DAS POTENZIAL VON SPRACHMODELLEN BEI DER ERSTELLUNG VON VERSTÄNDLICHEN DATENSCHUTZERKLÄRUNGEN

Bianca Steffes / Thorsten Conrad

Bianca Steffes, Wissenschaftliche Mitarbeiterin, Lehrstuhl für Rechtsinformatik, Universität des Saarlandes
Saarland Informatics Campus C3.1, 66123 Saarbrücken, DE, bianca.steffes@uni-saarland.de

Thorsten Conrad, Wissenschaftlicher Mitarbeiter, Lehrstuhl für Rechtsinformatik, Universität des Saarlandes
Saarland Informatics Campus C3.1, 66123 Saarbrücken, DE, thorsten.conrad1@uni-saarland.de

Schlagworte: ***Datenschutzerklärung, DSGVO, LLMs, grosse Sprachmodelle, Lesbarkeit***

Abstract: *Obwohl Datenschutzerklärungen einen hohen rechtlichen Stellenwert haben, sind sie oftmals schwer verständlich. Daher ist zu untersuchen, ob dies noch den Anforderungen der DSGVO entspricht und ob Sprachmodelle Verantwortliche dabei unterstützen können, verständliche Datenschutzerklärungen zu erstellen. Exemplarisch werden zwei Sprachmodelle zur Erstellung vereinfachter Datenschutzerklärungen herangezogen und ihre generierten Texte stichprobenartig auf rechtliche Korrektheit und inhaltliche Vollständigkeit überprüft. Die Ergebnisse zeigen, dass Sprachmodelle eine Hilfe in der Erstellung lesbarer Texte sein können, die generierten Texte jedoch noch Schwächen aufweisen.*

1. Einleitung

Datenschutzerklärungen sind im Alltag allgegenwärtig. Sie begegnen Betroffenen häufig auf Websites und sollen über die Verarbeitung von personenbezogenen Daten informieren. Dennoch belegen Umfragen unter Verbrauchern, dass die vorhandenen Datenschutzerklärungen zu lang und für viele unverständlich sind.[1] Verschiedene Studien im englischen Sprachraum, die sowohl vor[2] als auch nach dem Inkrafttreten der DSGVO durchgeführt wurden, zeigen durchgehend eine hohe Komplexität der Texte.[3] Auch im deutschsprachigen Raum konnten diese Ergebnisse bestätigt werden.[4] Dieses Ergebnis wirft Zweifel auf, ob diese Praxis im Einklang mit der DSGVO steht – insbesondere, wenn sich die Datenschutzerklärungen beispielsweise an Kinder richten, die lange komplexe Texte nur schwer verstehen können. Bei Erwachsenen könnte dies ebenfalls problematisch sein: Obwohl mehr als 50% aller Deutschen im Alter von 25 bis 64 Jahren einen Abschluss der Sekundarstufe II oder höher aufweisen[5], liegt die Lesekompetenz der Deutschen laut der PIAAC-Studie bei Stufe II von V[6], was ein Leseverständnis im unteren Mittelfeld der Skala darstellt. Dies wirft die Frage

[1] So die Umfrage der Verbraucherzentrale Niedersachen vom 04. Juli 2023: https://www.verbraucherzentrale-niedersachsen.de/themen/internet-telefon/datenschutz/ergebnisse-datenschutzumfrage (aufgerufen am 30. Oktober 2023).

[2] Fabian/Ermakova/Lentz, Large-Scale Readability Analysis of Privacy Policies, in: Proceedings of the International Conference on Web Intelligence, S. 8.

[3] Krumay/Klar, Readability of Privacy Policies. In: Singhal/Vaidya, (Hrsg.) Data and Applications Security and Privacy XXXIV. S. 388–399; Srinath/Wilson/Giles, Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), S. 6829–6839.

[4] Gerpott/Mikolas, Lesbarkeit von Datenschutzerklärungen grosser Internethändler in Deutschland, MMR 2021, S. 936–941.

[5] https://www.bmbf.de/SharedDocs/Downloads/de/2023/20230912-oecd-vergleichstudie-2023.pdf?__blob=publicationFile&v=1 (aufgerufen am 30. Oktober 2023).

