5), 1243–1260. https://doi.org/10.1093/bjc/azab035

Verbraucherzentrale. (2024, August). *Schnäppchen-App Temu: Aufpassen beim Online-Shoppen!* https://www.verbraucherzentrale.de/wissen/digitale-welt/onlinehandel/schnaeppchenapp-temu-aufpassen-beim-onlineshoppen-86905

Verduyn, P., Lee, D. S., Park, J., Shablack, H., Orvell, A., Bayer, J., Ybarra, O., Jonides, J., & Kross, E. (2015). Passive Facebook usage undermines affective well-being: Experimental and longitudinal evidence. *Journal of Experimental Psychology. General, 144*(2), 480–488. https://doi.org/10.1037/xge0000057

Vereinte Nationen. (2024). *Global Digital Compact- A comprehensive framework for global governance of digital technology and artificial intelligence*. Vereinte Nationen. https://www.un.org/global-digital-compact/sites/default/files/2024-09/Global%20Digital%20Compact%20-%20English_0.pdf

Wang, J.-L., Gaskin, J., Rost, D. H., & Gentile, D. A. (2018). The Reciprocal Relationship Between Passive Social Networking Site (SNS) Usage and Users' Subjective Well-Being. *Social Science Computer Review, 36*(5), 511–522. https://doi.org/10.1177/0894439317721981

Whittaker, J., Looney, S., Reed, A., & Votta, F. (2021). Recommender systems and the amplification of extremist content. *Internet Policy Review, 10*(2). https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content

Williams, E. M., & Carley, K. M. (2023). Search engine manipulation to spread pro-Kremlin propaganda. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-112

Woods, H. C., & Scott, H. (2016). #Sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *Journal of Adolescence, 51*(1), 41–49. https://doi.org/10.1016/j.adolescence.2016.05.008

World Economic Forum. (2023a, Mai 26). *Digital Safety Risk Assessment in Action: A Framework and Bank of Case Studies*. https://www.weforum.org/publications/digital-safety-risk-assessment-in-action-a-framework-and-bank-of-case-studies/

World Economic Forum. (2023b, August). *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*. World Economic Forum. https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf

World Economic Forum. (2024, Juni). *Making a Difference: How to Measure Digital Safety Effectively to Reduce Risks Online*. https://www3.weforum.org/docs/WEF_Making_a_Difference_2024.pdf

Xie, Z., Zhu, S., Li, Q., & Wang, W. (2016). You can promote, but you can't hide: Large-scale abused app detection in mobile app stores. *Proceedings of the 32nd Annual Conference on Computer Security Applications*, 374–385. https://doi.org/10.1145/2991079.2991099

Zannettou, S. (2021). *„I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter* (No. arXiv:2101.07183). arXiv. https://doi.org/10.48550/arXiv.2101.07183

Zhang, M. R., Lukoff, K., Rao, R., Baughan, A., & Hiniker, A. (2022). Monitoring Screen Time or Redesigning It?: Two Approaches to Supporting Intentional Social Media Use. *Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3517722

Anhang

Anhang 1: Weiterführende Literatur-Datenbank (siehe beigefügte Excel-Datei 1)

Anhang 2: Übersicht zu Risikokategorien im DSA (siehe beigefügte Excel Datei 2)

