Automatisierte Bewertungssysteme im Jugendmedienschutz

Abschlussbericht zum Modellversuch

Ein gemeinsames Pilotprojekt der Freiwilligen Selbstkontrolle Fernsehen (FSF) und der Freiwilligen Selbstkontrolle Multimedia-Diensteanbieter (FSM),
unter Begleitung der Kommission für Jugendmedienschutz (KJM)

Stand: 27. Juli 2026

Inhaltsverzeichnis

Inhaltsverzeichnis 2
Zusammenfassung 3
1 Ziele des Pilotprojekts, gesetzlicher Kontext und Ablauf des Modellversuchs 5
1.1 Ausgangslage und Zielsetzung 5
1.2 Gesetzliche Rahmenbedingungen 5
1.3 Projektstruktur und Beteiligte 7
1.4 Ablauf des Modellversuchs 7
2 Die Prüfkriterien für ABS 9
2.1 Prüfansatz und Anwendungsbereich 9
2.2 Autonomieabhängiger Prüfmaßstab 9
2.3 Bewertungsmaßstab und Umgang mit konservativen Systemeinstellungen 9
2.4 Die drei Prüfsäulen der ABS-Begutachtung 10
2.5 Bewertungslogik, Beurteilungsspielräume und Gesamtwürdigung 12
3 Pilotverfahren: Begutachtung und Anerkennung von Merm:ai:d 13
3.1 Prüfauftrag und Entscheidungsgrundlagen 13
3.2 Prüfgegenstand: Merm:ai:d 13
3.3 Einordnung der Eigenständigkeit des Systems 14
3.4 Ergebnisse der Säule 1: Technische Eignung und Performance 14
3.5 Ergebnisse der Säule 2: Kohärenz mit der FSF-Spruchpraxis 15
3.6 Ergebnisse der Säule 3: Organisatorische Einbettung des ABS 16
3.7 Gesamtergebnis und Anerkennungsempfehlung 17
4 Empfehlungen für zukünftige ABS-Verfahren 18
4.1 Antrag und Dokumentation 18
4.2 Zusammensetzung des Prüfausschusses 22
4.3 Anforderungen an das Verfahren 23
4.4 Rechtsfolgen des Verfahrens 24
4.5 Aufgaben der Selbstkontrolleinrichtung in einem zukünftigen Verfahren 24
4.6 Gesteigerte Anforderungen an weitgehend autonome ABS 25
4.7 Perspektiven einer gesetzlichen Verankerung von ABS-Anerkennungsverfahren 26
4.8 Formulierungsvorschlag für einen Anerkennungstatbestand: § 7a JMStV-E 28
4.9 Institutionalisierung eines selbstkontrollübergreifenden Benchmark-Korpus 30
4.10 Satzungsförmige Fortschreibung der Prüfkriterien 31
5 Fazit und Ausblick 32

Zusammenfassung

Künstliche Intelligenz verändert die Praxis des Jugendmedienschutzes. Systeme, die audiovisuelle Inhalte automatisiert analysieren, relevante Szenen markieren und im Hinblick auf ihre Jugendschutzrelevanz vorbewerten oder vollautomatisiert Altersbewertungen erstellen, sind bei Medienanbietern bereits im Einsatz. Damit stellt sich die regulatorische Frage, unter welchen Voraussetzungen der Einsatz solcher automatisierter Bewertungssysteme (ABS) mit den Anforderungen des Jugendmedienschutz-Staatsvertrags (JMStV) vereinbar ist und wie ihre Qualität und ihr fachkundiger Einsatz durch Jugendschutzbeauftragte verlässlich nachgewiesen werden kann. Die Freiwillige Selbstkontrolle Fernsehen (FSF) und die Freiwillige Selbstkontrolle Multimedia-Diensteanbieter (FSM) haben hierzu, begleitet von der Kommission für Jugendmedienschutz (KJM) und unter Einbeziehung der Bundeszentrale für Kinder- und Jugendmedienschutz (BzKJ), von Januar bis Juli 2026 einen Modellversuch durchgeführt. Der vorliegende Bericht dokumentiert Anlage, Verlauf und Ergebnisse dieses Pilotprojekts.

Im Zentrum des Modellversuchs standen drei Arbeitsstränge: Zum einen sollte ein Kriterienkatalog für die Prüfung und Anerkennung von ABS durch eine interdisziplinäre Pilotgruppe aus KI-Technik, KI-Recht und KI-Ethik entwickelt werden. Zum anderen sollten diese Kriterien in einem realen Anerkennungsverfahren am Beispiel des Systems *Merm:ai:d* von RTL Deutschland angewandt und validiert werden. Schließlich sollte die Pilotgruppe Eckpunkte und Empfehlungen für Anforderungen an ein künftiges Regelverfahren zur Zertifizierung von ABS erarbeiten. Die Prüfkriterien (Version 3.1, Stand 13.07.2026) folgen einem Drei-Säulen-Ansatz, der die technische Eignung des Systems (Säule 1), die Kohärenz seiner Bewertungen mit der jugendschutzrechtlichen Spruchpraxis im Rahmen einer qualitativen Stichprobe (Säule 2) und seine Einbettung in unternehmensinterne Entscheidungs- und Governancestrukturen (Säule 3) erfasst. Ein gleitender Prüfmaßstab gewichtet die Säulen dabei nach dem Autonomiegrad des Systems: Je stärker die finale Entscheidung beim Menschen liegt, desto bedeutsamer wird die organisatorische Einbettung; je autonomer das System bewertet, desto strenger fällt die technische Prüfung aus.

Das Pilotverfahren bestätigte die Praxistauglichkeit dieses Drei-Säulen-Ansatzes. *Merm:ai:d* wurde als assistierendes System mit einzelnen teilautomatisierten Ausgabeelementen eingeordnet. Von den 16 Prüfkriterien wurden fünfzehn als erfüllt und eins als erfüllt mit Auflagen bewertet. Der von FSF-Prüfenden durchgeführte Stichprobenabgleich über 37 Sendungen ergab eine hohe Übereinstimmung der Systembewertungen mit der Prüfpraxis, wobei die dominante Abweichungsrichtung die jugendschutzrechtlich unkritische, bewusst konservative Überbewertung durch das ABS war. Auf der Ebene der finalen Altersbewertungen zeigte sich keine systematische, unkompensierte Unterbewertung. Die Gutachter:innen empfahlen die Anerkennung des Systems in der geprüften Konfiguration, befristet auf zwei Jahre und verbunden mit einer Auflage.

Aus der Pilotierung der *Merm:ai:d*-Anerkennung leitet der Bericht Eckpunkte und Empfehlungen für künftige ABS-Anerkennungsverfahren ab. Sie betreffen die Antrags- und Dokumentationspflichten der Unternehmen, die Zusammensetzung des Prüfausschusses, den Verfahrensablauf einschließlich des Stichprobenabgleichs, die Rechtsfolgen der Anerkennung sowie die Aufgaben der FSF bzw. der FSM als Trägerin des Verfahrens. Dabei skizziert der Bericht die gesteigerten Anforderungen an Metriken, Erklärbarkeit und Auditierbarkeit für weitgehend autonome Systeme und auf welchen Wegen die Anerkennung de lege ferenda im JMStV selbst verankert werden könnte. Hier unterbreitet die

Pilotgruppe einen konkreten Formulierungsvorschlag für eine gesetzliche Verankerung einer ABS-Anerkennung (§ 7a JMStV-E) und skizziert die notwendigen institutionellen Voraussetzungen. Dazu gehören im Kern ein einheitliches Benchmarkset als ABS- und selbstkontrollübergreifender Referenzkorpus für Kohärenzprüfungen und die satzungsförmig legitimierte Fortschreibung der Prüfkriterien.

Insgesamt zeigt der Modellversuch, dass sich die Qualität menschlicher Prüfpraxis auf automatisierte Systeme übertragen und überprüfbar machen lässt, sofern die Prüfung das soziotechnische Gesamtsystem aus Technik, unternehmensinternen Verfahren und fachkundigen Menschen in den Blick nimmt. Er kann die Grundlage für zukünftig mögliche strukturierte, wiederholbare und rechtssichere Anerkennungsverfahren schaffen, die Anbietern bestenfalls Rechtssicherheit bieten und das Schutzniveau des JMStV auch unter den Bedingungen der KI-Transformation gewährleisten.

1 Ziele des Pilotprojekts, gesetzlicher Kontext und Ablauf des Modellversuchs

1.1 Ausgangslage und Zielsetzung

Die automatisierte, meist KI-basierte Analyse audiovisueller Inhalte hat in den vergangenen Jahren erhebliche Fortschritte gemacht. In Medienunternehmen sind zunehmend Systeme verfügbar, die bei der Altersbewertung und jugendschutzrechtlichen Einordnung von audiovisuellen Inhalten Verwendung finden und Jugendschutzbeauftragte bei der Sichtung, Annotation und Verschlagwortung („Tagging“) sowie bei der Kategorisierung von Material und beim Generieren von Altersbewertungsvorschlägen, Deskriptoren und gutachterlichen Begründungstexten unterstützen. Solche automatisierte Bewertungssysteme (ABS) versprechen Effizienzgewinne, werfen aber zugleich die Frage auf, ob und wie ihre Anwendung fachlich fundiert, nach anerkannten Jugendschutzkriterien und transparent erfolgt. Ohne verbindliche Qualitätsmaßstäbe besteht das Risiko, dass Systeme mit unzureichender Erkennungsleistung oder ohne angemessene menschliche Kontrolle jugendschutzrechtliche Entscheidungen treffen oder Bewertungen prägen.

Vor diesem Hintergrund haben FSF und FSM im Benehmen mit der KJM einen Modellversuch durchgeführt, um Qualitätsstandards für automatisierte und KI-gestützte Bewertungssysteme zu entwickeln und ein Zertifizierungs- bzw. Anerkennungsverfahren zu erproben. Kennzeichnend für den gewählten Ansatz ist, dass nicht allein die Softwaresysteme selbst geprüft werden, sondern auch ihre organisatorische Einbettung in den Unternehmen. Die Prüfperspektive richtet sich damit auf das soziotechnische Gesamtsystem aus technischer Anwendung, organisatorischen Verfahren und fachkundigen Nutzerinnen und Nutzern.

Die zentralen Ziele des Modellversuchs waren (1) die Entwicklung eines strukturierten, wiederholbaren und rechtssicheren Verfahrens zur Anerkennung und Zertifizierung von ABS für audiovisuelle Medieninhalte, (2) die Sicherstellung der jugendschutzrechtlichen Qualität und Zielorientierung automatisierter oder teilautomatisierter Bewertungen sowie (3) die Unterstützung von Anbietern beim verantwortungsvollen Einsatz automatisierter Systeme und von Jugendschutzbeauftragten im Hinblick auf ihre Qualifizierung nach § 7 Abs. 5 JMStV und ihre Zertifizierung nach § 14a Abs. 1 Nr. 2 JuSchG. Für Unternehmen, die automatisierte Systeme einsetzen, soll das Verfahren verbindliche Maßstäbe setzen und Rechtssicherheit schaffen; für die Selbstkontrolleinrichtungen und die Aufsicht soll es gewährleisten, dass nur Systeme Verwendung finden, die nach definierten Qualitätsmaßstäben verlässlich sind und unter fachkundiger Aufsicht in Unternehmensprozessen eingesetzt werden.

Die Anerkennung eines ABS soll dabei als Nachweis dienen, dass ein System an den Zielen des JMStV ausgerichtet ist und die jeweils aktuelle Spruchpraxis der anerkannten Einrichtung der Freiwilligen Selbstkontrolle abbildet, dass es eine ausreichende Qualität bei der jugendschutzrechtlichen Bewertung aufweist und dass durch fachkundige menschliche Aufsicht bzw. Letztentscheidungen gewährleistet ist, dass der Anbieter seiner Verantwortung nach dem JMStV nachkommt.

1.2 Gesetzliche Rahmenbedingungen

Der rechtliche Bezugsrahmen des Modellversuchs ist der JMStV. Er verpflichtet Anbieter von Rundfunk und Telemedien, dafür Sorge zu tragen, dass Kinder und Jugendliche entwicklungsbeeinträchtigende

Angebote üblicherweise nicht wahrnehmen (§ 5 JMStV). Für das Fernsehen wird diese Pflicht traditionell über Sendezeitbeschränkungen operationalisiert, die mit den Altersstufen 0, 6, 12, 16 und 18 korrespondieren: Inhalte mit einer Eignung ab 16 Jahren dürfen grundsätzlich erst im Spätabendprogramm (22 bis 23 Uhr), Inhalte ohne Jugendfreigabe erst im Nachtprogramm (23 bis 6 Uhr) ausgestrahlt werden, während das Tagesprogramm (6 bis 20 Uhr) und das Hauptabendprogramm (20 bis 22 Uhr) den niedrigeren Altersstufen vorbehalten sind. Ergänzend sieht der JMStV inhaltsbeschreibende Zusatzhinweise (Deskriptoren) vor (§ 5c JMStV).

Inhaltlicher Maßstab der Altersbewertungen ist die in der Spruchpraxis der Selbstkontrolleinrichtungen und der KJM entwickelte Systematik der Wirkungsdimensionen. Grundlage der Beurteilung ist die kontextsensitive Einordnung von Darstellungen entlang der Dimensionen Angst, Gewalt, Desorientierung, Sexualität und Sprache, mit denen die Risiken einer Entwicklungsbeeinträchtigung wie etwa die Förderung gewaltbefürwortender Einstellungen, übermäßige Ängstigung oder sozialethische Desorientierung erfasst werden. Der Modellversuch verlangt von einem ABS nicht, dass seine internen Graduierungen den gesetzlichen Altersstufen folgen; maßgeblich ist, dass sie auf diese rückführbar sind und dass die jugendschutzrechtliche Letztentscheidung auf Anbieterseite anhand der gesetzlichen Stufen erfolgt.

Normativer Anknüpfungspunkt des Modellversuchs und der ABS-Prüfung ist § 7 Abs. 5 JMStV und die dort verankerte Expertise der Jugendschutzbeauftragten. Anbieter länderübergreifenden Fernsehens und bestimmter Telemedien haben Jugendschutzbeauftragte zu bestellen, die über die erforderliche Fachkunde verfügen, weisungsfrei gestellt sind und den Anbieter in Fragen des Jugendschutzes beraten. Der Modellversuch geht davon aus, dass sich diese Fachkunde beim Einsatz eines ABS auch auf das System selbst, seine Möglichkeiten und seine Grenzen erstrecken muss. Je autonomer ein ABS entscheidet, desto stärker verlagert sich die Fachkundeprüfung zudem auf die technische Performanz des Systems selbst. Auf bundesrechtlicher Ebene korrespondiert die Zertifizierung von Jugendschutzfachkräften nach § 14a Abs. 1 Nr. 2 JuSchG damit. Unabhängig von einer Anerkennung des ABS bleibt die Verantwortung des Anbieters für die Einhaltung seiner jugendschutzrechtlichen Pflichten in vollem Umfang bestehen. Die Anerkennung ersetzt den Grundsatz der Anbieterverantwortung nicht, sondern dient dem Nachweis der erforderlichen Fachkunde der Jugendschutzbeauftragten in Bezug auf (teil-)automatisierte und assistierende KI-Technologien im Jugendmedienschutz.