[6] https://www.ssoar.info/ssoar/bitstream/handle/document/36068/ssoar-2013-rammstedt-Grundlegende_Kompetenzen_Erwachsener_im_internationalen.pdf?sequence=1&isAllowed=y&lnkname=ssoar-2013-rammstedt-Grundlegende_Kompetenzen_Erwachsener_im_internationalen.pdf (aufgerufen am 30. Oktober 2023).

auf, ob Sprachmodelle eine einfache Lösung darstellen können, um Datenschutzerklärungen verständlicher zu formulieren.

2. Sprache in Datenschutzerklärungen

Um festzustellen, ob Sprachmodelle Datenschutzerklärungen auch tatsächlich verständlicher oder lesbarer formulieren können, muss die Lesbarkeit eines Textes zunächst messbar gemacht werden. Zur Bestimmung der Lesbarkeit wurden im Verlauf der Zeit verschiedene Lesbarkeitsindizes erarbeitet. Während die meisten dieser Masse in der englischen Sprache entwickelt wurden (beispielsweise SMOG[7]), gibt es auch eine Reihe an Lesbarkeitsindizes für die deutsche Sprache. Tabelle 1 gibt eine Übersicht über die im folgenden betrachteten Masse und ihre Auswertung.

	LIX	G-SMOG	WSF
Wertebereich	15 – 80	4 – 15	4 – 15
Texte mit guter Lesbarkeit	< 50	< 9	< 9

Tabelle 1: Lesbarkeitsindices für die deutsche Sprache und ihre Auswertung.

LIX ist ein Lesbarkeitsmass, welches ursprünglich für die schwedische Sprache entwickelt wurde und auch auf die deutsche Sprache angewandt werden kann.[8] Es berechnet sich aus der durchschnittlichen Satzlänge und dem Prozentsatz der langen Wörter (mehr als sechs Buchstaben) eines Textes und liefert Werte im Bereich von 15 bis 80, wobei höhere Werte eine schlechtere Lesbarkeit darstellen. Texte mit einem Wert unter 40 werden als leicht, Texte mit einem Wert von 40 bis 50 als mittel, und Texte mit einem darüber liegenden Wert als schwer eingestuft.

Der auf der englischen Sprache erstellte Lesbarkeitsindex SMOG kann durch spezifische Änderungen auf die deutsche Sprache zugeschnitten werden[9] (auch **G-SMOG** genannt). Er betrachtet die Anzahl der mehrsilbigen Wörter (drei oder mehr Silben) im Verhältnis zur Anzahl der Sätze und liefert einen Wert, der in etwa den Schuljahren entspricht, die Leser für das Verständnis des Textes benötigen.[10] Ist etwa das Ziel, dass 15-Jährige einen Text verstehen können, so ist ein SMOG-Index von 8 oder geringer erstrebenswert.

Die (vierte) **Wiener Sachtextformel**[11] **(WSF)** ist dagegen ein Mass, welches konkret für die deutsche Sprache konzipiert wurde. Es basiert auf dem Verhältnis der Anzahl der Wörter zur Anzahl der Sätze im Text sowie auf dem prozentualen Anteil der Wörter, die mehr als drei Silben aufweisen. Ähnlich wie G-SMOG errechnet es einen Wert, der in etwa die absolvierten Schuljahre[12] abbildet, welche Leser für das Verständnis benötigen.

Mit den Fortschritten im Bereich des maschinellen Lernens in den letzten Jahren sind auch neue Methoden zur Bestimmung der Lesbarkeit in diesem Bereich erforscht worden. Im deutschsprachigen Raum steckt diese Forschung jedoch noch in den Kinderschuhen.[13]

---

7 Harry/Laughlin, SMOG Grading – A New Readability Formula. The Journal of Reading, 1969, S. 639–646.

8 Anderson, Analysing the Readability of English and non-English Texts in the Classroom with Lix.“ Paper presented at the Australian Reading Association Conference, Darwin, 1981.

9 Die Anpassung des Verfahrens an deutsche Schulstufen wurde von Bamberger/Vanecek, Lesen-Verstehen-Lernen-Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache. Wien: Jugend und Volk, 1984, S.78 vorgenommen.

