Digitalisierung & KI

# Wie sieht effektive Risikobewertung aus?

von Friederike Moraht, veröffentlicht am 14.08.2023

**Mit dem Anwendungsbeginn des Digital Services Act am 25. August müssen Twitter, Youtube und Co. der EU-Kommission ihre ersten Risikobewertungen vorlegen. Das Konzept ist in der Branche neu und es gab nur wenig Vorgaben. Zivilgesellschaft und Forschung befürchten Audit-Washing – und pochen auf mehr Beteiligung bei der Ausarbeitung zukünftiger Leitlinien.**

Am **25. August** ist es so weit: Der **Digital Services Act** (DSA), das erste umfangreiche europäische Regelwerk für digitale Plattformen und Suchmaschinen, gilt für Facebook, Twitter (X), Youtube und Co.
Zwar ist mit Anwendungsbeginn nicht vom einen auf den anderen Tag ein besseres Internet zu erwarten, doch zumindest mit einer Sache geht es direkt los: Den Risikobewertungen, die sehr große Online-Plattformen (VLOPs) und sehr große Online-Suchmaschinen (VLOSEs) dann zum ersten Mal vorlegen müssen.

Das Konzept der **Risikobewertung- und minderung** nach Artikel 34 und 35 gilt als das **Herzstück des DSA**. Demnach müssen VLOPs und VLOSEs mindestens einmal jährlich und vor der Einführung neuer Funktionen „systemische Risiken“, die sich aus ihren Diensten und Systemen ergeben, analysieren – und anschließend Maßnahmen ergreifen, um die ermittelten Risiken zu mindern.

Mindestens **vier Monate Zeit** werden die Dienstanbieter dafür gehabt haben, seitdem sie im April als VLOP/ VLOSE designiert wurden. Mit einer freiwilligen Veröffentlichung der Berichte Ende August ist allerdings nicht zu rechnen. Nach Abgabe der Risikobewertungen an die EU-Kommission müssen die Unternehmen sich auditieren lassen. Erst mit der Fertigstellung der Audits, die veröffentlicht werden müssen, erwarten Akteure aus Zivilgesellschaft und Forschung, **mittelbar Einblicke in die Risikobewertungen** zu bekommen.
Erwartet wird, dass die ersten Risikobewertungen sehr unterschiedlich ausfallen. „Das **Konzept des systemischen Risikos** wurde hauptsächlich im **Finanzwesen** entwickelt. In diesem Bereich ist es neu“, sagt **Andrea Calef**, der mit **Sally Broughton Micova** für das Centre on Regulation in Europe (Cerre) Elemente für eine effektive Risikobewertung analysiert hat, gegenüber Tagesspiegel Background. „Da es sich um die erste Runde der Risikobewertung

im Rahmen der DSA handelt, gibt es zudem **keinen vorherigen Benchmark**." Die Plattformen werden ihre eigenen Definitionen und Maßstäbe erstellen, da auch der DSA hier nur wenige Instruktionen gebe. (https://cerre.eu/news/cerre-on-dsa-systemic-risk-assessment/)

Befürchtet wird, dass diese Lücke die Ausbreitung von **„Audit-Washing"** begünstigen könnte. **Forschung und Zivilgesellschaft** üben deshalb schon seit längerem Druck auf die Plattformen aus und unterbreiten ihnen **Vorschläge für die Herangehensweise bei Risikobewertungen**. Es sei wichtig und notwendig, unabhängige Experten und Interessengruppen in die Gestaltung, Umsetzung und Überprüfung von Risikobewertungsmethoden und -prozessen einzubeziehen, heißt es einhellig. In einer kürzlich veröffentlichten Pressemitteilung von Algorithmwatch moniert man: „Der derzeitige **Mangel an Transparenz** und an der Einbeziehung der Zivilgesellschaftin diesen Prozess ist jedoch alarmierend". Auch die NGO hat vergangene Woche einen ersten Beitrag (https://algorithmwatch.org/en/wpcontent/uploads/2023/08/AlgorithmWatch_Risk_Assessment-DSA.pdf) dazu veröffentlicht, wie Risikobewertungen in der Praxis durchgeführt werden könnten.