Eingebettet ist das Verfahren in das System der regulierten Selbstregulierung. FSF und FSM sind von der KJM anerkannte Einrichtungen der Freiwilligen Selbstkontrolle, ihre Spruchpraxis konkretisiert die unbestimmten Rechtsbegriffe des JMStV und bildet den inhaltlichen Vergleichsmaßstab, an dem die Kohärenz automatisierter Bewertungen gemessen wird. Für die Befristung der ABS-Anerkennung orientiert sich der Kriterienkatalog an der Anerkennungssystematik des § 11 JMStV für Jugendschutzprogramme und sieht eine entsprechende Maximalfrist von drei Jahren vor.

Schließlich ordnet der Kriterienkatalog das Verhältnis zur europäischen KI-Regulierung ein. Da ABS regelmäßig nur unternehmensintern und ausschließlich von Fachkräften mit jugendmedienschutzrechtlicher Expertise eingesetzt werden, handelt es sich typischerweise weder um autonome oder final entscheidende Systeme noch um Hochrisiko-KI-Systeme im Sinne der KI-Verordnung. Die Prüfperspektive richtet sich daher weniger auf Risiken für die Rechte späterer Rezipientinnen und Rezipienten oder Entscheidungsbetroffener, sondern im Kern auf die Orientierung des Systems an den Bewertungsdimensionen der aktuellen JMStV-Spruchpraxis und auf die kritische Interpretation der

Systemausgaben durch fachkundige Jugendschutzbeauftragte. Gleichwohl greift der Modellversuch anerkannte Governance-Standards für KI-Systeme auf.

1.3 Projektstruktur und Beteiligte

Der Modellversuch war als kooperatives Projekt auf mehreren Ebenen angelegt. Die Kriterienentwicklung und die gutachterliche Prüfung lagen bei einer interdisziplinär besetzten „ABS-Pilotgruppe“, in der die Bereiche KI-Technik, KI-Recht und KI-Governance bzw. KI-Ethik vertreten waren. Mitglieder der ABS-Pilotgruppe waren Prof. Dr. Oksana Arnold (Fachhochschule Erfurt, Professorin für Theoretische Informatik und Künstliche Intelligenz), Dr. Stephan Dreyer (Leibniz-Institut für Medienforschung | Hans-Bredow-Institut, Senior Researcher für Medienrecht und Media Governance) und Prof. Dr. Judith Simon (Universität Hamburg, Professorin für Ethik in der Informationstechnologie). Diese Zusammensetzung spiegelt den Anspruch, technische Leistungsfähigkeit, rechtliche Anschlussfähigkeit und ethische Vertretbarkeit des Systemeinsatzes in einem integrierten Prüfansatz zusammenzuführen.

Der Austausch zu grundlegenden Kriterien und Prüfstandards erfolgte in Zusammenarbeit mit der für den Modellversuch berufenen „Expert:innengruppe Jugendmedienschutz“, die die Perspektiven der KJM und der BzKJ einbrachte und in Form regelmäßiger Werkstattgespräche tagte. Ihr gehörten Dr. Annika Baumann (technische Sachverständige der KJM), Henning Mellage (Landesanstalt für Medien Nordrhein-Westfalen, Leiter der KJM-Arbeitsgruppe „Technischer Jugendmedienschutz“), Lukas Neuerburg (Bundeszentrale für Kinder- und Jugendmedienschutz) und Katharina Uharek Langhanke (Medienanstalt Berlin-Brandenburg) an. Über diese Gruppe war ein enger und kontinuierlicher Austausch zwischen FSF und FSM sowie der KJM und der BzKJ fest im Projekt verankert.

Als Pilotanbieter stellte RTL Deutschland sein seit Januar 2025 im Produktivbetrieb befindliches System *Merm:ai:d* für das Anerkennungsverfahren zur Verfügung und legte die erforderliche technische und organisatorische Dokumentation vor. Die FSF übernahm dabei die Durchführung einer qualitativen Stichprobe: Ein Prüfteam aus sechs hauptamtlich Prüfenden und erfahrenen FSF-Prüfenden sichtete die Vergleichsinhalte nach den regulären Maßstäben der Prüfpraxis und glich die Ergebnisse anschließend mit den Systemausgaben ab. Die FSF-Geschäftsstelle koordinierte das Gesamtprojekt.

Für alle Beteiligten galt während des gesamten Verfahrens ein strenger Vertraulichkeitsgrundsatz, der insbesondere dem Schutz der Geschäfts- und Betriebsgeheimnisse des antragstellenden Unternehmens dient. Der vorliegende Bericht trägt diesem Grundsatz Rechnung, indem er unternehmensinterne Detailinformationen nur in der für das Verständnis der Prüfergebnisse erforderlichen, aggregierten Form wiedergibt.

1.4 Ablauf des Modellversuchs

Der Modellversuch war auf den Zeitraum vom 5. Januar bis zum 31. Juli 2026 angelegt und gliederte sich in drei Phasen. In der ersten Phase wurden nach dem Projektauftakt die Prüf- und Anerkennungskriterien für ABS entwickelt und abgestimmt. In der zweiten Phase wurden die Kriterien im Pilotverfahren erprobt und auf Basis der dabei gewonnenen Erkenntnisse weiterentwickelt und finalisiert. Die dritte Phase diente der Ableitung von Empfehlungen zu Verfahren und Qualitätsmanagement, die in den Abschlussbericht des Projekts einfließen.

Das Prüfverfahren selbst beruhte auf mehreren aufeinander aufbauenden Verfahrensbestandteilen: (1) einer schriftlichen Dokumentation des antragstellenden Unternehmens, in der Aufbau, Funktionsweise und Prüflogik des Systems dargelegt werden, (2) einer Präsentation des Systems vor der Pilotgruppe einschließlich einer Live-Vorführung sowie (3) qualitative Vergleichstests mit regulären Prüfverfahren, die von FSF-Prüfgruppen vorgenommen und deren Ergebnisse systematisch an die Projektbeteiligten zurückgespielt wurden, um eine belastbare Gegenüberstellung mit den automatisierten Bewertungen zu ermöglichen.

Im Einzelnen verlief das Pilotverfahren in mehreren dokumentierten Schritten. Auf Basis des Diskussionsstands zu den Prüfkriterien übergab die Pilotgruppe dem Anbieter zunächst die Anforderungen an die ABS-Dokumentation. RTL Deutschland übermittelte daraufhin die technische und organisatorische Dokumentation zu *Merm:ai:d* (Stand 13.04.2026). In der zweiten Sitzung der ABS-Pilotgruppe am 22.04.2026 wurden die Prüfkriterien diskutiert und weiterentwickelt, die Anbieterdokumentation gesichtet und bewertet sowie offene Punkte unmittelbar mit dem Anbieter erörtert; im Fokus standen dabei die Grenzen der Einsichtnahme in Einzelkomponenten und Fremdmodelle, die Maßnahmen zur Erklärbarkeit und Nachvollziehbarkeit sowie die Implementierung des Systems in die Governancestruktur der Jugendschutzbeauftragten. Im Nachgang übermittelte der Anbieter eine ergänzende Dokumentation mit Governance-Fokus (Stand 06.05.2026).

Parallel führte das FSF-Prüfteam die qualitative Stichprobe durch (s. dazu Abschnitt 3.5). Die Ergebnisse wurden am 18.05.2026 im zweiten Treffen der JMS-Expert:innengruppe vorgestellt und diskutiert. Die Tagesordnung umfasste neben dem Bericht aus der Pilotgruppe die Anwendung der Prüfkriterien auf das Pilotsystem, den Austausch zwischen Anbieter und Pilotgruppe sowie die Optimierung der Prüfkriterien. Auf dieser Grundlage wurden die Prüfkriterien zur Version 3.0 (Stand 02.07.2026) konsolidiert und das Anerkennungsgutachten verfasst. Beide Dokumente wurden bei der dritten Pilotgruppensitzung am 06.07.2026 diskutiert und vorläufig verabschiedet. Im Rahmen des Treffens wurden die letzten noch offenen Punkte mit RTL diskutiert; das Unternehmen legte kurzfristig letzte Dokumente und Nachweise vor. Den Abschluss des Modellversuchs bilden die drei Ergebnisdokumente: die finalen Prüfkriterien in der Version 3.1 (s. Abschnitt 2), das Anerkennungsgutachten zu Merm:ai:d (s. Abschnitt 3), sowie dieser Abschlussbericht mit den zentralen Projektaktivitäten und den Vorschlägen für ein mögliches künftiges FSF-Verfahren (s. Abschnitt 4).

2 Die Prüfkriterien für ABS

2.1 Prüfansatz und Anwendungsbereich

Der im Modellversuch entwickelte Kriterienkatalog (Version 3.1, Stand 13.07.2026) formuliert die Anforderungen an die Anerkennung eines ABS als aussagekräftiges Entscheidungsunterstützungssystem für Jugendschutzbeauftragte. Die Kriterien sind bewusst technikneutral angelegt: Sie sollen unabhängig von der konkreten technischen Ausgestaltung, von Art und Umfang der automatisierten Analysen, von der Form des Systemoutputs und von der Art seiner Einbindung in menschliche Entscheidungsverfahren anwendbar sein und zu nachvollziehbaren Prüfentscheidungen führen. Gleichzeitig ist ihr Anwendungsbereich klar begrenzt: Die Prüfkriterien gelten nur für ABS im unternehmensinternen Einsatz durch fachkundige Personen.

Gegenstand der Anerkennung ist stets das gesamte ABS in der geprüften Konfiguration einschließlich seiner organisatorischen Einbettung. Mit Blick auf die hohe Dynamik der KI-Entwicklungen wird eine Anerkennung befristet erteilt, in Anlehnung an § 11 JMStV mit einer Maximalfrist von drei Jahren. Außerdem lösen wesentliche Änderungen am System, an seinen Einzelkomponenten oder an der Entscheidungsarchitektur eine Pflicht zur Nachprüfung aus.

2.2 Autonomieabhängiger Prüfmaßstab

Angesichts der Vielfalt möglicher Systemgestaltungen ist ein zentraler Ansatz der Prüfung ein gleitender Prüfmaßstab, der von dem Autonomiegrad des Systems abhängt. Je eigenständiger ein ABS eine inhaltliche Prüfung oder eine konkrete Altersbewertung vornimmt, umso stärker zielt die Prüfung auf die Zuverlässigkeit und Performanz des Systems selbst. Je deutlicher die finale Jugendschutzentscheidung von einem menschlichen Jugendschutzbeauftragten getroffen wird, umso mehr verlagert sich die Prüfung auf die Eignung der Entscheidungsarchitektur aus fachkundigen Personen und ABS im Unternehmen.

Die Prüfenden ordnen das zu begutachtende System dafür zu Beginn auf einem Autonomie-Kontinuum ein, das von rein assistierenden Systemen (das System liefert relevante Hinweise, der Mensch bewertet) über teilautomatisierte Systeme (das System schlägt Bewertungen vor, der Mensch muss die Entscheidung unter Berücksichtigung der systemischen Bewertung als eigene deklarieren) bis zu weitgehend autonomen Systemen reicht (das System klassifiziert eigenständig, die Entscheidung wird größtenteils ohne menschliche Überprüfung übernommen). Als weitgehend autonom gilt ein System im Sinne der Prüfkriterien, wenn die Jugendschutzbeauftragten im Laufe eines beliebigen Programmtags weniger als 20 Prozent der ABS-Entscheidungen kontrollieren. Die Einordnung eines ABS auf diesem Kontinuum bestimmt dann die Gewichtung der drei Prüfsäulen (s. Abschnitt 2.4) in der Prüfung.

2.3 Bewertungsmaßstab und Umgang mit konservativen Systemeinstellungen

Entscheidungsmaßstab der technischen Bewertung ist nicht die optimale technische Performance eines ABS, sondern die Ausrichtung des Systemeinsatzes an den Zielen des JMStV. Hintergrund ist, dass der JMStV mit seinem Grundsatz der Anbieterverantwortung einen Anreiz für Anbieter darstellt, ein ABS „schärfer“ einzustellen, als die Spruchpraxis dies erfordert. Mit einer sensibleren Erkennung oder einem strenger bewertenden System können falsch-negative Ergebnisse und damit

jugendschutzrechtliche Haftungsrisiken reduziert werden. Bewusst in Kauf genommene, strengere Inhaltsbewertungen eines ABS sind zulässig, solange die Systemergebnisse so in die unternehmerischen Entscheidungsverfahren eingebunden sind, dass die beabsichtigten Ungenauigkeiten durch menschliche Entscheidungen berücksichtigt und kompensiert werden. Darüber hinaus sind auch geringe Ungenauigkeiten bzgl. der Altersklassifikation, d.h. Abweichungen von einer Altersstufe im Vergleich zur Bewertung durch die JMS-Expert:innen, zulässig.

Mit restriktiveren Systemeinstellungen korrespondieren insoweit erhöhte Anforderungen an die menschliche Entscheidungsseite. Um dem Risiko von Overblocking und einer schleichenden Normalisierung strengerer Entscheidungen zu begegnen, müssen sich die menschlichen Entscheiderinnen und Entscheider der konservativen Systemeinstellung bewusst sein und den ABS-Output entsprechend informiert und reflektiert interpretieren.

2.4 Die drei Prüfsäulen der ABS-Begutachtung

Die Bewertung eines ABS erfolgt in drei aufeinander bezogenen Prüfebenen, die das System aus technischer, qualitativ-vergleichender und organisatorischer Perspektive betrachten. Säule 1 (Technische Eignung und Performance) prüft die grundsätzliche Eignung, Datenbasis und Entscheidungsqualität des Systems, namentlich Architektur, Genauigkeit, Konsistenz, Robustheit, Nachvollziehbarkeit und Adaptivität. In Säule 2 (Kohärenz mit der Spruchpraxis) werden die ABS-Entscheidungen im Rahmen einer qualitativen Stichprobe mit den Prüfergebnissen einer Selbstkontrolleinrichtung verglichen und im Hinblick auf systemische Abweichungen oder Auffälligkeiten analysiert. Säule 3 (Organisatorische Einbettung) schließlich prüft die Entscheidungsarchitektur in der Unternehmenspraxis, die Fachkunde der ABS-Nutzenden und das Qualitäts- und Risikomanagement inklusive der unternehmensinternen Governance.

Das Bewertungsgewicht der Säulen folgt dem Autonomiegrad: Das Gewicht der Säule 1 steigt mit dem Autonomiegrad des ABS, das Gewicht der Säule 3 mit der Bedeutung der menschlichen Letztentscheidung. Säule 2 hat demgegenüber ein gleichbleibend hohes Gewicht, weil der Spruchpraxisabgleich für jede Systemklasse zentrale Hinweise auf JMStV-Orientierung und Zuverlässigkeit liefert. Die Gewichtungsmatrix in der Prüfkriterien konkretisiert den gleitenden Prüfmaßstab als Orientierung.