10 Im Original arbeitet G-SMOG mit einer Stichprobe von 30 Sätzen, in dieser Arbeit wurde jedoch der gesamte Text einbezogen wie von Wild/Pissarek, Ratte Regensburger Analysetool für Texte. Version 2.0. https://www.uni-regensburg.de/sprache-literatur-kultur/germanistik-did/downloads/ratte/index.html (aufgerufen am 30. Oktober 2023) vorgeschlagen.

11 Bamberger, Erfolgreiche Leseerziehung in Theorie und Praxis, Baltmannsweiler, 2000, S. 260.

12 In den höheren Werten spricht man auch von “Schwierigkeitsstufen”, da es sich nicht mehr um “Schulstufen” handeln kann. Siehe dazu Bamberger/Vanecek, Lesen-Verstehen-Lernen-Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache. Wien: Jugend und Volk, 1984, S. 79.

13 Ein Beispiel wäre etwa Weiss/Meurers, Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference?. In: *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, Seattle, 2022, S. 141–153.

Werden die zuvor beschriebenen Masse auf deutschsprachige Datenschutzerklärungen angewandt,[14] so zeigt sich, dass die Texte weit von einer guten Lesbarkeit entfernt sind: Im Durchschnitt liegt der LIX-Index etwa bei 56 und sowohl G-SMOG (12) als auch die Wiener Sachtextformel (12) weisen auf ein hohes benötigtes Bildungsniveau hin. Die Tatsache, dass eine kindgerechte Umsetzung juristischer Informationen möglich ist, zeigt jedoch etwa die Konvention über die Rechte des Kindes der UN: Während der Originaltext[15] eine hohe Komplexität aufweist (LIX: 83, G-SMOG: 19, WSF: 19[16]) erreicht eine speziell für Kinder erstellte Version[17] eine passende Lesbarkeit (LIX: 37, G-SMOG: 7, WSF: 7).[18] Die kindgerechte Version vereinfacht dabei zwar den Sachverhalt des Aussagekerns, die Information darüber, welche Rechte das Kind im Einzelnen hat, bleibt jedoch erhalten. Derartige Umsetzungen sind jedoch nicht die Regel. Die meisten juristischen Informationen sind, wie eingangs beschrieben, schlecht lesbar und benötigen ein hohes Bildungsniveau für das Verständnis.

3. Rechtliche Einordnung

Diese schlechte Lesbarkeit wirft die Frage auf, ob derartige Datenschutzerklärungen noch den rechtlichen Anforderungen entsprechen. Wie schon beim TMG[19] dienen Datenschutzerklärungen zur Erfüllung der gesetzlichen Informationspflichten der DSGVO. Erhebt der Verantwortliche beispielsweise personenbezogene Daten bei dem Betroffenen, muss er diesem gemäss Art. 13 DSGVO bestimmte Informationen zum Zeitpunkt der Erhebung zur Verfügung stellen.[20] Dazu gehören nach Art. 13 Abs. 1 DSGVO unter anderem Namen und Kontaktdaten des Verantwortlichen, der Zweck der Verarbeitung, die Rechtsgrundlage, die berechtigten Interessen und gegebenenfalls die Empfänger der Daten.[21] Zusätzlich zu diesen Informationen muss der Verantwortliche in der Regel auch die Informationen des Art. 13 Abs. 2 DSGVO, wie beispielsweise Informationen über Speicherdauer und Betroffenenrechte, zur Verfügung stellen.[22]

3.1. Anforderungen des Art 12 DSGVO

Art. 13 DSGVO regelt allerdings nicht, welche Anforderungen die bereitgestellten Informationen erfüllen müssen. Dies regelt Art. 12 DSGVO. Gemäss Art. 12 Abs. 1 DSGVO müssen die Informationen “in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache [übermittelt werden]; dies gilt insbesondere für Informationen, die sich speziell an Kinder richten.”. Diese Anforderungen gelten nicht nur für Art. 13 DSGVO, sondern “grundsätzlich für alle in Art. 13 bis 22 sowie in Art. 34 enthaltenen Informationspflichten.”[23] Art. 12 Abs. 1 DSGVO stellt sowohl formelle als auch inhaltliche Anforderungen an bereitzustellende Informationen.[24]

---

14 Gerpott/Mikolas, Lesbarkeit von Datenschutzerklärungen grosser Internethändler in Deutschland, MMR 2021, S. 936–941.

15 https://www.unicef.de/_cae/resource/blob/194402/3828b8c72fa8129171290d21f3de9c37/d0006-kinderkonvention-neu-data.pdf (aufgerufen am 30. Oktober 2023).