Anhang 3: Übersicht über Risikominderungsmaßnahmen in offiziellen EU-Dokumenten

Ziel	Minderungsmaßnahme	Spezifische Minderungsmaßnahme	Kategorie	Risiko (s. Anh. 1)	Risiko Dimension	Quelle
Sicherstellen, dass KI generierte Inhalte sowie andere Arten synthetischer und manipulierter Medien erkennbar und unterscheidbar sind; Nachweis der Herkunft und Authentizität von Inhalten	sichtbare Kennzeichen	Wasserzeichen, Metadatenkennzeichnungen, kryptografische Methoden, Logging Methoden, digitale Fingerabdrücke	Moderation	25, 28	Verhalten	Art. 35k) DSA; Leitlinien zur Minderung systemischer Risiken online für Wahlen Paragraph 3.3, 39a)
Sicherstellen, dass KI generierte Inhalte sowie andere Arten synthetischer und manipulierter Medien erkennbar und unterscheidbar sind; Nachweis der Herkunft und Authentizität von Inhalten	Bereitstellen von leicht zugänglichen Funktionen, die es Nutzer*innen des Services ermöglichen, KI generierte Inhalte als solche zu kennzeichnen		Design/ Nutzerfahrung	25, 28	Verhalten	Art. 35k) DSA
Nutzer*innen dabei unterstützen, die Vertrauenswürdigkeit von Informationsquellen zu bewerten	Bereitstellen von Werkzeugen und Informationen, um die Herkunft, Bearbeitungshistorie, Authentizität oder Genauigkeit digitaler Inhalte zu bewerten	Von unabhängigen Dritten entwickelte transparente Gütesiegel; Faktenprüfung: EU-weit und in allen EU-Sprachen verfügbar, bspw. durch die Stärkung der Zusammenarbeit mit lokalen Faktenprüfer*innen (insbesondere während Wahlperioden), Einsatz von Mechanismen, um die Wirkung von Faktenprüfung auf die Zielgruppen zu erhöhen, verfasst in leicht zugänglicher und leicht verständlicher Sprache	Moderation; Bildung/Information; Kooperation mit anderen Online-Plattformen/ Behörden/zivilgesellschaftlichen Organisationen	25, 33, 35	Verhalten	Leitlinien zur Minderung systemischer Risiken online für Wahlen, Paragraph 3.2.1; 27c) l; Paragraph 3.2.1, 27c) v.; Paragraph 3.2.1, 27c) vi
Erkennung und Management systemischer Risiken	Effektive interne Maßnahmen schaffen um den Missbrauch von Labels, Gütesiegeln oder anderen Maßnahmen, insbesondere zur Verifizierung von gelabelten Konten und Inhalten zu verhindern		Interne Prozesse	25, 33	Verhalten	Leitlinien zur Minderung systemischer Risiken online für Wahlen 3.2.1, 27 c) vii.
Erstellung falscher Inhalte verhindern, die das Nutzerverhalten stark beeinflussen könnten	Nutzer*innen vor potenziellen Fehlern in den von generativen KI-Systemen erstellten Inhalten warnen und ihnen vorschlagen, zuverlässigen Quellen zu konsultieren, um den Wahrheitsgehalt von Informationen zu überprüfen; Schutzmaßnahmen einführen	Soweit möglich, in den erzeugten Outputs die Quellen der als Eingangsdaten verwendeten Informationen angeben	Moderation; Bildung/Information; Algorithmische Systeme	25, 33	Verhalten	Leitlinien zur Minderung systemischer Risiken online für Wahlen, Paragraph 3.3 39c), 39g)