**Wie sieht systemisches Versagen von Diskurs aus?**

Im Zentrum steht die Frage, was **systemische Risiken** sind und wie sie gemessen werden. Im DSA ist zwar grob definiert, um welche Risiken es sich handelt – darunter die Verbreitung rechtswidriger Inhalte, nachteilige Auswirkungen auf Grundrechte, die gesellschaftliche Debatte, Wahlprozesse und die öffentliche Sicherheit sowie geschlechtsspezifische Gewalt – an welchem Punkt ein Risiko „systemisch" wird, sei laut Calef und Broughton Micova im Gegensatz zum Finanzwesen, wo man ein systemisches Risiko mit Formeln und Berechnungen bewerten kann, **viel komplizierter zu bestimmen**. „Wenn es um terroristische Inhalte oder Material zur sexuellen Ausbeutung von Kindern geht, ist der Maßstab im Idealfall Null. Aber wir wissen nicht, wie ein **systemisches Versagen des zivilgesellschaftlichen Diskurses** aussieht", so Broughton Micova. Die Risiken seien zudem **viel subtiler**. Im Laufe der Zeit könne es zu einer Häufung von Schocks kommen, die plötzlich eine gewisse kritische Masse erreichen, die dann systemisch wird.

Zwar fängt man auch hier nicht bei null an – die Plattformen etwa orientieren sich angeblich stark an den UN Guiding Principles on Business and Human Rights – doch braucht es laut Broughton und Calef zusätzlich einen Multi-Stakeholder-Ansatz, um bestehende Ansätze zusammenzubringen und das

normative Gleichgewicht zu ermitteln. Hier sollte die **EU-Kommission eine zentrale Vermittlerrolle einnehmen**, finden die Autor:innen.

**Netzwerkanalysen und Risikoszenarien**

Erste Ideen, wie systemische Risiken dieser Art empirisch quantifiziert werden könnten, liefert **Michele Loi** im neuen Bericht von Algorithmwatch. Seine Methodik verweist auf spezifische **Konzepte der Wahrscheinlichkeit**. Ihm nach sei es sogar möglich, ein Risikoverständnis für die Demokratie zu entwickeln, das „durch relativ  einfache Beobachtungen empirisch quantifiziert werden kann".

„Die **wahre Herausforderung liege jedoch in der Lösung der normativen Frage**, welche Beobachtungen von Bedeutung sind und warum", so Loi.

Bezüglich der Messung von Risiken warnen Expert:innen der **Action Coalition for Meaningful Transparency** davor, dass dieRisikobewertungen von den Plattformen **zu allgemein** gehalten werden. Ein allgemeines Risiko für das Recht auf freie Meinungsäußerung zubenennen, sei beispielsweise unzureichend, heißt es ihrem Policy Paper. Dienstanbieter sollten sich auch mit den **spezifischen Nuancen ihrer eigenen Produkte und Dienste** auseinandersetzen, zum Beispiel, wie sich eine Politik zur Monetarisierung für Content Creator auf das Vorherrschen von Desinformation in einem bestimmten Produkt auswirken könnte. (https://www.meaningfultransparency.tech/post/dsa-risk-assessment)

Für **Spezifität in den Risikobewertungen** spricht sich auch **Anna Semenova** von der Stiftung Neue Verantwortung (SNV) aus. In der datenbasierten Studie über Youtube kam heraus, dass bestimmte problematische Dynamiken erst bei der isolierten Betrachtung einzelner Komponenten wie der ‚Nächstes Video'-Funktion sichtbar werden. So war eine Untergruppe nur bei dieser Funktion Inhalten wie beispielsweise Esoterik und Impfskeptizismus besonders häufig ausgesetzt. Semenova kritisiert, dass die Plattformen meistens **aggregierte Zahlen veröffentlichen, die nicht aussagekräftig** sind. „Meine Studie zeigt, wie wichtig es ist, genauer hinzusehen, wie zum Beispiel Inhaltsaufrufe verteilt sind – sowohl in Bezug auf Personengruppen als auch Plattformkomponenten." (https://www.stiftung-nv.de/de/publication/thetreeof-complexity)

Die Autor:innen des Cerre plädieren für **Netzwerkanalysen**. „Die Plattformen müssen sehen, wie Funktionalitäten miteinander zusammenhängen, mit wem sie enge Beziehungen unterhalten und welche Risiken sich aus diesen Beziehungen ergeben könnten", so Broughton Micova. Als Beispiel nennt sie die Bedingungen von **Multi-Channel-Networks mit ihren unter Vertrag**