16 Obwohl in der Einführung der Wiener Sachtextformeln, als auch der G-SMOG Formel ein Wertebereich von 4 bis 15 angegeben wurde, zeigt sich hier, dass auch Werte über 15 möglich sind. Auch in der initialen Vorstellung der Masse finden sich Beispiele mit Werten ausserhalb des Definitionsbereichs (S.126, S. 171), es konnte jedoch keine nähere Erläuterung zur Bedeutung dieser Werte gefunden werden. Wir gehen von einer weiterhin steigenden Komplexität aus. (Beispiele zu finden in Bamberger/Vanecek, Lesen-Verstehen-Lernen-Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache. Wien: Jugend und Volk, 1984).

17 https://www.unicef.de/_cae/resource/blob/50770/b803ba01e7ad59fc9607c893b8800ede/d0007-krk-kinderversion-illustrationen-2014-pdf-data.pdf (aufgerufen am 30. Oktober 2023).

18 Es wurden nur die Artikel 1 bis 42 der Konvention betrachtet, da die übrigen Artikel nicht gleichwertig in der kindgerechten Version enthalten sind.

19 Conrad/Dovas, in: Forgó/Helfrich/Schneider, Betrieblicher Datenschutz, Teil IX Kapitel 2 Rn. 48.

20 Kühling/Klar/Sackmann, Datenschutzrecht, Heidelberg, 2021, Rn. 622.

21 Kühling/Klar/Sackmann, Datenschutzrecht, Heidelberg, 2021, Rn. 620.

22 Kühling/Klar/Sackmann, Datenschutzrecht, Heidelberg, 2021, Rn. 621.

23 Bäcker, in: Kühling/Buchner, DS-GVO BDSG, Art. 12 DSGVO Rn. 14.

24 Quaas, in: Wolff/Brink/v. Ungern-Sternberg, BeckOK Datenschutzrecht, Art. 12 DSGVO Rn. 11.

3.1.1. Formelle und Inhaltliche Kriterien

Alle Informationen sollen gemäss Art. 12 DSGVO in präziser, transparenter, verständlicher und leicht zugänglicher Form übermittelt werden. Das Merkmal der präzisen Form bedeutet, dass eine Information hinreichend genau und von anderen Informationen abgegrenzt ist.[25] Die Informationen sind "auf einfache Formel gebracht und griffig" darzustellen.[26] Sie dürfen kein Ausmass annehmen, das den Betroffenen überfordert.[27] Das Merkmal der Transparenz bedeutet, dass der Betroffene derart informiert werden soll, dass er den Umfang und die Folgen der Verarbeitung einschätzen kann und über die konkreten Risiken der Verarbeitung aufgeklärt ist.[28] Verständlichkeit setzt voraus, dass der Betroffene den Inhalt erfassen und die Informationen über die Verarbeitung seiner personenbezogenen Daten in seine Entscheidungen einbeziehen kann.[29] Das Kriterium der leichten Zugänglichkeit bedeutet, dass der Betroffene die Informationen "mit den [ihm] zur Verfügung stehenden (technischen) Mitteln (...) erreichen kann."[30]

Die Informationen sollen in einer einfachen und klaren Sprache übermittelt werden. Einfache und klare Sprache bedeutet, dass Fachbegriffe möglichst vermieden beziehungsweise verständlich erklärt werden.[31] Komplexe, lange Sätze mit mehreren Kommata und vielen Wörtern sind möglichst zu vermeiden.[32] Insbesondere muss der Betroffene erfahren, "ob eine Datenverarbeitung stattfindet oder nicht."[33]