Ziel	Minderungsmaßnahme	Spezifische Minderungsmaßnahme	Kategorie	Risiko (s. Anh. 1)	Risiko Dimension	Quelle
Schutz der Integrität der Wahlprozesse	Sicherstellen, dass die von KI-Systemen generierten Informationen zu Wahlprozessen so weit wie möglich auf zuverlässigen Quellen im Wahlkontext beruhen	Verwenden von offiziellen Informationen von den zuständigen Wahlbehörden, sicherstellen, sodass alle Zitate oder Verweise des Systems auf externe Quellen korrekt sind und den zitierten Inhalt nicht falsch wiedergeben	Design/ Nutzerfahrung	25, 28	Verhalten	Leitlinien zur Minderung systemischer Risiken online für Wahlen 39b)
Nutzer*innen kontextuelle Informationen bereitstellen	Klare, sichtbare und nicht irreführende Kennzeichnung von offiziellen Konten, Konten, die von Mitgliedstaaten und Drittländern kontrolliert oder finanziert werden, Konten von Einrichtungen, die von Drittländern kontrolliert oder finanziert werden, sowie Konten, die zuverlässliche Informationen über den Wahlprozess bereitstellen; Verhinderung von Imitationen; Kennzeichnung von identifizierten Desinformationen und FIMI-Inhalten; Ergreifung von Sensibilisierungsmaßnahmen und Anpassung der Nutzungsoberfläche, um den Nutzer*innen des Dienstes Zugang zu mehr Informationen zu geben	Überprüfung der Konten von Wahlbehörden, einschließlich der Grundlage, auf der eine solche Überprüfung vorgenommen wird. Die Kriterien, die zu einer »offiziellen« Kennzeichnung eines Kontos führen, sollten leicht zugänglich gemacht und in einer leicht verständlichen Sprache angegeben werden, um zu verhindern, dass solche Angaben Konten, die sich als offizielle Konten ausgeben (bspw. als Wahlbehörde), Glaubwürdigkeit verleihen	Moderation; Bildung/Information	25, 28, 33	Verhalten	Art. 35i) DSA; Leitlinien zur Minderung systemischer Risiken online für Wahlen 3.2.1; 27c) I, 3.2.1; 27c) III, 3.2.1; 27c) IV
Kinderrechte schützen		Tools zur Altersüberprüfung und elterlichen Kontrolle, Tools, die Minderjährigen unterstützen, Missbrauch zu melden oder gegebenenfalls Unterstützung zu erhalten	Design/Nutzererfahrung; Jugendschutz	2, 20,21,22	Inhalt, Nutzung	Art. 35j) DSA
Bekämpfung der Verbreitung von rechtswidrigen Inhalten im Internet	rasche Entfernung der gemeldeten Inhalte oder Sperrung des Zugangs dazu	Zusammenarbeit mit vertrauenswürdigen Hinweisgebern (Organisation von Trainingssitzungen und Austausch mit vertrauenswürdigen Hinweisgebern); Überprüfung der Mehrheit der gültigen Meldungen in Bezug auf die Entfernung rechtswidriger Hassreden in weniger als 24 Stunden & gegebenenfalls Entfernung oder Deaktivierung des Zugangs	Moderation; Kooperation mit anderen Online-Plattformen, Behörden und zivilgesellschaftlichen Organisationen	1,4, 27, 31	Inhalt, Verhalten	Erwägungsgrund 87; Art. 35 1) c), Art. 35 1) g) DSA; Verhaltenskodex zur Bekämpfung von rechtswidriger Hassrede im Internet

Ziel	Minderungsmaßnahme	Spezifische Minderungsmaßnahme	Kategorie	Risiko (s. Anh. 1)	Risiko Dimension	Quelle
Bekämpfung der Verbreitung von rechtswidrigen Inhalten im Internet	Förderung der Bereitstellung von Meldungen und Kennzeichnung von Inhalten, die zu Gewalt und aggressivem Verhalten aufstacheln durch Sachverständige, insbesondere durch Partnerschaften mit Organisationen der Zivilgesellschaft, indem sie klar über einzelne Unternehmensregeln und Community-Leitlinien und Vorschriften über die Meldungs- und Benachrichtigungsverfahren informieren.		Geschäftsbedingungen	1, 4, 27, 31	Inhalt, Verhalten	Verhaltenskodex zur Bekämpfung von rechtswidriger Hassrede im Internet
Kinderrechte schützen	Anpassung der Gestaltung von Diensten und ihrer Online-Oberfläche, insbesondere wenn diese auf Minderjährige ausgerichtet sind oder überwiegend von diesen genutzt werden; Sicherstellung, dass Dienste so organisiert sind, dass Minderjährige leicht Zugang zu den im DSA vorgesehenen Mechanismen erhalten; Ergreifen von Maßnahmen, um Minderjährige vor Inhalten zu schützen, die ihre körperliche, geistige oder moralische Entwicklung beeinträchtigen könnten; Bereitstellen von Tools, die einen bedingten Zugang zu solchen Informationen ermöglichen, einschließlich solcher, die durch selbstregulatorische Zusammenarbeit entwickelt wurden	Melde- und Handlungsverfahren sowie Beschwerdemechanismen; Verhaltenskodizes	Moderation; Jugendschutz; Kooperation mit anderen Online-Plattformen, Behörden und zivilgesellschaftlichen Organisationen	20, 21, 22, 23	Nutzung	Erwägungsgrund 89 DSA
Verbesserung der Wahlbeteiligung und Verhinderung der Verbreitung von Fehlinformation, Desinformation und ausländischer Informationsmanipulation und -intervention (»FIMI«) in Bezug auf den Wahlprozess	Erleichterten Zugang zu offiziellen Wahlinformationen über Online-Plattformen schaffen	Bereitstellen von Informationen, wie und wo gewählt werden kann, basierend auf offiziellen Angaben der Wahlbehörden der Mitgliedstaaten in Form von Bannern, Pop-Ups oder Links	Bildung/Information; Kooperation mit anderen Online-Plattformen, Behörden und zivilgesellschaftlichen Organisationen	25	Nutzung	Leitlinien zur Minderung systemischer Risiken online für Wahlen 27a)
Bereitstellung zusätzlicher kontextbezogener Informationen für Nutzer*innen über die Inhalte und Accounts, mit denen sie interagieren	Aufforderungen und Hinweise, die Nutzer*innen dazu anregen, Inhalte zu lesen und deren Genauigkeit sowie die Quelle zu bewerten, bevor sie diese teilen		Bildung/Information	25, 28	Nutzung	Leitlinien zur Minderung systemischer Risiken online für Wahlen 3.2.1; 27c) II