Autonomiegrad	**Säule 1 (technisch)**	**Säule 2 (Kohärenz)**	**Säule 3 (Organisation)**
assistierend	ergänzend	hoch	maßgeblich
teilautomatisiert	hoch	hoch	hoch
weitgehend autonom	maßgeblich	hoch	ergänzend

Tabelle 1: Gewichtungsmatrix der Prüfsäulen nach Autonomiegrad (Prüfkriterien, Abschnitt F)

Die insgesamt 16 Einzelkriterien der drei Säulen sind vergleichbar aufgebaut: Jedes Kriterium benennt den Prüfgegenstand, die Prüfanweisung, strukturierende Leitfragen, relevante Indikatoren für die Erfüllung sowie den bestehenden Bewertungsspielraum. Die Indikatoren sind dabei ausdrücklich Anhaltspunkte und keine abschließende Checkliste. Das Fehlen eines Kriteriums begründet für sich genommen keine Nichterfüllung, sondern ist in eine begründete Gesamteinschätzung einzustellen und kann durch andere, strenger umgesetzte Aspekte oder Prozessarchitekturen kompensiert werden. Die folgende Übersicht fasst die Kriterien zusammen.

Nr.	Kriterium	Prüfgegenstand (Kurzfassung)
Säule 1 – Technische Eignung und Performance		
1.1	**Systemüberblick**	Angemessenheit und Nachvollziehbarkeit der Gesamtarchitektur als Pipeline verbundener Komponenten
1.2	**Grundsätzliche Eignung**	Geeignetheit des Gesamtsystems zur Analyse bewegter Bild-, Ton- und Sprachinhalte einschließlich des Szenenkontexts
1.3	**JMStV-Zielorientierung**	Ausrichtung der Systemlogik an den Schutzzielen des JMStV und Abbildung der fünf Wirkungsdimensionen in den Modellen
1.4	**Datenbasis und Modelltraining**	Qualität, Herkunft und Jugendschutzbezug der Trainings-, Kalibrierungs- und Validierungsdaten
1.5	**Genauigkeit**	Zuverlässigkeit der Erkennung jugendschutzrelevanter Inhalte; Benchmarks bzw. Metriken
1.6	**Konsistenz und Fairness**	Gleichbehandlung gleicher und vergleichbarer Inhalte über Zeit und Nutzung hinweg; keine sachfremden Differenzierungen; Maßnahmen gegen Drift
1.7	**Robustheit**	Widerstandsfähigkeit gegen Fehlbedienung und bewusste Täuschung
1.8	**Nachvollziehbarkeit und Transparenz**	Protokollierung und Nachvollziehbarkeit der Systembewertungen
1.9	**Adaptivität und Change Management**	Fähigkeit zur kontrollierten Anpassung an neue Modelle, Inhaltsformen und eine sich wandelnde Spruchpraxis; Erkennung und Korrektur fehlerhafter Bewertungen im laufenden Betrieb
Säule 2 – Kohärenz der Systembewertungen mit der jugendschutzrechtlichen Spruchpraxis		
2.1	**Methodik des Stichprobenabgleichs**	Sachgerechte Anlage und Durchführung des qualitativen Abgleichs; Stichprobe und Sichtungsverfahren
2.2	**Übereinstimmungs- und Abweichungsanalyse**	Systematisierung von Abweichungen nach Richtung, Genre und Darstellungsform; besonderes Gewicht auf ABS-Abweichungen nach unten und mögliche „blinde Flecken“
2.3	**Umsetzung von Transparenz und Nachvollziehbarkeit für Fachkräfte**	Erkennbarkeit der Bewertungsgründe des ABS und ihre Anschlussfähigkeit an die Maßstäbe der JMStV-Spruchpraxis
Säule 3 – Einbettung in Verfahren, Governance und Qualitätsmanagement		
3.1	**Entscheidungsarchitektur**	Zusammenspiel von ABS und fachkundigen Personen; Letztentscheidung entlang der gesetzlichen Altersstufen

Nr.	Kriterium	Prüfgegenstand (Kurzfassung)
3.2	**Fachliche Qualifikation und Ressourcen**	Jugendschutzfachliche Bewertungskompetenz, technische Betriebskompetenz und ausreichende Ressourcen
3.3	**Qualitätsmanagement**	Einbindung des ABS-Einsatzes in ein anerkanntes oder gleichwertig gelebtes Qualitätsmanagementsystem
3.4	**Anpassung an die JMStV-Spruchpraxis**	Verstetigter Abgleich mit der sich verändernden JMStV-Spruchpraxis und verstetigter Austausch mit einer Selbstkontrolleinrichtung

Tabelle 2: Übersicht der 16 Prüfkriterien (Prüfkriterien, Abschnitte C bis E)

2.5 Bewertungslogik, Beurteilungsspielräume und Gesamtwürdigung

Jedes Kriterium wird auf einer dreistufigen Skala beurteilt, wobei die Einstufung jeweils zu begründen und zu dokumentieren ist. Ein Kriterium ist „erfüllt“, wenn es nach Überzeugung der Prüfenden gewahrt ist und etwaige Schwächen unerheblich oder anderweitig kompensiert sind. Es ist „mit Auflagen erfüllt“, wenn es im Grundsatz gewahrt ist, aber im Nachgang behebbare Schwächen aufweist. Die Erfüllung wird dann mit konkret benannten, fristgebundenen Auflagen festgestellt. „Nicht erfüllt“ ist ein Kriterium, wenn es nicht gewahrt und nicht durch andere Vorkehrungen kompensiert erscheint. Eine Anerkennung kommt dann ohne Nachbesserung in der Regel nicht in Betracht.

Die Prüfung und Anerkennungsentscheidung ergehen dann in drei Schritten. Zunächst erfolgt eine Einzelbewertung der Kriterien, die in einer säulenweisen Würdigung münden. Abschließend erfolgt eine Gesamtwürdigung entlang der Gewichtungsmatrix, d.h. welche Säule erhält welches Gewicht angesichts der Einordnung der Eigenständigkeit der ABS-Entscheidungen. Kennzeichnend ist dabei die Kompensationslogik des Katalogs: Schwächen einer Säule können durch Stärken einer anderen ausgeglichen werden, etwa technische Ungenauigkeit durch eine robuste menschliche Letztkontrolle oder eine schwächere menschliche Kontrolle durch hohe technische Zuverlässigkeit. Eine absolute Grenze zieht der Katalog bei der Kohärenzprüfung im Rahmen der qualitativen Stichprobe. Ein systematisches, nicht kompensiertes Unterbewerten relevanter Inhalte, insbesondere mit mehr als einer Altersstufendifferenz, führt zum Aussetzen einer Anerkennung und zur Aufforderung zur Nachbesserung. Das Prüfgutachten dokumentiert die Einzelbewertungen, die gewählte Gewichtung, die genutzten Bewertungsspielräume und die relevanten Überlegungen und Gründe.

3 Pilotverfahren: Begutachtung und Anerkennung von Merm:ai:d

3.1 Prüfauftrag und Entscheidungsgrundlagen

Gegenstand der im Rahmen des Modellversuchs pilotierten ABS-Begutachtung war die Prüfung des automatisierten Bewertungssystems *Merm:ai:d* von RTL Deutschland anhand der ABS-Prüfkriterien in der Version 3.1. Zu prüfen war, ob das System in der dokumentierten Konfiguration und organisatorischen Einbettung als aussagekräftiges Entscheidungsunterstützungssystem für Jugendschutzbeauftragte gemäß der Prüfkriterien anerkannt werden kann. Die Begutachtung erfolgte unter Berücksichtigung der Bewertungsspielräume und des beschriebenen gleitenden Prüfmaßstabs. Festgestellte Schwächen oder Dokumentationslücken wurden, soweit erforderlich, in konkrete Auflagen überführt.

Der Begutachtung zugrunde lagen neben den Prüfkriterien drei Anbieter- bzw. Verfahrensdokumente: Von der Anbieterin wurde eine technische und organisatorische Systemdokumentation (Stand 13.04.2026) sowie eine Dokumentation zur Erklärbarkeit, Nachvollziehbarkeit und unternehmensinternen Governance (Stand 06.05.2026) vorgelegt. Von der FSF wurden die Ergebnisse der qualitativen Stichprobe am 18.05.2026 präsentiert und diskutiert. In der dritten Pilotgruppen-Sitzung am 06.07.2026 wurden letzte Informationen und Nachweise von RTL im Rahmen eines gegengezeichneten Protokolls in das Verfahren einbezogen.

3.2 Prüfgegenstand: Merm:ai:d

Bei *Merm:ai:d* handelt es sich um eine in der Cloud betriebene, modulare Analysepipeline zur Erkennung, Klassifizierung, Kontextualisierung und Aufbereitung potenziell jugendschutzrelevanter Inhalte in audiovisuellen Medien. Das System betrachtet Bild und Text in mehreren KI-gestützten Analysestufen bezüglich ihrer Jugendschutzrelevanz, aggregiert die Ergebnisse zeitcodiert zu Szenen und bereitet sie zu einer Ergebnisdatei auf. Diese enthält eine Inhaltszusammenfassung, ein Kurzgutachten zur Jugendschutzrelevanz, eine zeitcodierte Liste aller relevanten Fundstellen mit Bild- und Dialogbeschreibungen, getrennte Bild- und Sprachkategorien, Vorschläge für Deskriptoren nach § 5c JMStV, eine sogenannte Watch Ratio als Anteil der als relevant erkannten Laufzeit an der Gesamtlänge sowie einen ausdrücklich als derzeit noch experimentell gekennzeichneten Vorschlag zur Altersbewertung mit einer Wahrscheinlichkeitsverteilung über die Stufen 0, 6, 12, 16 und 18.

Die Bewertungslogik des Systems ist aus der realen Jugendschutzpraxis der Anbieterin abgeleitet: Das Kategoriensystem mit 17 Kategorien in den Bereichen Gewalt, Verletzung, Ängstigung, Suizid, Alkohol und Drogen, Sexualität, Sprache und Diskriminierung sowie die Prompts der LLM-Komponenten wurden aus der Analyse und Faktorisierung von Jugendschutz-Schnittlisten gewonnen, die in den vergangenen 15 Jahren bei RTL zur Entfernung problematischer Szenen entstanden sind, und unter ausdrücklicher Bezugnahme auf die Prüfgrundsätze der FSF entwickelt. Das System ist seit Januar 2025 im Produktivbetrieb und wird ausschließlich als Werkzeug für ausgebildete Jugendschutzbeauftragte eingesetzt. Die abschließende jugendschutzrechtliche Bewertung verbleibt ausnahmslos bei den fachlich zuständigen Personen. Der Anbieter formuliert den Grundsatz, dass am Anfang und am Ende der KI-Pipeline immer ein Mensch steht. Nach den Erfahrungswerten aus dem Produktivbetrieb erlaubt die Ergebnisdatei im Zusammenspiel mit dem gezielten Blick in die Sendung eine fundierte Beurteilung bei einer Sichtzeitersparnis von rund 80 Prozent.

3.3 Einordnung der Eigenständigkeit des Systems

Das Gutachten ordnet *Merm:ai:d* als assistierendes System mit einzelnen teilautomatisierten Ausgabeelementen ein. Im Kern liefert das System relevante Hinweise in Form von Fundstellen, Timecodes, Kategorien und Zusammenfassungen, die eine Jugendschutz-Fachkraft bewertet. Dabei dienen die Inhaltszusammenfassung und das Kurzgutachten ausdrücklich nicht als abschließendes jugendschutzrechtliches Fazit, sondern vor allem als Einstiegspunkt für die menschliche Bewertung. Zwar enthält der Output mit Kurzgutachten, Deskriptorvorschlägen und Altersbewertungsvorschlag durchaus bewertende Elemente, die der Erscheinungsform eines teilautomatisierten Systems nahekommen; der konkrete Altersbewertungsvorschlag ist bislang aber als rein experimentell gekennzeichnet und nicht für den ungeprüften Export in Produktivsysteme bestimmt. Der Vorschlag soll eine mögliche Tendenz aufzeigen, während die Deskriptoren durch die Jugendschutzbeauftragten im weiteren Verfahren plausibilisiert werden und jederzeit veränderbar sind. Die Schwelle zu einer weitgehenden Autonomie des Systems ist nicht erreicht: Jede Ergebnisdatei wird durch die Jugendschutzbeauftragten geprüft, relevante Sendungen mit einer Watch Ratio von über 50 Prozent werden stets vollständig gesichtet, und die finale Freigabeentscheidung wird ausschließlich menschlich getroffen und dokumentiert.

Für die Gewichtungsmatrix (s. Abschnitt 2.2) bedeutet diese Einordnung, dass die organisatorische Einbettung (Säule 3) maßgeblich ist, die Kohärenz mit der Spruchpraxis (Säule 2) hohes Gewicht hat und die technische Bewertung (Säule 1) lediglich ergänzendes Gewicht trägt. Vorsorglich hat das Gutachten die technische Säule gleichwohl mit erhöhter Prüftiefe behandelt, um Anwendungserfahrung im Hinblick auf die Prüfkriterien sammeln zu können. Außerdem bliebe auf diese Weise die Bewertung auch bei einer Einordnung als teilautomatisiertes System tragfähig.

3.4 Ergebnisse der Säule 1: Technische Eignung und Performance

Von den zehn Kriterien der technischen Säule wurden alle als erfüllt bewertet. Als Stärken hebt das Gutachten die bis auf Komponentenebene dokumentierte Gesamtarchitektur hervor, die systematische Verarbeitung der für den Einsatzbereich maßgeblichen Bild- und Sprachebene einschließlich des Szenen- und Gesamtwerkkontexts, die vollständige Adressierung der JMStV-Wirkungsdimensionen über das Kategoriensystem und die dokumentierte Kategorienkombinatorik sowie die auf 15 Jahren realer Spruch- und Schnittpraxis beruhende, jugendschutzfachlich annotierte Datenbasis. Der bewusste Verzicht auf die Auswertung der Geräusch- und Musikebene ist empirisch begründet und vom Kriterienkatalog gedeckt; bekannte Systemgrenzen sind offengelegt und werden durch Nutzungsregeln organisatorisch aufgefangen.

Positiv gewürdigt werden ferner der bewusst nicht selbstlernende Systemansatz, der reproduzierbare Ergebnisse gewährleistet und Änderungen des Systemverhaltens einem kontrollierten Änderungsakt vorbehält. Das A/B-Testverfahren mit einem genreübergreifenden Benchmarkset nach jedem externen Modellwechsel dient als zuverlässiger Regressionsschutz. Die Nachvollziehbarkeit der Systembewertungen unmittelbar am Originalmaterial über Timecodes, Originaltextauszüge und Kategorien erleichtert die Identifizierung von Fehlverhalten und Artefakten für Fachkräfte. Definierte Korrekturwege auf Einzelfall- wie auf Systemebene verfügen über spezielle Feedback-Tools und Wartungsverfahren.

Im Rahmen der Prüfung der Kriterien Datenbasis (1.4) und Genauigkeit (1.5) wurde der Anbieterin empfohlen, im Rahmen eines gemeinsamen Workshops mit der FSF mögliche Hinweise der

qualitativen Stichprobe auf Unterrepräsentation bestimmter Formate und Genres in den Trainingsdaten zu diskutieren und ggf. eine Bias-Analyse durchzuführen.