3.1.2. Anforderungen an Informationen für Kinder

Art. 12 Abs. 1 HS. 2 DSGVO stellt klar, dass diese aufgeführten Anforderungen insbesondere dann gelten, wenn sich die Informationen speziell an Kinder richten. Kinder im Sinne der DSGVO sind alle Personen, die das 18. Lebensalter noch nicht vollendet haben.[34] Anhand der Art und Gestaltung eines Angebots lässt sich erkennen, ob es sich an Kinder richtet.[35] Indizien sind insbesondere die Art der Sprache und die Bilder, die das Angebot enthält.[36] Bei Verarbeitung von personenbezogenen Daten von Kindern, denen die "betreffenden Risiken, Folgen und Garantien und ihre Rechte bei der Verarbeitung personenbezogener Daten möglicherweise weniger bewusst sind", bedarf es eines besonderen Schutzes.[37] Für die Informationspflichten bedeutet dies, dass die genannten Informationen in einer für Kinder verständlichen Sprache vorliegen müssen.[38] Ein gutes Beispiel für die Übersetzung von juristischen Texten in kindgerechte Sprache bildet die in Abschnitt 2 angesprochene Konvention über die Rechte des Kindes der UN.[39]

3.2. Gestalterische Vorgaben

Genaue gestalterische Vorgaben, wie die eben beschriebenen Anforderungen umzusetzen sind, enthält Art. 12 Abs. 1 DSGVO nicht. Dies stellt an die Umsetzung besondere Herausforderungen. So können umfangreiche Erklärungen den Betroffenen schon aus Zeitgründen leicht überfordern.[40] Der Verantwortliche muss genaue

---

25 Artikel 29-Gruppe, Leitlinien für Transparenz gemäss der Verordnung 2016/679, WP 260 rev. 01, S. 7.
26 Ibid.
27 Ibid.
28 Franck, in: Gola/Heckmann, DSGVO – BDSG, Art. 12 DSGVO Rn. 19.
29 Paal/Hennemann, in: Paal/Pauly, DS-GVO BDSG, Art. 12 DSGVO Rn. 30.
30 Paal/Hennemann, in: Paal/Pauly, DS-GVO BDSG, Art. 12 DSGVO Rn. 32.
31 Paal/Hennemann, in: Paal/Pauly, DS-GVO BDSG, Art. 12 DSGVO Rn. 33.
32 Franck, in: Gola/Heckmann, DSGVO – BDSG, Art. 12 DSGVO Rn. 22.
33 Pöhle/Spittka, in: Taeger/Gabel, DSGVO – BDSG – TTDSG, Art. 12 DSGVO Rn. 12.
34 Klement, in: Simitis/Hornung/Spiecker gen. Döhmann, Datenschutzrecht, Art. 8 DSGVO Rn. 13.
35 Greve, in: Sydow/Marsch, DS-GVO/BDSG, Art. 12 DSGVO Rn. 17.
36 Ibid.
37 ErwGr 38 DSGVO.
38 Erwägungsgrund 58 S. 4 der DSGVO; Heckmann/Paschke, in: Ehmann/Selmayr, DS-GVO, Art. 12 Rn. 21.
39 Artikel 29-Gruppe, Leitlinien für Transparenz gemäss der Verordnung 2016/679, WP 260 rev. 01, S. 12.
40 Bäcker, in: Kühling/Buchner, DS-GVO BDSG, Art. 12 DSGVO Rn. 12.

und vollständige Informationen "mit einer anschaulichen, unkomplizierten und nachvollziehbaren Darstellung in Einklang bringen, (...)".[41] Art. 12 Abs. 1 DSGVO lässt dem Verantwortlichen einen Gestaltungsspielraum, wie er die eben beschriebenen Anforderungen konkret erfüllt.[42] Demnach lässt sich eine Verletzung von Art. 12 Abs. 1 DSGVO nur feststellen, wenn die Informationen grob ungenau oder unverständlich sind.[43]

Um eine Verletzung der Vorschrift festzustellen, muss zuerst der Massstab festgelegt werden, mit dem die Verständlichkeit zu messen ist. Aus der Vorschrift selbst wird jedoch nicht deutlich, wie verständlich die Anforderungen sein müssen und wie viele Informationen den Betroffenen überfordern.