Ziel	Minderungsmaßnahme	Spezifische Minderungsmaßnahme	Kategorie	Risiko (s. Anh. 1)	Risiko Dimension	Quelle
Transparente und faire algorithmische Systeme	Regelmäßige Bewertung der Leistung und Auswirkungen von Empfehlungssystemen und bei Bedarf Anpassung algorithmischer Systeme (einschließlich Empfehlungssysteme); Entwicklung von Systemen, die Medienpluralismus und inhaltliche Vielfalt fördern	Schaffung von Transparenz über die Gestaltung und Funktionsweise von Empfehlungssystemen, insbesondere in Bezug auf die Daten und Informationen, die bei der Entwicklung der Systeme verwendet werden, um eine Prüfung und Forschung durch Dritte zu ermöglichen	Algorithmische Systeme	7, 8, 10, 12, 14, 16, 18, 22, 23, 24, 26, 29, 30, 32, 34, 36, 37	Verhalten, Nutzung	Art. 35 d) DSA; Leitlinien zur Minderung systemischer Risiken online für Wahlen 27d) IV., V.; Erwägungsgrund 88, 94 DSA
Transparente und faire algorithmische Systeme	Sicherstellen, dass algorithmische Systeme so gestaltet und angepasst sind, dass Nutzer*innen sinnvolle Wahlmöglichkeiten und Kontrolle über ihre Feeds haben.	Sicherstellen, dass Nutzer*innen alternative Optionen zur Verfügung stehen, die nicht auf Profiling basieren. Diese sollten direkt über die Benutzeroberfläche zugänglich sein, auf der die Empfehlungen präsentiert werden.	algorithmische Systeme; Design/ Nutzererfahrung	7, 8, 10, 12, 14, 16, 18, 22, 23, 24, 26, 29, 30, 32, 34, 36, 37	Verhalten, Nutzung	Leitlinien zur Minderung systemischer Risiken online für Wahlen 27d) I
Reduzierung der Sichtbarkeit von Desinformation	Einschränkung der Verbreitung von irreführenden, falschen oder täuschenden Inhalten, die von KI generiert wurden, als falsch überprüft wurden oder von Konten stammen, die wiederholt durch die Verbreitung von Desinformation auffällig geworden sind		algorithmische Systeme	25, 26, 28, 29, 33, 38	Verhalten, Nutzung	Leitlinien zur Minderung systemischer Risiken online für Wahlen 27d) II
Transparente und faire algorithmische Systeme	Zusammenarbeit mit externen Parteien zur Durchführung von Adversarial Testing und Red-Team-Tests		Kooperation mit anderen Online-Plattformen, Behörden und zivilgesellschaftlichen Organisationen	8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 21, 22, 24, 29, 30	Nutzung, Verhalten	Leitlinien zur Minderung systemischer Risiken online für Wahlen 39 d), 27 d) vi., 53
Bekämpfung der Verbreitung von rechtswidrigen Inhalten im Internet	Regeln oder Community-Leitlinien, in denen klargestellt wird, dass die Aufstachelung zu Gewalt und aggressivem Verhalten verboten ist.		Geschäftsbedingungen	1, 2, 4	Inhalt	Verhaltenskodex zur Bekämpfung von rechtswidriger Hassrede im Internet
Reduzierung der Sichtbarkeit von Desinformation und ausländischer Informationsmanipulation und -intervention (FIMI)	Durch gezielte Richtlinien und Systeme sicherstellen, dass die Platzierung von Werbung keine finanziellen Anreize für die Verbreitung von Desinformation (einschließlich generierter KI-Inhalte) und FIMI im Zusammenhang mit Wahlprozessen sowie von hasserfüllten, (gewalttätigen) extremistischen oder radikalisierenden Inhalten bietet, die die Wahlentscheidungen von Personen beeinflussen können		algorithmische Systeme; Geschäftsbedingungen	25, 26, 27, 28	Verhalten	Leitlinien zur Minderung systemischer Risiken online für Wahlen 27 g), 27 e), IV.