3.5 Ergebnisse der Säule 2: Kohärenz mit der FSF-Spruchpraxis

Den Kern der Prüfung von Säule 2 bildete der von der FSF durchgeführte Stichprobenabgleich. Geprüft wurden 37 Sendungen, die vorrangig aus einer Sendewoche von RTL Deutschland stammen und die Bandbreite der Genres und Sendezeitschienen abbilden: Fiktion (unter anderem Krimi, Action, Thriller, Drama, Fantasy, Komödie, Krankenhausserien), Unterhaltung (Comedy, Reality, Casting-Shows, Scripted Reality, Soaps), Information (Reportagen, Dokumentationen, True Crime), Lifestyle sowie Kinder- und Jugendformate einschließlich Animationsserien. Ausgeschlossen wurden nicht vorlagefähige Livesendungen und Nachrichtenformate (Berichterstattungsprivileg des § 5 Abs. 6 JMStV). Lifestyle-Formate wie etwa Kochsendungen waren wegen fehlender Jugendschutzrelevanz nur in geringer Zahl in der Stichprobe vertreten. Die Prüfung erfolgte durch ein Team von sechs Prüfenden mit zufallsbasierter Zuteilung der Sendungen. Die klassische Sichtung wurde ausdrücklich ohne vorherige Einsicht in die ABS-Ergebnisse durchgeführt und mit dem Notiz- und Dokumentationsinstrumentarium YouKit der FSF dokumentiert. Ergänzend erfolgte eine Gegenprüfung an Sendungen mit niedriger Watch Ratio und einzelnen Sendungen höherer Relevanz, um zu ermitteln, ob sich die Jugendschutzrelevanz allein aus den Systemprotokollen erfassen lässt, sowie ein Gegencheck mit den tatsächlich von RTL gewählten Sendezeitschienen. Das Kriterium der methodischen Durchführung (Kriterium 2.1) wurde als erfüllt bewertet.

Die inhaltliche Analyse (Kriterium 2.2) ergab insgesamt einen positiven Befund: Die Inhaltszusammenfassungen des ABS liefern eine treffende Beschreibung und Analyse des Materials und erfassen relevante Kontextfaktoren weitgehend. Die gelisteten Szenen weisen eine hohe Schnittmenge mit den Notizen der Prüfenden auf, und die Anzahl relevanter Szenen ist im Durchschnitt ähnlich. Zuverlässig erkannt werden nach dem Befund der Prüfenden deutlicher Rassismus, Diskriminierung, Stereotype, die Beteiligung von Kindern, Gewaltszenen und Sexualität. Weniger zuverlässig erkannt werden die Ästhetisierung von Gewalt (etwa durch Zeitlupe), implizite, über die Bildmontage vermittelte Botschaften und Alkoholkonsum. Zugleich zeigt das System – erwartete – Übersensibilitäten, etwa bei der Emotionserkennung über Gesichter, bei vermeintlicher Nacktheit, bei Feuer und Explosionen sowie bei buchstäblicher Textauffassung („jemanden auf die Folter spannen“). Die Watch Ratio erwies sich als grober, aber valider Indikator mit starkem Zusammenhang zur Tagesprogramm-Tauglichkeit einer Sendung.

Bei der Gegenüberstellung der notierten Szenen war die dominante Abweichungsrichtung die überinklusive Markierung durch das System, also eine strengere Notation als in der menschlichen Prüfung. Diese Überbewertungen sind mit Blick auf die bewusst konservative Systemkonfiguration und den Bewertungsmaßstab der Prüfkriterien unkritisch, da sie das jugendschutzrechtliche Haftungsrisiko nicht erhöhen, sondern lediglich erhöhte Anforderungen an die kritische menschliche Befassung stellen. Prüfungsrelevant waren demgegenüber Abweichungen nach unten: Auf der Ebene der Szenenzählung fielen mehrere Sendungen mit deutlicher Unterdeckung auf, die sich formatbezogen bündeln lassen und zeithistorische bzw. dokumentarische Formate, True-Crime-Formate mit dokumentarischer Machart, animierte Formate sowie einzelne Actionfilme mit ästhetisierter Gewaltdarstellung betreffen. Bei konkretem Abgleich wurde klar, dass die szenenbasierte zahlenmäßige Abweichung nach unten bei dem ABS an vom System zusammengefassten Szenen lag. Die Beschreibung relevanter Szenen hat in diesen Fällen mehrere aufeinander

folgenden Szenen als einen durchgehend jugendschutzrelevanten Abschnitt verarbeitet. Eine Fehlerkennung oder -bewertung lag in diesen Fällen nicht vor.

Für die Anerkennungsfrage relevant war zudem, dass sich die Szenenzählung und -bewertung auch nicht in systematischen Unterbewertungen auf der Ebene der finalen jugendschutzrechtlichen Einordnung niedergeschlagen hat. Der Vergleich der FSF-Ergebnisse mit den tatsächlichen Sendezeitschienen bzw. Altersbewertungen von RTL zeigte ganz überwiegend Übereinstimmung. In einer Reihe von Fällen ordnete RTL die Inhalte sogar strenger ein als die FSF-Prüfung, während mildere Einordnungen durch das ABS bzw. die RTL-Letztentscheidung gar nicht beobachtbar waren. Ein systematisches, unkompensiertes Unterbewerten relevanter Inhalte lag entsprechend nicht vor. Dies wurde als zusätzliches Indiz dafür gewertet, dass die beobachteten Szenenzahlabweichungen durch die menschliche Entscheidungsarchitektur mit ihrer Gesamtsichtungslogik, dem Kontextwissen der Nutzenden und dem Kontrollblick auf die Intensität in jedem Fall aufgefangen werden. Das Kriterium wurde deshalb als erfüllt bewertet.

Das Kriterium der Transparenz (Kriterium 2.3) der Bewertungsgründe des ABS bewährte sich in der Prüfpraxis: Die Szenen werden zusammen mit Kategorien und Beschreibungen eingeordnet, die Deskriptoren benennen die wesentlichen Gründe der Alterseinstufung, und das Kurzgutachten weist be- und entlastende Faktoren aus. Die verwendeten Fachbegriffe sind dabei nicht immer präzise auf die Prüfsprache der Selbstkontrolle abgestimmt, verbleiben aber auf einer verständlichen und anschlussfähigen Diktion. Die Prüfenden hielten zugleich fest, dass das Protokoll als Ganzes zu lesen ist und dass erst die Kombination der einzelnen Ausgabekomponenten ein relevantes Gesamtbild ergibt. Das System ist insoweit nicht selbsterklärend, seine Nutzung durch geschulte Jugendschützer:innen bleibt erforderlich. Das Fazit der Stichprobenuntersuchung lautete, dass das System ein hilfreiches Instrument zur jugendschutzrechtlichen Einschätzung darstellt und einen treffenden Eindruck des Medieninhalts liefert.

3.6 Ergebnisse der Säule 3: Organisatorische Einbettung des ABS

Die für die assistierende Funktion von *Merm:ai:d* maßgebliche dritte Säule wurde im Schwerpunkt anhand der Governance-Dokumentation geprüft, die die organisatorische Einbettung des ABS als soziotechnisches Gesamtsystem beschreibt. Das Kriterium der Entscheidungsarchitektur (Kriterium 3.1) wurde als erfüllt bewertet: Die finale jugendschutzrechtliche Bewertung wird ausnahmslos durch die zuständigen, weisungsfrei gestellten Jugendschutzbeauftragten entlang der gesetzlichen Altersstufen bzw. der korrespondierenden Sendezeitschienen getroffen und dokumentiert. Die Verantwortungszuordnung ist in einer RACI-Matrix[1] formalisiert und die Eingriffspunkte der Fachkräfte (Prüfung der Ergebnisdatei, Entscheidung über die Sichtungstiefe, Korrektur der Deskriptoren, Feedbackwege) sind konkret dokumentiert.

Als strukturell besonders wirksam wird die obligatorische Vollsichtung bei einer Watch Ratio über 50 Prozent oder sonst erkennbar hoher Relevanz eingeschätzt. Gerade dort ist eine rein selektive Sichtung nicht vorgesehen; das wirkt etwaigen Anreizen zur unkritischen Übernahme von Systembewertungen bei besonders jugendschutzrelevanten Inhalten entgegen. Der beschriebene Gegencheck der Stichprobe lieferte den empirischen Beleg für die tatsächliche Wirksamkeit dieser Governance: Die Freigabepraxis des Anbieters weicht erkennbar eigenständig von den

[1] Eine RACI-Matrix ist eine klassische Tabelle im Projektmanagement, um Aufgaben, Rollen und Verantwortlichkeiten klar zuzuordnen und Missverständnisse zu vermeiden.

Systemausgaben ab und liegt im Vergleich zu den FSF-Ergebnissen auf der gleichen Stufe oder ist strenger. Der in der Governance-Dokumentation ausdrücklich adressierte Automation Bias wird durch den Human-in-the-Loop-Grundsatz, Schulungen, definierte Systemgrenzen und die Watch-Ratio-Regel eingehegt.

Auch die fachliche Qualifikation und die Ressourcen (Kriterium 3.2) wurden als erfüllt bewertet: Das System wird ausschließlich durch ausgebildete bzw. erfahrene Jugendschutzbeauftragte genutzt, die weite Teile des Systems mitkonzipiert, -trainiert und -validiert haben; die technische Kompetenz ist zulässigerweise auf gesonderte Rollen (Datenteam, IT-Betrieb) verteilt. Insbesondere für künftige, an der Systementwicklung unbeteiligte Mitarbeitende ist eine Einarbeitung anhand eines Leitfadens bzw. einer Bedienungsanleitung vorgesehen. Das unternehmensinterne Qualitätsmanagement (Kriterium 3.3) wird als kontinuierlicher, fachlich-technischer Regelkreis mit Verfahrensbeschreibungen, mehrstufiger Feedbackschleife und A/B-Regressionstests gelebt; ein übergreifender, zertifizierter Qualitätsmanagement-Rahmen ist nicht nachgewiesen, die (Weiter-)Entwicklung und der Produktivbetrieb des ABS wird aber durch Fachgremien auf mehreren Ebenen begleitet und reflektiert (Einbindung von Compliance- und Legal-Abteilung; Governance-Runde bestehend aus Jurist:innen, Techniker:innen und inhaltlich arbeitenden Mitarbeitenden; Ethikrat). Zudem wurden im Rahmen der Compliance-Aktivitäten mit dem EU AI Act Risikoprüfungen durchgeführt.

Die Anpassung an die JMStV-Spruchpraxis (Kriterium 3.4) ist in *Merm:ai:d* angelegt (anpassbare Deskriptorensets, Wortlisten, Prompts und Schwellwerte). Es fehlt aber bislang ein systematischer, formeller Austauschkanal mit der FSF über den Modellversuch hinaus. Die Schaffung eines solchen Kanals wurde als Auflage in das Prüfgutachten aufgenommen.

3.7 Gesamtergebnis und Anerkennungsempfehlung

In der Gesamtwürdigung gelangt das Gutachten zu dem Ergebnis, dass *Merm:ai:d* und seine Einbettung im Unternehmen die Prüfkriterien erfüllen. Das System ist konsequent an den Zielen des JMStV ausgerichtet; Kategoriensystem, Prompts und Deskriptoren sind aus den Wirkungsdimensionen der Spruchpraxis und aus 15 Jahren dokumentierter unternehmensinterner Jugendschutzpraxis hergeleitet. Der unabhängige Stichprobenabgleich hat die Kohärenz der Systembewertungen mit der JMStV-Spruchpraxis bestätigt, und die organisatorische Einbettung mit gewährleisteter menschlicher Letztentscheidung, der Watch-Ratio-Vollsichtungsregel, dem Regressionsschutz durch A/B-Tests und dem nicht selbstlernenden, reproduzierbaren Systemansatz erfüllt die Anforderungen an ABS und ihre Einbettung. Ein verbleibender Punkt – der systematische Austausch mit der Selbstkontrolle – ist fristgerecht behebbar.

Die Gutachter:innen haben der FSF und der FSM entsprechend empfohlen, *Merm:ai:d* in der geprüften Konfiguration und mit dieser organisatorischen Einbettung als aussagekräftiges Entscheidungsunterstützungssystem für Jugendschutzbeauftragte im Sinne der ABS-Prüfkriterien anzuerkennen. Dem Erstverfahren entsprechend sollte die Anerkennung zunächst auf zwei Jahre befristet und mit der angesprochenen Auflage versehen sein. Die Anerkennung bezieht sich auf die dokumentierte Konfiguration, wesentliche Änderungen am System, an seinen Einstellungen oder an der Entscheidungsarchitektur lösen eine Nachprüfungspflicht aus. Insbesondere ist der experimentelle Altersbewertungsvorschlag von der Anerkennung nur in seiner dokumentierten Funktion als nicht exportfähiger Tendenzhinweis umfasst; eine spätere Aufwertung zu einem regulären Bewertungselement wäre eine wesentliche, nachprüfungspflichtige Änderung.

4 Empfehlungen für zukünftige ABS-Verfahren

Das dritte Projektziel des Modellversuchs war die Ableitung von Empfehlungen für ein mögliches verstetigtes Anerkennungsverfahren in der Zukunft. Absehbar werden ABS bei Mitgliedsunternehmen und bei Nichtmitgliedern Eingang in die Jugendschutzpraxis finden. Mit der Schaffung eines Anerkennungsverfahrens können Selbstkontrolleinrichtungen entsprechende Zertifizierungsanfragen von Unternehmen gleichförmig und mit den Vorgaben des JMStV abgestimmten Kriterien bearbeiten.

Auf Basis der Prüfkriterien und den dortigen Verfahrensregeln (Stichprobe, Bewertungslogik, Befristung, Nachprüfung), den im Laufe der pilotierten Anerkennung erprobten Verfahrensbestandteilen sowie den hierbei gewonnenen praktischen Erfahrungen können Empfehlungen in den Prozess einer etwaigen Verfahrensentwicklung eingespeist werden. Dabei werden einzelne Punkte wie die erforderliche Anbieterdokumentation, die Zusammensetzung eines Prüfausschusses, die möglichen Konsequenzen einer Anerkennung und der Zuschnitt der FSF-Aufgaben im Folgenden als begründete Empfehlungen im Folgenden nachvollziehbar entwickelt (Abschnitte 4.1 bis 4.5).

Die Abschnitte 4.6 bis 4.10 gehen über die reine Projektdokumentation hinaus: Sie enthalten Einordnungen der Pilotgruppe zu den (gesteigerten) Anforderungen an weitgehend autonome ABS, zu möglichen Wegen einer gesetzlichen Verankerung von ABS-Anerkennungsverfahren im JMStV, zur Erstellung einer unabhängigen Prüfreferenz in Form eines einheitlichen Benchmark-Sets sowie die Möglichkeit einer satzungsförmigen Fortschreibung der Prüfkriterien.