Ein möglicher Massstab ist die Fähigkeit und das Verständnis eines durchschnittlichen Adressaten.[44] Dieser wird anhand des "typischen Angehörigen des Zielpublikums" ermittelt.[45] Gibt es unterschiedliche Adressatenkreise, hat der Verantwortliche für jeden Adressatenkreis unterschiedliche Darstellungen bereitzustellen.[46] Dieser adressatenorientierte Massstab stützt sich auf den Wortlaut des Art. 12 Abs. 1 DSGVO, der sich durch das "insbesondere" nicht nur auf Kinder bezieht, sondern auch darüber hinaus reicht.[47] Durch dieses "insbesondere" im Wortlaut wird allerdings auch deutlich, dass die "Regelungen der einfachen Sprache nicht pauschal für alle Empfänger angewendet werden müssen."[48] Vielmehr ist, um eine "intendierte Stärkung der Autonomie des Betroffenen und seiner datenschutzrechtlichen Betroffenenrechte effektiv" zu gewährleisten, auf den "durchschnittlich verständigen Betroffenen" abzustellen.[49] Für die gängigen Datenschutzerklärungen, die ein hohes Bildungsniveau erfordern, bedeutet das Folgendes: Ist der durchschnittliche Adressat eine Person mit mittlerem oder hohem Bildungsniveau, sind die Anforderung des Art. 12 DSGVO noch erfüllt. Handelt es sich bei dem Zielpublikum um Menschen, die ein niedriges oder ein sich noch entwickelndes Leseverständnis aufweisen, erfüllen die gängigen Datenschutzerklärungen nicht die Anforderungen des Art. 12 DSGVO.

Eine andere Auslegung geht von einem "weit unterdurchschnittliche[n] Maßstab" aus.[50] Demnach bestimmt sich der Adressat ebenfalls aus der "Benutzergruppe".[51] Es dürfe aber nicht an den durchschnittlichen Adressaten angeknüpft werden, sondern es muss ein "weit unterdurchschnittlicher Maßstab" angesetzt werden.[52] Diese Auslegung stützt sich insoweit auf primärrechtliche Vorschriften. So gilt Art. 8 GRCh für "jede Person" und ein "weit unterdurchschnittlicher Maßstab" würde "insoweit allen Menschen gerecht werden".[53] Die Folge dieses Massstabs wäre, dass Datenschutzerklärungen, die ein hohes Leseverständnis erfordern, grob unverständlich sind und damit gegen Art. 12 Abs. 1 DSGVO verstossen.

Im Ergebnis ist grundsätzlich der Massstab eines durchschnittlichen Adressaten vorzuziehen. Dies liegt zum einen daran, dass dieser Massstab zuverlässiger erfüllt werden kann: Wird die Datenschutzerklärung so gestaltet, dass sie von einem typischen Angehörigen des Zielpublikums verstanden werden kann, ist es möglich, dies anhand der Zielgruppe auch zu testen. Zum anderen würde ein unterdurchschnittlicher Massstab zu erheblichen Unsicherheiten führen, weil der Verantwortliche nicht wissen kann, wann dieser "weit unterdurchschnittliche Maßstab" genau erfüllt ist. Das Abweichen von der bisherigen Zielgruppe würde auch den Erfüllungsaufwand erhöhen, weil zur Feststellung nicht nur der durchschnittliche Adressat ermittelt werden

41 Schneider/Schwartmann, in: Schwartmann/Jasper/Thüsing/Kugelmann, DSGVO/BDSG, Art. 12 DSGVO Rn. 23.
42 Bäcker, in: Kühling/Buchner, DS-GVO BDSG, Art. 12 DSGVO Rn. 12.
43 Ibid.
44 Bäcker, in: Kühling/Buchner, DS-GVO BDSG, Art. 12 DSGVO Rn. 11.
45 Wörtlich Artikel 29-Gruppe, Leitlinien für Transparenz gemäss der Verordnung 2016/679, WP 260 rev. 01, S. 8; Greve, in: Sydow/Marsch, DS-GVO/BDSG, Art. 12 DSGVO Rn. 12.
46 Bäcker, in: Kühling/Buchner, DS-GVO BDSG, Art. 12 DSGVO Rn. 11.
47 Quaas, in: Wolff/Brink/v. Ungern-Sternberg, BeckOK Datenschutzrecht, Art. 12 DSGVO Rn. 11.
48 Franck in Gola/Heckmann, DS-GVO – BDSG, Art. 12 DSGVO Rn. 22.
49 Greve, in: Sydow/Marsch, DS-GVO/BDSG, Art. 12 DSGVO Rn. 12.
50 Heckmann/Paschke, in: Ehmann/Selmayr, DS-GVO, Art. 12 Rn. 13.
51 Ibid.
52 Ibid.
53 Ibid.

müsste, sondern auch der "weit unterdurchschnittliche". Dies würde zu einem gesteigerten Erfüllungsaufwand führen, der einer effizienten datenschutzrechtlichen Gestaltung entgegensteht.