Ziel	Minderungsmaßnahme	Spezifische Minderungsmaßnahme	Kategorie	Risiko (s. Anh. 1)	Risiko Dimension	Quelle
Reduzierung der Sichtbarkeit von Desinformation und ausländischer Informationsmanipulation und -intervention (»FIMI«)	Sicherstellen, dass Manipulationen des Dienstes rechtzeitig und wirksam erkannt und unterbunden werden, wenn diese als relevantes systemisches Risiko identifiziert wurden, unter Berücksichtigung der bestmöglichen verfügbaren Beweise	Aufstellen von spezifischen Regeln gegen die Erstellung von unauthentischen Konten oder Bot-Netzen (einschließlich automatisierter, teilweise automatisierter oder nicht automatisierter Konten) sowie gegen die irreführende Nutzung eines Dienstes	interne Prozesse; Geschäftsbedingungen	25, 26, 27, 28, 29, 33	Verhalten	Leitlinien zur Minderung systemischer Risiken online für Wahlen 27 h)
Schutz der Integrität von Wahlprozessen	Förderung eines zeitnahen Informationsaustauschs bei Risiken, die den Wahlprozess betreffen; Verbesserung der Geschwindigkeit und Effektivität der Kommunikation zwischen den Behörden der Mitgliedstaaten und IT-Unternehmen, insbesondere bei Benachrichtigungen sowie beim Sperren oder Entfernen rechtswidriger Hassrede im Internet	Organisation von Treffen vor den Wahlen und Einrichtung regelmäßiger Kommunikationskanäle mit relevanten Akteur*innen; Aufbau von Kontaktstellen zu relevanten Behörden (auf regionaler, nationaler und EU-Ebene)	Kooperation mit anderen Online-Plattformen, Behörden und zivilgesellschaftlichen Organisationen; Moderation	25, 27, 31	Verhalten	Leitlinien zur Minderung systemischer Risiken online für Wahlen 3.4
Bekämpfung der Verbreitung von rechtswidrigen Inhalten im Internet	Unterstützung von zivilgesellschaftlichen Organisationen bei der Durchführung wirksamer Gegenkampagnen	Bereitstellen von Best-Practice-Schulungen zur Bekämpfung von hasserfüllter Rhetorik und Vorurteilen sowie Ausweitung proaktiver Maßnahmen zur Kontaktaufnahme mit zivilgesellschaftlichen Organisationen; Austausch über bewährte Verfahren	Kooperation mit anderen Online-Plattformen, Behörden und zivilgesellschaftlichen Organisationen; Bildung/ Information	1, 31	Inhalt, Verhalten	Verhaltenskodex zur Bekämpfung von rechtswidriger Hassrede im Internet
Bekämpfung der Verbreitung von rechtswidrigen Inhalten im Internet	Anbieten von regelmäßigen Schulungen für das Plattformpersonal zu aktuellen gesellschaftlichen Entwicklungen und Förderung des Austauschs von Ideen zur weiteren Verbesserung der Dienste		Interne Prozesse	1,31	Inhalt, Verhalten	Verhaltenskodex zur Bekämpfung von rechtswidriger Hassrede im Internet