4.1 Antrag und Dokumentation

Jedes künftige ABS-Verfahren sollte mit einem förmlichen Antrag des Unternehmens bei der Geschäftsstelle der Selbstkontrolle eingeleitet werden. Der Antrag sollte das System, seinen Einsatzbereich und das abgedeckte Inhaltsspektrum beschreiben und – bestenfalls – eine Selbsteinordnung des ABS in eine der drei Autonomie-Grade enthalten. Da diese Selbsteinordnung zugleich die Prüftiefe und die Säulengewichtung vorstrukturiert, kann der Anbieter antizipieren, welche Teile der Dokumentation er in welcher Breite und Tiefe vorzulegen hat. Hier könnte die Geschäftsstelle im Vorfeld begleiten bzw. beraten.

Das Pilotverfahren hat mit Blick auf die Dokumentation gezeigt, dass sich die Antragsdokumentation grundsätzlich gut in drei Teile gliedern lässt, die mit den Prüfsäulen korrespondieren:

- Eine technische Systemdokumentation mit Architekturübersicht, Beschreibung der bewertungsrelevanten Komponenten und ihrer informationsverarbeitenden Technologie, Datenbasis und Annotationsgrundlagen, verfügbaren Evaluationsmetriken, Fallbackverhalten, Konfigurations- und Wartungsmöglichkeiten und bekannten Systemgrenzen;
- die für Säule 2 (qualitative Stichprobe) erforderlichen Systemzugänge bzw. Systemausgaben zu den Vergleichsinhalten;
- eine governancebezogene Dokumentation, die die organisatorische Einbettung als soziotechnisches Gesamtsystem beschreibt und auf Verantwortungszuordnungen, Entscheidungsarchitektur und Human-in-the-Loop-Prozesse, Schulungs- und Qualifikationspraktiken sowie Qualitäts- und Änderungsmanagement und den systematischen Abgleich mit Veränderungen in der Spruchpraxis der Selbstkontrolle zum Inhalt hat.

Für spätere Verfahren sollte die jeweilige Geschäftsstelle der Selbstkontrolle hierfür standardisierte Dokumentationsanforderungen und ggf. Vorlagen bereitstellen, wie sie im Piloten auf Basis des Diskussionsstands der Prüfkriterien an den Anbieter übergeben wurden. Wichtig ist, dass die Geschäftsstelle der Selbstkontrolle bereits im Rahmen der Planung einer Antragstellung dem Anbieter beratend zur Seite steht.

Tabelle: Überblick über mögliche Inhalte der Anbieterdokumentation

Nr.	Bezeichnung des Prüfkriteriums	Mögliche Inhalte der Dokumentation	Skalierung nach Autonomiegrad
Vorgelagerte Anforderungen			
A.1	**Zweck und Anwendungsbereich – Stammangaben und Einsatzbereich**	Antragsteller mit Rechtsform und Anschrift; benannter Jugendschutzbeauftragter; bei Sitz außerhalb Deutschlands Zustellungsbevollmächtigter. Beschreibung des Einsatzbereichs: Inhaltsspektrum nach fiktionalen, non-fiktionalen und nutzergenerierten Inhalten, Genres, Sprachfassungen, Mengengerüst, Ausspielwege sowie etwaige ausgenommene Inhaltsklassen.	Unverändert für alle Autonomiegrade. Bei Spezialisierung auf einzelne Genres ist die Begrenzung des praktischen Einsatzbereichs verbindlich zu erklären.
A.3	**Gleitender Prüfmaßstab – Selbsteinstufung des Autonomiegrads**	Einordnung des Autonomiegrads: assistierendes, teilautomatisiertes oder weitgehend autonomes ABS.	Unverbindliche Selbsteinstufung, die in erster Linie zur Abschätzung der Dokumentationstiefe dient.
Säule 1 – Technische Eignung und Performance			
1.1	**Systemüberblick**	Architekturdokumentation im Überblick mit Pipeline und Funktion der Komponenten, eingesetzter informationsverarbeitender Technologie (Modelltyp, Version, Hersteller, Hostingform), Zusammenspiel und internen Schnittstellen der Komponenten; Datenflüsse sowie Ein- und Ausgabeformate; Kennzeichnung der Komponenten, die jugendschutzrelevante Bewertungen treffen oder beeinflussen; Verortung menschlicher Prüf- und Eingriffspunkte. Überblicksartige Dokumentation der gespeicherten und der manuellen Überprüfung zugänglichen Daten, der Logging- und Überwachungsprozesse sowie der Konfigurationsmöglichkeiten zur Änderung des Systemverhaltens einschließlich Berechtigungskonzept und Versionierung.	Assistierend: funktionaler Überblick. Teilautomatisiert und weitgehend autonom: Einblick bis auf Komponentenebene.
1.2	**Grundsätzliche Eignung zur Bewertung audiovisueller Inhalte**	Verarbeitete Modalitäten (Bild, Bewegtbild, Ton, Sprache) und Art ihrer Einbeziehung und ggf. Kombination; Verfahren der Kontextberücksichtigung durch sequenzielle und szenenbasierte Analyse, Einordnung der Gesamterzählung, gegebenenfalls Genreerkennung; erklärte blinde Flecken.	Multimodalität ist nicht zwingend, sofern alle relevanten Modalitäten systematisch berücksichtigt werden. Spezialisierung zulässig, wenn sie offengelegt und der Einsatzbereich entsprechend begrenzt wird.
1.3	**JMStV-Zielorientierung und Abbildung der Wirkungsdimensionen**	Überblick über die intern genutzten Kategorien und ihren Zuordnungen zu den Wirkungsdimensionen Angst, Gewalt, Desorientierung, Sexualität und Sprache; Abbildungsvorschrift der internen Graduierung von Altersbewertungen auf die Altersstufen 0, 6, 12, 16 und 18; Verfahren zur Erkennung unzulässiger Darstellungen; dokumentierte Abdeckungslücken.	Assistierend; teilautomatisiert: Vollständige 1:1-Abbildung nicht erforderlich.
1.4	**Datenbasis und Grundlage des Modelltrainings**	Herkunft und Umfang der Trainings-, Kalibrierungs- und Validierungsdaten; bei vortrainierten Modellen die zum Finetuning und zur Validierung genutzten eigenen Daten; Annotationsrichtlinie und Schulung der annotierenden Personen; Datenqualität nach Vollständigkeit,	Assistierend: Herkunftsangabe. Teilautomatisiert: zusätzlich Annotationsmaßstab. Weitgehend autonom:

Nr.	Bezeichnung des Prüfkriteriums	Mögliche Inhalte der Dokumentation	Skalierung nach Autonomiegrad
		Genauigkeit, Konsistenz, Korrektheit, Einheitlichkeit; Dokumentation der Trennung von Trainings- und Evaluationsdaten; ggf. bekannte Verzerrungen und Repräsentationslücken.	zusätzlich Qualitäts- und Biasanalysen.
1.5	**Genauigkeit (Erkennungsqualität)**	Ermittelte Metriken einschließlich Trefferquote, Präzision und Konfusionsmatrix; verwendete Benchmark-Sets mit Herkunft und Dokumentation der Ausschließlichkeit der Trainingsdaten; Verteilung der Fehler über Wirkungsdimensionen und Altersstufen; gesonderter Ausweis der falschnegativen Ergebnisse; etwaige systematische blinde Flecken; Angabe von Stichprobengrößen.	Nicht verlangt wird optimale technische Leistung. Schwächen sind hinnehmbar, wenn sie bekannt sind und durch Einstellungen oder menschliche Kontrolle kompensiert werden. Bei weitgehend autonomen Systemen sind ermittelte Metriken zwingend.
1.6	**Konsistenz und Fairness**	Reproduktionstests bei identischen Eingangsinformationen, bei stochastischen Modellen mit Angabe von Wiederholungszahl und Streuungsmaß; Stabilitätstests gegen bewertungsirrelevante Variationen wie Format, Auflösung, Sprachfassung oder Farbraum; Prüfung auf sachfremde Bewertungsunterschiede.	Streuung ist bei lernenden Systemen (DeepLearning; GenAI) unvermeidlich; maßgeblich ist, dass Falschklassifikationen erkannt, strukturell untersucht und im Betrieb nicht jugendschutzrelevant verzerrend wirksam werden.
1.7	**Robustheit gegen Fehlbedienung und Umgehung**	Dokumentierte Testfälle zu Normalbetrieb, Fehlkonfiguration, Grenzsituation und Ausnahmesituation; Sicherheit der Schnittstellen gegenüber manipulierenden oder fehlerhaften Eingaben; Systemverhalten im Störfall (Abbruch, Eskalation, Rückfall auf rein menschliche Prüfung); Monitoring- und Eskalationsstrukturen; Ressourcen für den unterbrechungsfreien Betrieb.	Die Anforderungen orientieren sich am realistischen internen Einsatz durch Fachpersonal; bei weitgehend autonomen Systemen ist eine vollständige Testmatrix vorzulegen.
1.8	**Nachvollziehbarkeit und Transparenz**	Loggingumfang, Loggingkonfiguration und Zugänglichkeit für die relevanten Personen; Angabe verwendeter interpretierbarer Modelle oder nachgelagerter Erklärverfahren; Darstellung der Begründungselemente und Deskriptoren im Bewertungsvorschlag; mindestens drei vollständige Beispielprotokolle einer Prozesskette bzw. eines vollständigen Outputs.	Der erreichbare Erklärbarkeitsgrad hängt vom Modelltyp ab. Assistierend: Beschreibung genügt. Teilautomatisiert und weitgehend autonom: Musterausgaben zwingend. Fehlende Erklärbarkeit kann durch Benchmarking mit ausgewiesenen Metriken oder durch Maßnahmen der Säule 3 teilweise kompensiert werden.
1.9	**Adaptivität und Change Management**	Definierter Update-Prozess mit Turnus und Freigabestufen; Regressionstestkonzept; Monitoring neuer Inhaltsphänomene und der Spruchpraxis; definierte Melde-, Feedback- und Korrekturwege für Fehlbewertungen im laufenden Betrieb; Rollback-Fähigkeit. Bei der Nutzung von Fremdmodellen zusätzlich Verfahren zur Erkennung stillschweigender Modellwechsel.	Erstprüfung: plausibler Nachweis geeigneter Prozesse. Folgeprüfung: Nachweis der Wirksamkeit anhand der tatsächlichen Pflegepraxis.
Säule 2 – Kohärenz der Systembewertungen mit der jugendschutzrechtlichen Spruchpraxis (hier: Mitwirkungs- und Bereitstellungsleistungen)			
2.1	**Methodische Durchführung des Stichprobenabgleichs**	Vollständiger Katalog- oder Programmwochenauszug mit Metadaten je Inhalt als Grundgesamtheit für die Stichprobenziehung; technischer Zugang und Sichtungsrechte für die Einzelprüferinnen und Einzelprüfer; Bereitstellung des ABS für den Parallellauf; Bewertungen bereits geprüfter Inhalte für den rückwärtsgerichteten Abgleich. Erklärung, dass der Parallellauf in der fixierten Produktivkonfiguration erfolgt.	Umfang und Zusammensetzung der Stichprobe können angepasst werden; bei autonomeren Systemen oder heterogenen Katalogen kann eine größere, gezielt um Grenzfälle

Nr.	Bezeichnung des Prüfkriteriums	Mögliche Inhalte der Dokumentation	Skalierung nach Autonomiegrad
			angereicherte Stichprobe angezeigt sein.
2.2	**Übereinstimmungs- und Abweichungsanalyse**	Vollständige ABS-Rohausgaben zu den Stichprobeninhalten einschließlich Zwischenergebnissen, Konfidenz- und Schwellenwerten sowie Begründungselementen; je Inhalt die finale Jugendschutzentscheidung mit Zeitpunkt, entscheidender Person und Begründung etwaiger Abweichung vom Systemvorschlag. Die Systematisierung der Abweichungen und die Bildung von Ursachenhypothesen erfolgen prüferseitig.	Abweichungen nach oben sind regelmäßig unkritischer als Abweichungen nach unten; Differenzen um eine Klasse liegen eher im akzeptablen Bereich als solche um zwei.
2.3	**Umsetzung von Nachvollziehbarkeit und Transparenz der ABS-Bewertungskriterien für Fachkräfte**	Nachweis, dass die für eine Bewertung maßgeblichen Faktoren im ausgelieferten Output ausgewiesen werden; Darstellung der Deskriptoren und Zusatzhinweise; Beleg der Anschlussfähigkeit der Systembegründungen an die JMStV-Spruchpraxis; Aufbereitung der Systementscheidungen für das Fachpersonal.	Assistierend; teilautomatisiert: Der erreichbare Transparenzgrad hängt vom Modelltyp ab. Maßgeblich ist, dass die Entscheidenden die Ergebnisse fachlich einordnen und plausibilisieren können. Weitestgehend automatisiert: Die Determiniertheit der Entscheidungen ist erforderlich; ggf. sind Temperaturregler des Modells entsprechend einzustellen. Zusätzlich sind post-hoc XAI-Methoden zur Nachvollziehbarkeit und Transparenz von Entscheidungen notwendig und zur Überprüfung von Entscheidungen vorzulegen.
Säule 3 – Einbettung in Verfahren, Governance und Qualitätsmanagement			
3.1	**Entscheidungsarchitektur und menschliche Letztentscheidung**	Prozessdarstellung mit Rollen, Informationsgrundlagen, Eingriffs- und Überstimmungspunkten sowie Eskalationswegen; Festlegung, wer die gesetzliche Altersstufe final verantwortet; Regeln für Übernahme und Abweichung von ABS-Entscheidungen; Pflicht zur Einzelfalldokumentation der Letztentscheidung. Maßnahmen gegen Automatisierungsverzerrung: Schulungen, Blindprüfquoten, Vier-Augen-Regelungen.	Gewicht des Kriteriums steigt mit der Bedeutung der menschlichen Letztentscheidung. Die konkrete Architektur ist frei wählbar, solange sie dem Autonomiegrad des ABS angemessen ist. Bei bewusst strenger eingestellten Systemen ist ein erhöhter Nachweis kritischer Befassung mit dem Systemoutput zu führen.
3.2	**Fachliche Qualifikation und Ressourcen**	Zertifizierung der Jugendschutzbeauftragten durch eine Einrichtung der Freiwilligen Selbstkontrolle im Sinne des § 14a Abs. 1 S. 2 Nr. 2 JuSchG; Nachweis der technischen Betriebskompetenz; Schulungsplan und Schulungshistorie; Dokumentation der KI-Kompetenz des eingesetzten Personals; Personalkapazität im Verhältnis zum Prüfvolumen.	Kompetenzprofil und Ressourcenniveau bemessen sich nach Autonomiegrad, Verfügbarkeits- und Durchsatzanforderungen sowie dem Inhaltsspektrum. Technische und jugendschutzfachliche Kompetenz dürfen auf mehrere Rollen / Teams verteilt sein.
3.3	**Qualitätsmanagement**	Zertifikat nach ISO 9001 oder Nachweis eines gleichwertigen Standards; Prozessbeschreibungen mit ausgewiesenem ABS-Bezug; ggf. Berichte der internen Audits der letzten Periode; Korrektur- und Vorbeugemaßnahmen.	Eine formale Zertifizierung ist nicht zwingend; ein nachweislich gleichwertiges, gelebtes Qualitätsmanagement genügt, insbesondere bei kleineren Anbietern.