**stehenden Influencern** und den daraus resultierenden Werbeeinnahmen. Gibt es Standards in Bezug auf die Ansprache von Kindern und in Bezug auf die Verwendung von Sprache? Eine Abhilfemaßnahme liege möglicherweise nicht in der Inhaltsmoderation auf der Plattform, sondern in einem **Code of Conduct für Multi-Channel-Networks** und deren Kunden.
Anna-Katharina Meßmer und Martin Degeling (SNV) haben sich insbesondere mit dem Auditieren und Bewerten der Empfehlungssysteme auseinandergesetzt. Sie schlagen vor , **mit Szenarien zu arbeiten**: Abstrakte Risiken wie Hatespeech sollen „**inkonkrete, überprüfbare Risiko-Szenarien** umgewandelt [werden], indem die betroffene Partei und ihre Merkmale, der Schaden, die beteiligten Elemente der Plattform und die weiteren Auswirkungendefiniert werden.“(https://www.stiftung-nv.de/en/publication/auditing-recommendersystemsoverview-existing-audits-risk-assessments-and-studies)

„Unser Ansatz war der Versuch, den Plattformen ein Schema anzubieten. Die Rückmeldung von Plattformseite war, dass es zu viel erwartet sei, für jedes Risiko, jeden Schaden und jedes Produkt einzelner Plattformen ein eigenes Risiko-Szenario zu entwickeln und zu testen“, sagt Meßmer gegenüber Tagesspiegel Background. Dafür hat sie nur wenig Verständnis. „Die **Facebook-Papers von Frances Haugen** haben gezeigt, dass intern ganz viele Tests laufen. Durch den DSA ändert sich, dass es systematischer sein und bestimmten Vorgaben folgen muss.“ Sie verstehen den Aufwand, den die neuen Ansätze des Artikel 34 für sehr große Plattformen nach sich ziehen, aber habe angesichts von deren Milliardeneinnahmen „wenig Mitleid“ bei konkreten Umsetzungsfragen.

Digitalisation & AI

# What does effective risk assessment look like?

by Friederike Moraht, published on 14.08.2023

**With the start of application of the Digital Services Act on 25 August, Twitter, Youtube and Co. have to submit their first risk assessments to the EU Commission. The concept is new in the industry and there was little guidance. Civil society and research fear audit-washing - and insist on more participation in the development of future guidelines.**

On 25 August, the time has come: the Digital Services Act (DSA), the first comprehensive European regulatory framework for digital platforms and search engines, will apply to Facebook, Twitter (X), Youtube and Co.

Although a better internet cannot be expected from one day to the next with the start of application, at least one thing will start right away: the risk assessments that very large online platforms (VLOPs) and very large online search engines (VLOSEs) will have to submit for the first time.

The concept of risk assessment and mitigation under Articles 34 and 35 is considered the core of the DSA. It requires VLOPs and VLOSEs to analyse "systemic risks" arising from their services and systems at least once a year and before introducing new features - and then take action to mitigate the identified risks.

Service providers will have had at least four months to do this since they were designated as VLOP/VLOSE in April. However, a voluntary publication of the reports at the end of August is not to be expected. After submitting the risk assessments to the EU Commission, the companies must have themselves audited. Only with the completion of the audits, which have to be published, do actors from civil society and research expect to gain indirect insights into the risk assessments.

It is expected that the initial risk assessments will be very different. "The concept of systemic risk was developed mainly in finance. It is new in this field," says Andrea Calef, who with Sally Broughton Micova has analysed elements for

effective risk assessment for the Centre on Regulation in Europe (Cerre), to Tagesspiegel Background. "Moreover, as this is the first round of risk assessment under the DSA, there is no previous benchmark." Platforms will create their own definitions and benchmarks, he said, as the DSA also gives few instructions in this regard. (https://cerre.eu/news/cerre-on-dsa-systemic-risk-assessment/)

It is feared that this gap could encourage the spread of audit-washing. Research and civil society have therefore been exerting pressure on the platforms for a long time, making suggestions to them on how to approach risk assessments. It is important and necessary to involve independent experts and stakeholders in the design, implementation and review of risk assessment methods and processes, they say unanimously. In a recent press release from Algorithmwatch, they criticise: "However, the current lack of transparency and civil society involvement in this process is alarming". Last week, the NGO also published a first contribution (https://algorithmwatch.org/en/wpcontent/uploads/2023/08/AlgorithmWatch_Risk_Assessment-DSA.pdf) on how risk assessments could be carried out in practice.