Ziel	Minderungsmaßnahme	Spezifische Minderungsmaßnahme	Kategorie	Risiko (s. Anh. 1)	Risiko Dimension	Quelle
Sicherstellung von Transparenz bei zielgerichteter (politischer) Werbung		Verbesserung der Sichtbarkeit vertrauenswürdiger Quellen; strukturelle Anpassung von Werbesystemen, um die Verbreitung risikobehafteter Inhalte zu verhindern; Sicherstellung des öffentlichen Zugangs zu Werbe-Datenbanken; Bereitstellung detaillierter Informationen zu Werbeanzeigen, einschließlich Daten zum Werbetreibenden und Sponsoren; klare, auffällige und eindeutige Kennzeichnung; Bereitstellung von Informationen für Nutzer*innen über: die Identität des Sponsors, gegebenenfalls die letztlich den Sponsor kontrollierende Stelle; den Zeitraum, in dem die politische Werbung veröffentlicht werden soll; die aggregierten Beträge und den Gesamtwert anderer Vorteile, die die Anbieter*innen von Dienstleistungen im Bereich politischer Werbung erhalten haben; sowie aussagekräftige Informationen über die Hauptparameter, die zur Bestimmung der Zielgruppe verwendet wurden, an die die Werbung gerichtet ist	Moderation; Bildung/Information; algorithmische Systeme	18,19, 25, 28	Nutzung, Verhalten	Erwägungsgrund 88 DSA; Leitlinien zur Minderung systemischer Risiken online für Wahlen 27e)
Sicherstellung von Transparenz bei zielgerichteter (politischer) Werbung		Bereitstellung einer Funktion, die Influencer*innen ermöglicht, anzugeben, ob der von ihnen bereitgestellte Inhalt politische Werbung enthält, einschließlich der Identität des Sponsors und, falls zutreffend, die den Sponsoren kontrollierende Entität	Bildung/Information; Design/Nutzererfahrung	18,19,25,28	Nutzung, Verhalten	Leitlinien zur Minderung systemischer Risiken online für Wahlen 27f)
Aufbau von Resilienz gegenüber möglichen und erwarteten Desinformationsnarrativen und Manipulationstechniken	In-App-Initiativen zur Medienkompetenz	Gamifizierte Interventionen und Videos, unter Berücksichtigung lokaler Kontexte und ergänzt durch verlässliche Informationen, um Nutzer*innen über Inhalte aufzuklären, die gemäß den Regeln und Gemeinschaftsrichtlinien nicht erlaubt sind; Entwicklung neuer Ideen und Initiativen sowie Unterstützung von Bildungsprogrammen, die kritisches Denken fördern	Bildung/Information	1, 2, 3, 4, 5, 6, 7, 25, 27, 30, 31	Inhalt, Verhalten	Leitlinien zur Minderung systemischer Risiken online für Wahlen 27 b)

Ziel	Minderungsmaßnahme	Spezifische Minderungsmaßnahme	Kategorie	Risiko (s. Anh. 1)	Risiko Dimension	Quelle
Schutz der Integrität von Wahlprozessen	Interne Vorfallsreaktionsplanung für Notfallsituationen	Einrichtung, Abstimmung und Testen eines Notfallplans, u.a. durch Red-Teaming-Übungen; Einbezug der obersten Führungsebene sowie eine Kartierung der innerhalb der Organisation am Notfallplan beteiligten Stakeholder	Interne Prozesse	25, 26	Verhalten	Leitlinien zur Minderung systemischer Risiken online für Wahlen 53