Nr.	Bezeichnung des Prüfkriteriums	Mögliche Inhalte der Dokumentation	Skalierung nach Autonomiegrad
3.4	**Anpassung an die JMStV-Spruchpraxis**	Nachhaltige Kommunikationsverbindung zur Selbstkontrolleinrichtung; Teilnahme an Weiterbildungen; sonstige Quellen des Spruchpraxis-Monitorings; ggf. Protokolle des Austauschs mit Selbstkontrolleinrichtungen.	Regelmäßigkeit und Tiefe können risikoorientiert gestaffelt werden. Bei stabiler Übereinstimmung sind größere Abstände vertretbar, nach festgestellten Schwächen oder Systemänderungen sind engmaschigere Abgleiche geboten.
Anforderungen ohne Entsprechung im Kriterienkatalog (Empfehlung)			
	Geheimschutz- und Veröffentlichungskonzept	Kennzeichnung der als Betriebs- oder Geschäftsgeheimnis eingestuften Bestandteile der Dokumentation; Einverständnis mit einer veröffentlichungsfähigen Kurzfassung des Prüfergebnisses (in Anlehnung an § 11 Abs. 4 S. 3 JMStV).	Unverändert für alle Autonomiegrade.

Der Modellversuch hat gezeigt, dass Anbieter in innovativen Produktbereichen wie der Entwicklung von ABS ein hohes Geheimhaltungsinteresse haben, damit unternehmenseigene Geschäftsgeheimnisse nicht im Rahmen eines ABS-Anerkennungsverfahrens „leaken“ oder durch Wettbewerber einsehbar sind. Für die eingereichten Unterlagen sollte daher grundsätzlich ein strenger Vertraulichkeitsgrundsatz zum Schutz der Geschäfts- und Betriebsgeheimnisse der antragstellenden Unternehmen gelten. Darüber hinaus bedarf es einer Einverständniserklärung in die Veröffentlichung einer aggregierten Kurzfassung der Begutachtung, um notwendige Öffentlichkeit und Nachvollziehbarkeit einer ABS-Prüfung sicherzustellen.

4.2 Zusammensetzung des Prüfausschusses

Für die Zusammensetzung eines künftigen ABS-Prüfausschusses empfiehlt die Pilotgruppe eine im Vergleich zu reinen inhaltsprüfenden Ausschüssen interdisziplinärere Besetzung als Mindeststandard. Aus Sicht der Verfasser:innen hat sich die interdisziplinäre Zusammensetzung im Modellversuch bewährt. Die Kombination aus technischer bzw. Informatik-Expertise im Bereich der Künstlichen Intelligenz, aus jugendmedienschutzrechtlicher Expertise und aus ethisch-soziotechnischer Expertise deckt die bei ABS-Prüfungen relevanten Prüfungsfelder ab. Ein Prüfausschuss sollte vor diesem Hintergrund mindestens drei Mitglieder umfassen, die zusammen die drei Kompetenzfelder abdecken. Ergänzend sollte die FSF-Prüfpraxis vertreten sein, sei es in Form eines weiteren Ausschussmitglieds und/oder durch die enge Zusammenarbeit mit der Geschäftsstelle und den hauptamtlich Prüfenden auch zur Koordinierung des Stichprobenabgleichs in Säule 2. Dabei hat sich im Modellversuch die Trennung zum eigenständigen Prüfteam der Selbstkontrolleinrichtung zur Durchführung des Stichprobenabgleichs in Säule 2 bewährt. Sie sollte auch in zukünftigen Verfahren beibehalten werden, so dass die Inhalte ohne Kenntnis des Systemoutputs unvoreingenommen gesichtet und bewertet werden. Der Prüfausschuss diskutiert und würdigt die Ergebnisse der Stichprobe anschließend gutachterlich. Diese Trennung sichert die Expertise der menschlichen Vergleichsbewertung und belässt die medieninhaltliche Prüfkompetenz auf Seiten der Selbstkontrolleinrichtung, während die Bewertung des ABS im Drei-Säulen-Ansatz dem Prüfausschuss obliegt.

Die Mitglieder des ABS-Prüfausschusses sollten unabhängig und frei von Interessenkonflikten gegenüber dem antragstellenden Unternehmen sein. Entsprechende Befangenheitsregeln und Verschwiegenheitspflichten sind in der Verfahrensordnung bzw. den Prüfvereinbarungen mit den Ausschussmitgliedern vorzusehen. Für den nachhaltigen Aufbau und die Kontinuität einer Spruchpraxis der Prüfausschüsse im ABS-Bereich kann als Ansatz eine Bestellung der Mitglieder für eine feste Amtszeit angedacht werden, ggf. mit der Möglichkeit, für einzelne Verfahren weitere Sachverständige hinzuzuziehen.

4.3 Anforderungen an das Verfahren

Nach Eingang und Vollständigkeitsprüfung des Antrags eines Anbieters bei der Selbstkontrolleinrichtung wird der ABS-Prüfausschuss einberufen. Der Ausschuss nimmt zunächst die Einordnung des Systems hinsichtlich der Autonomie der Systementscheidungen vor, da sie die Säulengewichtung und die erforderliche Prüftiefe bestimmt. Es folgen die Dokumentationsprüfung aus Perspektive der Säulen 1 und 3, eine Präsentation des Systems vor dem Ausschuss einschließlich einer Live-Vorführung sowie ein strukturiertes Dialogformat mit dem Anbieter zur Klärung offener Punkte, wie es sich im Piloten insbesondere zu den Grenzen der Einsichtnahme in Fremdmodelle und zu den Erklärbarkeitsmaßnahmen bewährt hat.

Parallel führt das Prüfteam der Selbstkontrolleinrichtung den Stichprobenabgleich in Säule 2 nach den Vorgaben des Kriterienkatalogs durch: Zunächst erfolgen die Auswahl der Stichprobe und die unabhängige Sichtung durch mindestens zwei Einzelprüfer:innen ohne Kenntnis des Systemoutputs. Danach wird der systematische Abgleich mit Analyse der Abweichungsrichtung und -muster durchgeführt. Im Anschluss ist eine Gegenprüfung bereits bewerteter Inhalte und ein Gegencheck mit den finalen Anbieterentscheidungen vor allem bei assistierten und teilautomatisierten ABS zu empfehlen. Für weitestgehend automatisierte ABS ist vor allem interessant, wie stichprobenartige Prüfungen in den automatisierten Prozess integriert werden und wie Fehlerkorrekturen in den Change-Management-Prozess einfließen. Gerade die Gegenprüfung sowie der Gegencheck mit den finalen Anbieterentscheidungen haben sich im Modellversuch als aufschlussreich erwiesen, weil sie die tatsächliche Wirksamkeit der menschlichen Entscheidungsarchitektur empirisch belegen und damit Verschränkungen von Säule 2 und Säule 3 in der Praxis sichtbar machen. Auch wird dadurch besser nachvollziehbar, über welche Qualifikationen und Kontextwissen Jugendschutzbeauftragte in der Entscheidungsarchitektur haben müssen. In der Diskussion zwischen Prüfteam und ABS-Prüfausschuss kann auch noch einmal ein besonderes Augenmerk auf falschnegative Ergebnisse gelegt werden.

Das Verfahren mündet in ein schriftliches Gutachten, das die Prüfkriterien anwendet, ihre Erfüllung, Teilerfüllung oder Nichterfüllung begründet, etwaige genutzte Bewertungsspielräume transparent macht und etwaige festgestellte (behebbare) Schwächen in fristgebundene Auflagen überführt. Für die Verfahrensdauer sollte ein Regelzeitrahmen definiert werden, um Planungssicherheit für die Antragstellenden zu gewährleisten. Der Modellversuch wurde innerhalb eines halben Jahres durchgeführt, was für Regelverfahren mit standardisierten Unterlagen eher als Obergrenze erscheint. Bei guter Dokumentation, überschaubarer System- und Einbettungskomplexität ist ein Durchlaufen des Verfahrens theoretisch auch in deutlich kürzerer Zeit denkbar. Bei besonders komplexen Prüfungen muss die Geschäftsstelle die Verfahrensdauer begründet verlängern können.

Zudem müssten Kosten- und Gebührenregelungen, Fristen für Nachreichungen und die Behandlung von Verfahrensabbrüchen bzw. Antragsrücknahmen in einer Verfahrensordnung geregelt werden.

4.4 Rechtsfolgen des Verfahrens

Die Anerkennung durch einen Prüfausschuss bescheinigt dem Antragstellenden in erster Linie, dass ein ABS in der geprüften Konfiguration und Einbettung als aussagekräftiges Entscheidungsunterstützungssystem für dessen Jugendschutzbeauftragte geeignet ist. Sie wirkt als Nachweis des Vorliegens der Anerkennungsvoraussetzungen (JMStV-Zielorientierung, Abbildung der Spruchpraxis, ausreichende Bewertungsqualität, fachkundige menschliche Aufsicht). Sie wird befristet erteilt, im Erstverfahren auf maximal zwei Jahre, im Übrigen – in Anlehnung an § 11 JMStV – auf maximal drei Jahre, und kann mit fristgebundenen Auflagen verbunden werden. Werden Auflagen innerhalb der Frist nicht erfüllt – die Kontrolle der Umsetzung sollte der Geschäftsstelle der Selbstkontrolleinrichtung obliegen –, liegen die Voraussetzungen der ABS-Anerkennung nicht vor (auflösende Bedingung).

Die Anerkennung selbst ist versions- und konfigurationsgebunden: Wesentliche Änderungen am System, an seinen Einstellungen oder an der Entscheidungsarchitektur lösen eine Nachprüfungspflicht aus, die je nach Umfang auch durch die Geschäftsstelle und ggf. nach Aktenlage durchgeführt werden kann. Stellt der Stichprobenabgleich ein systematisches, nicht kompensiertes Unterbewerten relevanter Inhalte fest, insbesondere mit mehr als einer Altersstufendifferenz, kann die Anerkennung ausgesetzt und der Anbieter zur Nachbesserung aufgefordert werden. Den Anbieter trifft insoweit eine Informationspflicht gegenüber der Geschäftsstelle im Hinblick auf relevante Systemveränderungen.

Über die verfahrensinternen Wirkungen hinaus ist die rechtliche Reichweite der Anerkennung klar zu begrenzen und entsprechend zu kommunizieren: Die Anerkennung ersetzt weder die grundsätzliche Anbieterverantwortlichkeit nach dem JMStV noch die sonstige individuelle Fachkunde der Jugendschutzbeauftragten nach § 7 Abs. 5 JMStV. Sie konkretisiert aber, dass deren Fachkunde beim Einsatz eines ABS auch das System, seine Möglichkeiten und seine Grenzen umfasst, und dokumentiert, dass die betrieblichen Voraussetzungen hierfür vorliegen. Innerhalb des Systems der regulierten Selbstregulierung im JMStV kann die Anerkennung daneben als qualifiziertes Indiz für die Erfüllung von Sorgfaltsanforderungen des Anbieters bei der Wahrnehmung der Anbieterverantwortlichkeit nach § 5 Abs. 1 JMStV wirken (vgl. § 24 Abs. 1 Nr. 4 JMStV). Eine förmliche Bindungswirkung gegenüber der KJM oder den Landesmedienanstalten entfaltet sie nach geltender Rechtslage nicht und sollte auch nicht suggeriert werden (zu Möglichkeiten einer gesetzlichen Verankerung s. unten Abschnitt 4.7). Perspektivisch erscheint es sinnvoll, die Berücksichtigung anerkannter ABS in den Prüfgrundsätzen und Verfahrensordnungen der Selbstkontrolleinrichtungen zu verankern und den Dialog mit der KJM über die aufsichtsrechtliche Einordnung fortzuführen.

4.5 Aufgaben der Selbstkontrolleinrichtung in einem zukünftigen Verfahren

Als mögliche Trägerin eines künftigen ABS-Verfahrens übernimmt die Einrichtung der Freiwilligen Selbstkontrolle ein Bündel neuer Aufgaben. Die Geschäftsstelle wäre dann Herrin der Anerkennungsverfahren. Sie nähme Anträge entgegen, prüfte die Vollständigkeit der Unterlagen, stellte die standardisierten Dokumentationsanforderungen bereit, organisierte die Sitzungen des ABS-

Prüfausschusses und die Systempräsentationen, wählte die Stichprobeninhalte nach der vom Prüfausschuss vorgegebenen Anzahl und Formatvielfalt aus und wickelte Nachprüfungen nach Aktenlage ab. Die Prüfabteilung stellte das Team für den Stichprobenabgleich und führte die Gegenprüfungen durch. Dafür wären Prüfer:innen ggf. speziell für die Arbeit mit ABS-Protokollen zu qualifizieren, da sich im Piloten gezeigt hat, dass die Systemausgaben nicht selbsterklärend sind und als Ganzes gelesen werden müssen.

Zu den fortlaufenden Aufgaben gehörten daneben die Pflege und Fortschreibung der Prüfkriterien – im Zusammenspiel mit den ABS-Prüfausschussmitgliedern und dem Kuratorium –, die regelmäßig an die fortschreitende KI-Entwicklung und die Erfahrungen in der Prüfpraxis anzupassen wären. Hinzu träte der wiederkehrende Austausch mit den anerkannten Anbietern. Die Selbstkontrolle sollte zudem ein Register der anerkannten Systeme mit Konfigurationsstand, Befristung und Auflagenstatus führen, die Erfüllung der Auflagen fristbezogen überwachen und über die Verfahren in geeigneter, den Vertraulichkeitsgrundsatz wahrender Form gegenüber der KJM und der Öffentlichkeit berichten. Schließlich würde sich mit dem Verfahren ein Qualifizierungsauftrag ergeben: Die im Modellversuch gewonnenen Erkenntnisse über Möglichkeiten und Grenzen von ABS sollten in Schulungsangebote für Jugendschutzbeauftragte einfließen und damit das dritte Projektziel, die Qualifizierung der Anwendergruppe, dauerhaft absichern. Interessierten Unternehmen sollten die entsprechenden Informationen und Antragsunterlagen bereitgestellt werden.

4.6 Gesteigerte Anforderungen an weitgehend autonome ABS

Die im Modellversuch entwickelten Prüfkriterien sind vom Ansatz für eine ganze Bandbreite von Systemen mit unterschiedlichen Autonomiegraden ausgelegt. Erprobt wurden sie im Rahmen des Modellversuchs aber an einem „nur“ assistierenden System. Für weitgehend autonome Systeme, bei denen die Jugendschutzbeauftragten im Laufe eines Programmtags weniger als 20 Prozent der Entscheidungen kontrollieren – das ABS also in den überwiegenden Fällen vollständig automatisiert Altersbewertungen vornimmt, die von dem Anbieter ungeprüft übernommen werden –, verschieben sich die Prüfanforderungen: Das Kompensationsargument, dass bei einem assistierenden System etwa Teilschwächen oder Dokumentationslücken der technischen Säule durch die organisatorische Einbettung aufgefangen werden können, trägt in diesen Fällen nicht mehr. Maßnahmen in Säule 3 als Sicherheitsnetz für automatisierte Entscheidungen treten hier zurück. Das System muss seine „Fachkunde“ vielmehr aus sich heraus belegen: An die Stelle einer Plausibilisierung der Erkennungsqualität durch abstrakte architekturbezogene Betrachtungen, Produktivbetriebserfahrung und Stichprobe müsste eine förmliche Validierung der Systemperformanz treten. Diese beinhaltete eine Evaluationsstudie gegen ein unabhängiges Benchmark-Set (s. Abschnitt 4.9), Metriken und Kennzahlen wie Recall und Präzision je Altersstufe und Format, definierte Mindestwerte für die Erkennungsraten – insbesondere bei Inhalten der Stufen 16 und 18 – sowie hinreichende Stichprobengrößen.