What does systemic failure of discourse look like?

At the heart of this is the question of what systemic risks are and how they are measured. While the DSA roughly defines what risks are - including the dissemination of unlawful content, adverse impacts on fundamental rights, social debate, electoral processes and public safety, and gender-based violence - at what point a risk becomes "systemic" is much more complicated to determine, according to Calef and Broughton Micova, in contrast to finance, where one can assess a systemic risk with formulas and calculations.

"When it comes to terrorist content or child sexual exploitation material, the standard is ideally zero. But we don't know what a systemic failure of civil society discourse looks like," says Broughton Micova. The risks are also much more subtle, she added. Over time, he said, there can be an accumulation of shocks that suddenly reach a certain critical mass, which then becomes systemic.

While not starting from scratch - the platforms, for example, are said to be strongly guided by the UN Guiding Principles on Business and Human Rights -

Broughton and Calef say that an additional multi-stakeholder approach is needed to bring together existing approaches and identify the normative balance. The authors believe that the EU Commission should play a central role as a facilitator.

## Network analyses and risk scenarios

The first ideas on how systemic risks of this kind could be empirically quantified are provided by Michele Loi in the new report by Algorithmwatch. His methodology refers to specific concepts of probability. According to him, it is even possible to develop an understanding of risk for democracy that "can be empirically quantified through relatively simple observations".

"The real challenge, however, lies in resolving the normative question of which observations matter and why," Loi says.

Regarding the measurement of risks, experts from the Action Coalition for Meaningful Transparency warn that platforms' risk assessments are too general. For example, stating a general risk to the right to freedom of expression is insufficient, they say in their policy paper. Service providers should also look at the specific nuances of their own products and services, for example how a policy on monetisation for content creators might affect the prevalence of disinformation in a particular product. (https://www.meaningfultransparency.tech/post/dsa-risk-assessment)

Anna Semenova from the Foundation for New Responsibility (SNV) is also in favour of specificity in risk assessments. In the data-based study on Youtube, it came out that certain problematic dynamics only become visible when looking at individual components such as the 'next video' function in isolation. For example, a subgroup was particularly frequently exposed to content such as esotericism and vaccination scepticism only in this function. Semenova criticises that the platforms mostly publish aggregated figures that are not meaningful. "My study shows how important it is to look more closely at how content views, for example, are distributed - both in terms of groups of people and platform components." (https://www.stiftung-nv.de/de/publication/thetreeof-complexity)

The Cerre authors advocate network analysis. "Platforms need to see how functionalities are interrelated, who they have close relationships with and what risks might arise from those relationships," says Broughton Micova. As an

example, she cites the terms of multi-channel networks with their contracted influencers and the resulting advertising revenue. Are there standards in terms of targeting children and in terms of the use of language? A remedy might not lie in content moderation on the platform, but in a code of conduct for multi-channel networks and their customers.

Anna-Katharina Meßmer and Martin Degeling (SNV) have dealt in particular with auditing and evaluating recommendation systems. They suggest working with scenarios: Abstract risks like Hatespeech should be "transformed into concrete, verifiable risk scenarios by defining the affected party and its characteristics, the damage, the involved elements of the platform and the further impact"(https://www.stiftung-nv.de/en/publication/auditing-recommendersystemsoverview-existing-audits-risk-assessments-and-studies).

"Our approach was to try to offer a scheme to the platforms. The feedback from the platform side was that it was too much to expect to develop and test a separate risk scenario for every risk, every damage and every product of individual platforms," Meßmer tells Tagesspiegel Background. She has little understanding for this. "The Facebook papers by Frances Haugen showed that there are a lot of tests going on internally. The DSA changes that it has to be more systematic and follow certain guidelines."

She understands the effort that the new Article 34 approaches entail for very large platforms, but in view of their billions in revenue, she has "little sympathy" for concrete implementation issues.