Minderungsmaßnahmen, die auf alle Risiken anwendbar sind

Ziel	Minderungsmaßnahme	Spezifische Minderungsmaßnahme	Kategorie	Risiko (a. Anh. 1)	Risiko Dimension	Quelle
Anpassung interner Ressourcen zur Minderung systemischer Risiken	Anpassung relevanter Entscheidungsprozesse und bereichsspezifischer Ressourcen (einschließlich Content-Moderation-Personal, Schulungen und lokaler Expertise)		Moderation	alle		Erwägungsgrund 87 DSA
Zusammenarbeit bei der Minderung systemischer Risiken	Teilen von Informationen über die Risikobewertung und Maßnahmen zur Risikominderung während Wahlprozessen; Zusammenarbeit mit nationalen Behörden und nichtstaatlichen Akteur*innen (Wissenschaft, unabhängige Expert*innen, zivilgesellschaftliche Organisationen, Gemeinschaftsvertreter, unabhängige Faktenprüfer*innen).	Zusammenarbeit mit dem Digital Services Coordinator des jeweiligen Mitgliedstaates; Engagement in nationalen Wahlnetzwerken und Förderung der Kooperation mit relevanten Akteur*innen; Zusammenarbeit mit Faktenprüfer*innen, die hohe Standards in Methodik, Ethik und Transparenz einhalten, wie beispielsweise die des European Fact-Checking Standards Network (EFCSN) gemäß dessen Verhaltenskodex; während der Wahlperiode; Entwicklung eines verfahrensrechtlichen Rahmen für die Kooperation und Koordination bei Wahlen unter Gewährleistung eines raschen Feedbacks und angemessener Folgemaßnahmen durch die Mechanismen der Plattform für die Reaktion auf Zwischenfälle im Rahmen des Schnellreaktionssystems; Sammlung von öffentlichem Feedback, um bestehende Maßnahmen zur Risikominderung zu verbessern oder erfolgreiche Maßnahmen mit anderen Anbieter*innen zu teilen.	Kooperation mit anderen Online-Plattformen, Behörden und zivilgesellschaftlichen Organisationen; Interne Prozesse	alle		Leitlinien zur Minderung systemischer Risiken online für Wahlen 3.4
Zusammenarbeit bei der Minderung systemischer Risiken	Zusammenarbeit mit anderen Anbieter*innen (z.B. Beitritt zu oder Initiierung von Verhaltenskodizes); Austausch über bewährte Verfahren		Kooperation mit anderen Online-Plattformen, Behörden und zivilgesellschaftlichen Organisationen	alle		Erwägungsgrund 88, 89; Art. 35 h) DSA
Erkennung und Management systemischer Risiken	Verstärkung der internen Prozesse, Ressourcen, Tests, Dokumentation oder Überwachung der Tätigkeiten		Interne Prozesse	alle		Art. 35 f) DSA

Ziel	Minderungsmaßnahme	Spezifische Minderungsmaßnahme	Kategorie	Risiko (a. Anh. 1)	Risiko Dimension	Quelle
Stetige Verbesserung der Wirksamkeit der Minderungsmaßnahmen	Anpassung der ergriffenen Maßnahmen basierend auf internen Metriken zur Erfassung der Wirksamkeit	Metriken: Angaben zur durchschnittlichen Reaktionszeit auf Verstöße gegen die Nutzungsbedingungen, gekennzeichnete Inhalte durch Nutzer*innen und nichtstaatliche Akteur*innen, und Reichweite, sowie zum Engagement von Inhalten, zur Anzahl der Verstöße gegen bestimmte Wahlgrundsätze, zu Fällen von Informationsmanipulation und zur Reichweite bestimmter Maßnahmen wie Initiativen zur Förderung der Medienkompetenz und behördlicher Initiativen.	Interne Prozesse	alle		Leitlinien zur Minderung systemischer Risiken online für Wahlen 61

ISD | Institute for Strategic Dialogue

Amman | Berlin | London | Paris | Washington DC

www.isdgermany.org

Im Auftrag von:

Bundesnetzagentur