Im laufenden Betrieb müsste regelmäßig bzw. enger beobachtet werden, ob sich das Bewertungsverhalten über die Zeit verschiebt („model drift“). Die Schwellenwerte, ab deren Überschreiten eine Warnung ausgelöst wird, müssen dabei restriktiver gesetzt sein als bei assistierenden ABS. Die Stichprobe für solche ABS müsste größer ausfallen und gezielt auch schwierige Grenzfälle des Jugendschutzes enthalten. Auch sollten Tests regelmäßig durchlaufen werden.

Zudem müssten die Anforderungen an die Erklärbarkeit steigen. Bei assistierenden Systemen genügt die Nachvollziehbarkeit der Hinweise am Originalmaterial, weil final stets ein Mensch die Entscheidung trägt. Übernimmt ein automatisiertes System die finale Entscheidung, müssen die Erklärungen selbst entscheidungstragend sein, das heißt fallbezogene Begründungen liefern, die die maßgeblichen Faktoren entlang der Wirkungsdimensionen in einer subsumtionsfähigen, nachträglich fachlich überprüfbaren Struktur ausweisen. Da nachträglich generierte Begründungen von Sprachmodellkomponenten (Post-hoc XAI-Methoden) nicht notwendigerweise den tatsächlichen Entscheidungsweg abbilden, sollte die Erklärbarkeit des ABS auf struktureller Ebene verlangt und geprüft werden.

Schließlich müsste die Auditierbarkeit des Systems umfassend möglich sein, z.B. durch manipulationssichere, vollständige Prüfprotokolle und Testschnittstellen, über die das Prüfteam zum Stichprobenabgleich nach Säule 2 bestenfalls eigene Benchmark-Läufe und Tests durchführen kann. Hinzukommen könnten turnusmäßige externe Audits, eine deutlich kürzere Anerkennungsbefristung sowie Melde- und Korrekturpflichten bei relevanten Fehlfunktionen. In der Folge einer Anerkennung würde der Prüfung eine begleitende Systemaufsicht durch die Geschäftsstelle folgen.

4.7 Perspektiven einer gesetzlichen Verankerung von ABS-Anerkennungsverfahren

Nach geltendem Recht ist die ABS-Anerkennung ein rein selbstregulatives Angebot, das sich im Modellversuch im Kontext des Fachkundebegriffs von § 7 Abs. 5 S. 1 JMStV bewegt. Eine Anerkennung dokumentiert insoweit, dass die Fachkunde der Jugendschutzbeauftragten das eingesetzte ABS, seine Möglichkeiten und Grenzen umfasst, und kann im Aufsichtsverfahren als Indiz der Fachkunde dienen, die sich auch auf den Einsatz und die Einordnungen der ABS-Ergebnisse bezieht. Bei der Vermutung fehlender Fachkunde eines Jugendschutzbeauftragten kann die zuständige Landesmedienanstalt wegen eines möglichen Verstoßes gegen § 7 Abs. 5 JMStV ein Beanstandungsverfahren einleiten, sanktionsbewehrt ist ein Verstoß gegen das Fachkundeerfordernis aber derzeit nicht.

Von der Privilegierung des § 20 Abs. 3 bzw. Abs. 5 JMStV kann das derzeitige Verfahren nicht profitieren. Die Erstreckung einer Privilegierung auf Entscheidungen eines anerkannten ABS wäre auch mehr als eine Analogie: Sie würde das inhaltsbezogene Vertrauen in eine tatsächlich erfolgte Fachprüfung durch ein system- und verfahrensbezogenes Vertrauen in eine überprüfte Entscheidungsarchitektur ersetzen. Das aber bedeutete eine signifikante Privilegierungserweiterung, die einer eigenständigen Legitimation bedürfte. Einen dogmatischen Anker dafür bietet der Staatsvertrag allerdings bereits: Bei nichtvorlagefähigen Sendungen im Rundfunkbereich greift die beschränkte Aufsicht, obwohl der Einzelinhalt vor Ausstrahlung nicht geprüft werden konnte. Getragen wird die Privilegierung dann von der institutionellen Anbindung des Anbieters und der nachgelagerten Befassung der Selbstkontrolleinrichtung (§ 20 Abs. 3 Satz 3 JMStV) mit der betreffenden Sendung. Diese Struktur aus Systemvertrauen ex ante und gesicherter Ex-post-Befassung könnte durch die qualitative Stichprobe in Säule 2 und einen verstetigten Abgleich mit der Spruchpraxis funktional vergleichbar gebildet werden. Die rechtssystematische Konsequenz aus einer solchen Analogie aber wäre die Gleichsetzung der Privilegierung nachgelagerter menschlicher Inhaltsprüfungen durch eine Selbstkontrolle mit der vorherigen automatisierten Jugendschutzbewertung durch ein ABS und einen Jugendschutzbeauftragten. Eine derartige Vorverlagerung der Absicherung der JMStV-Orientierung in eine technik- und governancebezogene

Prüfung durch die Selbstkontrolle erscheint mit Blick auf die bisherige Ko-Regulierungssystematik sehr weit. Wünschenswert(er) wäre daher eine ausdrückliche Aufnahme von ABS-Prüfverfahren in den JMStV mit dem Ziel einer rechtssystematisch passenden, klaren Beschreibung der Rechtsfolgen einer ABS-Anerkennung.

Für eine solche gesetzliche Ausgestaltung kommen aus Sicht der Pilotgruppe zwei Ansätze mit aufsteigender Regelungstiefe in Betracht. Ein zurückhaltendes Modell würde die Möglichkeit der ABS-Anerkennung durch eine Selbstkontrolle gesetzlich vorsehen, ohne aber Rechtsfolgen daran zu knüpfen. Ein neuer Tatbestand – in Anlehnung an die Anerkennung von Jugendschutzprogrammen nach § 11 JMStV – definierte automatisierte Bewertungssysteme, stellte Anbietern frei, ein eigenes oder lizenziertes ABS zur Eignungsprüfung vorzulegen, wiese die Anerkennungskompetenz den nach § 19 JMStV anerkannten Selbstkontrolleinrichtungen zu und legte die wesentlichen Voraussetzungen fest (Ausrichtung an Schutzzielen und Spruchpraxis, Kohärenzstichprobe, menschliche Letztentscheidung oder äquivalente Sicherungen, Befristung, Konfigurationsbindung, Auflagen-, Nachprüfungs- und Widerrufsregime sowie Berichtspflichten gegenüber der KJM). Die Norm würde die Möglichkeit der Anerkennungsprüfung und ihre Maßstäbe gesetzlich fixieren. Über die Regelung der Zuständigkeit hinaus aber hätten Anerkennungsentscheidungen keine unmittelbaren Rechtsfolgen. Das schmälerte den Anreiz für Anbieter, das eigenen ABS anerkennen zu lassen, erheblich.

Ein anderer Ansatz wären zwei an § 7 JMStV angedockte Vermutungsregeln: Die Anerkennung eines ABS könnte als formeller Nachweis der ABS-bezogenen Fachkunde eines Jugendschutzbeauftragten dienen. Zusammen mit einer – ebenfalls noch nicht existierenden – und an die Zertifizierung aus § 14a Abs. 1 Nr. 3 JuSchG angelehnten Regelung des Nachweises der „übrigen“ Fachkunde und der Einführung einer Ordnungswidrigkeitennorm bei Verstößen gegen § 7 Abs. 5 JMStV entstünde ein relevanter Anreiz für Anbieter, ein ABS überprüfen zu lassen. Als weiterer Anreiz könnte sich anbieten, eine weitere Vermutungsregel aufzunehmen: Wer ein anerkanntes ABS in der anerkannten Konfiguration einsetzt und die dokumentierte Entscheidungsarchitektur beachtet, für den würde widerleglich vermutet, dass er seine jugendschutzrechtlichen Organisations- und Sorgfaltspflichten erfüllt, vor allem mit Wirkung auf der Ebene der Vorwerfbarkeit bei Verstößen gegen Altersbewertungspflichten. Anbieter, die das ABS-Anerkennungsverfahren erfolgreich durchlaufen haben, könnten sich bei einer Falschbewertung auf die Anerkennungswirkung berufen und entsprechend exkulpieren, insbesondere mit Blick auf § 24 Abs. 1 Nr. 4 JMStV. Ähnlich der Situation bei nicht-vorlagefähigen Sendungen könnte in diesen Fällen eine Vorbefassungspflicht der Selbstkontrolle statuiert werden.

Insbesondere der zweite Ansatz müsste gesetzlich geregelten Leitplanken unterliegen. So folgt aus der verfassungsrechtlichen Schutzpflicht im Jugendmedienschutz ein Untermaßverbot: Die Absenkung der Aufsichtsintensität muss stets durch funktional gleichwertige Sicherungen kompensiert werden, hier gesetzlich fixierte Anerkennungsmindeststandards, Befristung, Auflagen, den Aussetzungstatbestand bei systematischer Unterbewertung und einen möglichen Widerruf. Die Wesentlichkeitslehre verlangt, dass solche Mindeststandards im JMStV selbst verankert werden müssten und nicht allein in privaten Kriterienkatalogen stehen. Kommunikationsgrundrechtlich müsste die Regelung auch das Overblocking-Problem mitberücksichtigen. Eine Regelung, die falsch-positive Bewertungen folgenlos stellt, würde überkonservative Systeme mittelfristig und systematisch zulasten der Informationsfreiheit auch jugendlicher Rezipient:innen zulassen. Die im Kriterienkatalog

angelegte Pflicht zur kritischen Reflexion konservativer Systemeinstellungen gehörte deshalb ebenfalls in den gesetzlichen Anerkennungstatbestand.

4.8 Formulierungsvorschlag für einen Anerkennungstatbestand: § 7a JMStV-E

Nicht als Forderung, sondern als Impuls für eine medienpolitische Diskussion zur besseren gesetzlichen Einbeziehung von KI-Systemen in den gesetzlichen Jugendmedienschutz unterbreitet dieser Bericht einen ausformulierten Vorschlag für die Aufnahme eines Anerkennungstatbestands für ABS in den JMStV.

Ein Tatbestand sollte dabei auf Begriffe wie „künstliche Intelligenz“, „maschinelles Lernen“ oder „Sprachmodell“ verzichten, da diese technisch altern und – wie der Auslegungsstreit um die Systemdefinition des Art. 3 Nr. 1 KI-Verordnung zeigt – eigene Abgrenzungsprobleme erzeugen. Der Formulierungsvorschlag folgt daher begrifflich einem offenen Ansatz. Maßgeblich sind die Funktion eines ABS (automatisierte Analyse, Bewertung oder Vorbereitung menschlicher Bewertungen von Angeboten) und der Einsatzkontext (unternehmensinterner Einsatz durch fachkundige Personen zur Erfüllung der Pflichten aus dem Staatsvertrag).

Dem Bestimmtheitsgebot genügt der Tatbestand durch eine gestufte Regelungsdichte: Die wesentlichen Anerkennungsvoraussetzungen, die Rechtsfolgen und die Schutzmechanismen (Befristung, Konfigurationsbindung, Aussetzung, Widerruf) stehen im Normtext selbst; quantitative Konkretisierungen wie die Schwelle der weitgehend selbständigen Bewertung – im Kriterienkatalog operationalisiert als Kontrolle von weniger als 20 Prozent der Entscheidungen – werden dagegen nicht im Staatsvertrag geklärt, sondern als gesetzlicher Typusbegriff angelegt und auf der staatlich verantworteten Richtlinien- bzw. Prüfordnungsebene konkretisiert (s. dazu Abschnitt 4.10). Bestimmtheit ist dabei gewahrt, weil die Rechtsfolgen an gesetzlich definierte Merkmale anknüpfen und die Konkretisierung publiziert, staatlich mitverantwortet und gerichtlich überprüfbar bleibt. Der folgende Vorschlag setzt diese Prinzipien um.

§ 7a JMStV-E – Anerkennung anbieterseitiger automatisierter Bewertungssysteme

(1) Automatisierte Bewertungssysteme sind Systeme, die auf der Grundlage automatisierter Informationsverarbeitung Angebote im Hinblick auf ihre Relevanz nach diesem Staatsvertrag analysieren, bewerten oder menschliche Bewertungen vorbereiten und die von Anbietern zur Erfüllung ihrer Pflichten nach diesem Staatsvertrag unternehmensintern durch fachkundige Personen eingesetzt werden. Für automatisierte Bewertungssysteme, die nach § 12 Abs. 4 von der KJM anerkannt werden, gilt diese Vorschrift nicht; die Zuständigkeiten nach § 12 Abs. 4 und 6 sowie § 16 Abs. 1 Satz 3 Nr. 7 bleiben unberührt.

(2) Eine anerkannte Einrichtung der Freiwilligen Selbstkontrolle kann ein automatisiertes Bewertungssystem gemäß Abs. 1 in einer bestimmten Konfiguration einschließlich seiner organisatorischen Einbettung zeitlich befristet anerkennen, wenn

1. das System und sein Einsatz an den Zielen dieses Staatsvertrages ausgerichtet sind,
2. die Bewertungen des Systems mit der Spruchpraxis der Einrichtung übereinstimmen und dies durch unabhängige Stichprobenvergleiche nachgewiesen ist,

3. die Zuverlässigkeit der Bewertungen in einem dem Grad der selbständigen Bewertung entsprechenden Umfang nachgewiesen ist und
4. durch fachkundige menschliche Aufsicht gewährleistet ist, dass der Anbieter seine Verantwortung nach diesem Staatsvertrag wahrnimmt.

Eine Einrichtung der Freiwilligen Selbstkontrolle darf Anerkennungen nach Satz 1 erst aussprechen, nachdem sie der KJM nachgewiesen hat, dass sie über die für die Prüfung automatisierter Bewertungssysteme erforderliche technische und methodische Kompetenz verfügt. Sie kann sich hierzu unabhängiger Sachverständiger bedienen. Der Nachweis ist der KJM auf Verlangen, mindestens jedoch alle drei Jahre, erneut zu erbringen.

(3) Die zeitliche Befristung beträgt höchstens drei Jahre, bei erstmaliger Anerkennung höchstens zwei Jahre. Die Anerkennung kann mit Nebenbestimmungen verbunden werden. Wesentliche Änderungen des Systems, seiner Einstellungen oder seiner organisatorischen Einbettung sind der Einrichtung der Freiwilligen Selbstkontrolle unverzüglich anzuzeigen. Sie können eine Nachprüfung auslösen.

(4) Die KJM überprüft Anerkennungsentscheidungen der Einrichtung nach Absatz 2 auf Überschreitung der rechtlichen Grenzen des ihr insoweit zustehenden Beurteilungsspielraums. § 19b Abs. 2 gilt entsprechend. Erklärt die KJM eine Anerkennung nach Satz 1 für unwirksam, entfällt die Wirkung nach Absatz 5 für den betroffenen Anbieter ab dem Zeitpunkt der Entscheidung der KJM.

(5) Ergeben Stichprobenvergleiche eine systematische, nicht anderweitig ausgeglichene Unterschreitung des durch diesen Staatsvertrag gebotenen Schutzniveaus durch ein anerkanntes automatisiertes Bewertungssystem, fordert die Einrichtung der Freiwilligen Selbstkontrolle den Anbieter zur Nachbesserung auf. Wird der Mangel nicht innerhalb angemessener Frist behoben, widerruft sie die Anerkennung.

(6) Weist ein Anbieter nach, dass ein Angebot unter Einsatz eines anerkannten automatisierten Bewertungssystem geprüft und bewertet wurde, ist vor Maßnahmen der KJM wegen eines behaupteten Verstoßes gegen § 5 Abs. 1 S. 1 JMStV die Einrichtung der freiwilligen Selbstkontrolle, die das automatisierte Bewertungssystem anerkannt hat, zu befassen.

(7) Die Einrichtung der freiwilligen Selbstkontrolle regelt die Einzelheiten der Anforderungen, des Verfahrens und der Stichprobenvergleiche in einer Prüfordnung. Die Prüfordnung sieht eine Vorlageregelung für die Anerkennung nach Absatz 2 und die Stichprobenvergleiche nach Absatz 4 vor, die die Größe der Stichprobe am Nutzungsumfang des anerkannten Systems bemisst. Die Auswahl der Stichproben erfolgt risikoorientiert und berücksichtigt Beschwerden sowie neuartige Angebotskategorien in einem angemessenen, in der Prüfordnung festzulegenden Mindestumfang. Die Prüfordnung und der zugrunde liegende Kriterienkatalog bedürfen der Bestätigung durch die KJM.

(8) Die Einrichtung führt ein öffentliches Verzeichnis der anerkannten automatisierten Bewertungssysteme mit Angaben zu Konfigurationsstand, Befristung und Nebenbestimmungen und berichtet der KJM jährlich über ihre Anerkennungspraxis.

(9) Die Anerkennung eines automatisierten Bewertungssystems führt zu dem widerleglichen Nachweis der Fachkunde des Jugendschutzbeauftragten gemäß § 7 Abs. 5 Satz 1. Die Fachkunde wird vermutet, solange die Zertifizierung nicht widerrufen und ihre Frist nicht abgelaufen ist.

Absatz 1 definiert funktional und technikneutral und begrenzt den Anwendungsbereich auf den unternehmensinternen Einsatz durch fachkundige Personen, wie ihn auch der Kriterienkatalog voraussetzt. Absatz 2 kodifiziert die vier Anerkennungsvoraussetzungen des Modellversuchs und schreibt in Nummer 3 den gleitenden Prüfmaßstab fest, indem der geforderte Zuverlässigkeitsnachweis mit dem Grad der selbständigen Bewertung wächst. Absatz 4 verschränkt

die Anerkennungen mit einer (beschränkten) Überprüfungsmöglichkeit durch die KJM in Anlehnung an § 19b Abs. 2 JMStV. Die Absätze 3 und 5 überführen Befristung, Konfigurationsbindung, Nachprüfungspflicht und die Aussetzungsschwelle bei systematischer, nicht kompensierter Unterbewertung in gesetzliche Schutzmechanismen. Absatz 6 enthält den Vorschlag für die Exkulpationsmöglichkeit und die Vorbefassungspflicht im Falle von behaupteten Falschbewertungen. Absatz 7 regelt die Anforderungen an zu schaffende Prüfordnungen für ABS, Absatz 8 sieht selbstkontrollbezogene Transparenzvorschriften und Berichtspflichten für abgeschlossene Anerkennungsverfahren vor. Absatz 8 stellt die Vermutungswirkung in Bezug auf die in § 7 Abs. 5 JMStV geforderte Fachkunde des Jugendschutzbeauftragten auf.

4.9 Institutionalisierung eines selbstkontrollübergreifenden Benchmark-Korpus

Die in Abschnitt 4.6 geforderte unabhängige Referenz in Form eines Benchmark-Sets bei der wiederholbaren Prüfung selbständig entscheidender ABS wirft die Frage der dauerhaften Organisation und Vorhaltmöglichkeit auf. Verfahrensbezogene ad-hoc-Stichproben, wie sie der Modellversuch aus einer Programmwoche gebildet hat, leisten einen guten Erstnachweis der Kohärenz. Sie tragen aber weder die wiederkehrenden Kohärenzprüfungen des laufenden Betriebs noch die metrische Validierung weitgehend autonomer Systeme, und sie erschweren die Vergleichbarkeit über Verfahren, Systeme und Einrichtungen hinweg. Ein gemeinsamer, selbstkontrollübergreifender Benchmark-Korpus würde Prüfkapazitäten bündeln, Bewertungsstandards vereinheitlichen und Doppelstrukturen vermeiden.

Inhaltlich müsste der Korpus den Anforderungen genügen, die Abschnitt 4.6 an die Referenzbildung stellt: Kuratierung durch als Einzelprüfer:in befugte Personen, verblindete Mehrfachannotation mit dokumentierter Interrater-Reliabilität, Beachtung unterschiedlicher Formate und Genres, Wirkungsdimension, Altersstufe und Darstellungsform sowie gezielte Anreicherung um jugendschutzrechtliche Grenzfälle. Ein existierender Korpus wäre mit der sich entwickelnden Spruchpraxis fortzuschreiben. Die Nutzungsrechte am Prüfmaterial sollten rechtssicher eingeräumt werden.

Das Gestaltungsproblem eines Benchmark-Sets ist die Rückwirkung der Messung auf das Gemessene: Sobald eine Messgröße zur Zielgröße wird, kann sie ihre Messfunktion einbüßen. Ein bekannter, statischer Korpus würde dazu einladen, vorzulegende Systeme auf den Benchmark-Test anstatt auf die allgemeine Aufgabe der Bewertung unbekannter Inhalt zu optimieren. Das kann gezielt oder unbemerkt dadurch erfolgen, dass Korpusinhalte in Trainings- oder Kalibrierungsdaten gelangen. Dagegen kann ein Bündel sich ergänzender Vorkehrungen helfen: Der Korpus müsste vollständig bei der Prüfinstanz bzw. den verfahrensführenden Selbstkontrollen verbleiben und die Tests müssten über Audit-Schnittstellen erfolgen. An die Anbieter selbst dürften keine Einzelfallergebnisse zurückfließen, sondern nur aggregierte Kennzahlen und Abweichungsmuster. Aufwändig, aber weiterführend erschiene zudem ein Rotationsdesign, das einen statischen Kern für Längsschnittvergleiche mit rollierend erneuerten Tranchen und je Prüfzyklus frisch hinzugezogenen, dem Anbieter unbekannten Inhalten kombiniert. Schließlich bedürfte es einer vertraglichen Zusicherung nebst Auditrecht, dass Korpusinhalte nicht in Trainings-, Kalibrierungs- oder Feinabstimmungsdaten einfließen.

Die durch dabei entstehende Doppelrolle der Selbstkontrolleinrichtungen als Prüfinstanz und Korpusverwalter könnte durch eine spezifische Governance-Struktur abgesichert werden, etwa durch eine gemeinsame Trägerschaft mit einem wissenschaftlichen Beirat, eine Finanzierung der Kosten über Verfahrensentgelte, Berichtspflichten gegenüber der KJM sowie ggf. einen geordneten Datenzugang für unabhängige Begleitforschung.

4.10 Satzungsförmige Fortschreibung der Prüfkriterien

Eine Anschlussfrage im Falle eines gesetzlichen Tatbestands für die Anerkennung von ABS ist, ob die Prüfkriterien als Konkretisierung der gesetzlichen Anerkennungsvoraussetzungen selbst einer stärkeren Legitimation bedürften. Der JMStV sieht die Möglichkeit vor, dass die Landesmedienanstalten übereinstimmende Satzungen und Richtlinien zur Konkretisierung des Staatsvertrags erlassen (§ 15 Abs. 2 JMStV). Diese wären dann auch von den anerkannten Selbstkontrolleinrichtungen zu beachten. Auf dieser Grundlage ließe sich die Fortschreibung der ABS-Prüfkriterien als dreistufige Normkonkretisierung organisieren: Auf der ersten Stufe legt der Staatsvertrag die wesentlichen Eckpunkte fest – Anerkennungsvoraussetzungen, Rechtsfolgen, Aussetzungs- und Widerrufstatbestände, Verfahrensrechte der Anbieter (s. Abschnitt 4.8). Auf der zweiten Stufe konkretisiert eine ABS-Richtlinie der Landesmedienanstalten die rechtsfolgenrelevanten Maßstäbe, insbesondere etwa die Schwellen des Autonomiegrads, die Mindestanforderungen an Stichprobenmethodik und Metriken oder den Auditturnus. Auf der dritten Stufe schreibt die Selbstkontrolleinrichtung ihren Kriterienkatalog mit den technisch-operativen Indikatoren, Leitfragen und Bewertungsspielräumen dynamisch fort, angezeigt bei oder bestätigt durch die KJM. Diese Arbeitsteilung könnte Innovationsoffenheit auf der operativen Ebene mit demokratischer Legitimation auf den beiden darüber liegenden Ebenen verbinden. Virulenter werden diese Überlegungen aber „erst“ mit dem zunehmenden Einsatz vollständig autonom eingesetzter ABS.

Die Richtlinienkompetenz kann so jedenfalls ein Scharnier sein, das die Weiterentwicklung des Kriterienkatalogs bzw. dessen Adaptivität an den technologischen Fortschritt mit der Gewährleistungsverantwortung des Staates kombiniert. Der Katalog bliebe als fachliches Instrument dynamisch erweiter- bzw. aktualisierbar, seine rechtsfolgenrelevanten Kernaussagen aber wandern auf staatlich verantwortete Ebenen. Die KJM sichert die Übereinstimmung der Prüfkriterien mit den gesetzlichen Vorgaben über ihr Benehmen bzw. ihre Bestätigung, über die Möglichkeit der Beanstandung und durch die Kenntnisnahme von Berichten der Selbstkontrolle.

5 Fazit und Ausblick

Der Modellversuch hat sein zentrales Ziel erreicht: Mit dem Kriterienkatalog in der Version 3.1 liegt ein strukturiertes, wiederholbares und rechtssicher handhabbares Prüfinstrument für die Anerkennung automatisierter Bewertungssysteme im Jugendmedienschutz vor, dessen Praxistauglichkeit in einem vollständigen Anerkennungsverfahren an einem produktiv eingesetzten System (*Merm:ai:d*) nachgewiesen wurde.

Als tragfähig erwiesen haben sich insbesondere drei konzeptionelle Grundentscheidungen. 1) Der *Drei-Säulen-Ansatz* nimmt das ABS als soziotechnisches Gesamtsystem in den Blick und verhindert damit sowohl eine technikfixierte Prüfung, die die organisatorische Praxis ausblendet, als auch eine bloße Prüfung der Unternehmensverfahren ohne Gewährleistung der Systemqualität. 2) Der *gleitende Prüfmaßstab* entlang des Autonomiegrads des ABS erlaubt es, sehr unterschiedliche Systemgestaltungen mit einem einheitlichen Katalog zu prüfen und die Prüftiefe dort zu konzentrieren, wo jeweils die Risiken liegen. 3) Die *Kompensationslogik* mit dreistufiger Bewertung und fristgebundenen Auflagen ermöglicht differenzierte Entscheidungen jenseits eines binären Bestehens oder Scheiterns, ohne die absolute Grenze bei systematischen, unkompensierten Unterbewertungen aufzugeben.

Das Pilotverfahren hat zugleich inhaltliche Erkenntnisse von grundsätzlicher Bedeutung erbracht. Es zeigt, dass ein aus langjähriger realer Prüfpraxis abgeleitetes System eine hohe Kohärenz mit der Spruchpraxis der Selbstkontrolle erreichen kann. Erkennungsschwächen und Systemeigenheiten machen aber eine fachkundige menschliche Kontextualisierung unverzichtbar. Es belegt zudem, dass bewusst konservative Systemeinstellungen ein legitimes Gestaltungsmittel sein können, ihr Preis aber in erhöhten Anforderungen an die kritische menschliche Befassung besteht. Der Umgang mit Overblocking-Risiken ist nicht nur eine reine Frage der Systemgüte, sondern auch eine der Entscheidungsarchitektur, in die das System eingebettet ist. Dass die Freigabepraxis des Pilotanbieters im Gegencheck erkennbar eigenständig von einigen Systemausgaben abwich, ist insoweit ein relevanter Einzelbefund des Modellversuchs, weil er die Wichtigkeit und Wirksamkeit menschlicher Letztentscheidung empirisch zeigt. Auch hat sich bestätigt, dass die Fachkunde der Jugendschutzbeauftragten ein zentraler Punkt des Regelungsmodells bleibt: Automatisierung verschiebt die Anforderungen an diese Fachkunde, ersetzt sie aber nicht; sie muss künftig auch die Kompetenz umfassen, Systemausgaben zu interpretieren, ihre Grenzen zu kennen und ihnen begründet zu widersprechen.

Für die Weiterentwicklung zeichnen sich nun mehrere Linien ab. Das Verfahren könnte zunächst in den Regelbetrieb der FSF überführt werden. Die hierfür erforderlichen Bausteine (Kriterienkatalog, Verfahrensmuster, Dokumentationsanforderungen) liegen mit den Ergebnissen des Modellversuchs vor. Der Kriterienkatalog selbst ist als transparentes, systematisches und nachvollziehbares Instrument angelegt. Er wird mit der technischen Entwicklung fortzuschreiben sein, etwa im Hinblick auf zunehmend autonome Systeme, für die die technische Säule maßgeblich werden würde; damit würden die Anforderungen an Metriken, Erklärbarkeit und Auditierbarkeit deutlich steigen.

Der Modellversuch zeigt schließlich auch, wie regulierte Selbstregulierung im JMStV auf technologische Umbrüche reagieren kann. Durch die Entwicklung überprüfbarer Qualitätsstandards, die Innovation ermöglichen und zugleich das Schutzniveau wahren, kann der staatliche Schutzauftrag im Jugendmedienschutz im Rahmen der regulierten Selbstregulierung modernisiert und nachhaltig

gestellt werden. Wenn automatisierte Systeme künftig einen wachsenden Teil der inhaltlichen Vorsortierung im Jugendmedienschutz übernehmen, entscheidet die Qualität genau dieser Anerkennungs- und Kontrollverfahren mit darüber, ob dieser Wandel das Vertrauen in die Alterskennzeichnung stärkt – oder untergräbt. Das im Modellversuch entwickelte Verfahren bietet hierfür eine belastbare Grundlage.