Wird dieser Massstab nun auf die gängigen Datenschutzerklärungen angewendet, die den durchschnittlichen Deutschen als Durchschnittsadressaten voraussetzen, ergibt sich Folgendes: Der durchschnittliche Betroffene in Deutschland hat, wie die oben erläuterten Ergebnisse der PIAAC-Studie zeigen, ein niedriges bis mittleres Leseverständnis. Die gängigen Datenschutzerklärungen, die ein hohes Leseverständnis erfordern, sind daher bereits für den durchschnittlichen Deutschen unverständlich bis grob unverständlich. Im Einzelfall dürften Datenschutzerklärungen, die ein hohes Leseverständnis erfordern, daher nicht mehr den Anforderungen der DSGVO entsprechen. Dem Verantwortlichen ist folglich nahezulegen, seine Datenschutzerklärungen verständlicher zu gestalten. Grosse Sprachmodelle könnten den Verantwortlichen dabei unterstützen.

4. Nutzung von Sprachmodellen zur Erstellung von lesbaren Datenschutzerklärungen

Um das Potential von Sprachmodellen für die Vereinfachung von vorhandenen Datenschutzerklärungen zu untersuchen, müssen die von Sprachmodellen erzeugten Texte analysiert werden. Die zu vereinfachenden Inhalte sollen dabei in einfacher Sprache möglichst ohne komplizierte Sätze und Fachwörter formuliert werden. Dies deckt sich auch mit den in Abschnitt 2 eingeführten Lesbarkeitsindizes. Der Inhalt der Informationspflicht darf jedoch nicht verloren gehen. Im Folgenden soll nun untersucht werden, ob Kinder die erzeugten Datenschutzerklärungen verstehen können. Dies basiert darauf, dass die DSGVO diesen speziellen Massstab in Art. 12 DSGVO hervorhebt und bereits auf Erkenntnisse der Forschung über das Leseverständnis von Kindern zurückgegriffen werden kann. Vorliegend wird sich auf Kinder im Alter zwischen 13 und 16 Jahren konzentriert.[54] Für die Untersuchung werden daher im Folgenden zwei Beispieldatensätze zuerst auf ihre Lesbarkeit untersucht und im Anschluss daran mit Hilfe zweier Sprachmodelle vereinfacht. Für die Analyse der Lesbarkeit wurden jeweils die Überschriften entfernt, da diese oftmals keine vollständigen Sätze ergeben.

Datensatz A: Datenschutzerklärungen bekannter Websites

Im ersten Datensatz wurden die Datenschutzerklärungen der 20 meistbesuchten Websites in Deutschland[55] betrachtet. Sie umfassen auch für Kinder und Jugendliche besonders relevante Websites wie etwa Seiten von Google, Amazon, Spotify oder auch Valve (Steam). Daher sollten diese Datenschutzerklärungen eine adäquate Lesbarkeit für Kinder und Jugendliche aufweisen. Die verschiedenen eingangs beschriebenen Lesbarkeitsmasse bestätigen dies jedoch nicht: im Durchschnitt werden die Datenschutzerklärungen als äusserst komplizierte Texte bewertet, die eine langjährige Schulbildung bis zum Ende der Oberstufe voraussetzen (LIX: 59, G-SMOG: 12, WSF: 13). Auch die (entsprechend der Masse) am besten lesbare Datenschutzerklärung weist noch keine kinderfreundlichen Texte auf (LIX: 51, G-SMOG: 9, WSF: 11).

Datensatz B: Muster für Datenschutzerklärungen

Der zweite Datensatz umfasst fünf Muster, die Unternehmen zur Erstellung ihrer Datenschutzerklärungen nutzen können. Diese Muster sind über eine weit verbreitete juristische Datenbank erreichbar. Dadurch besteht die Möglichkeit, dass sie oft als Vorlage genutzt werden, weil Verantwortliche davon ausgehen, dass die Muster den Anforderungen des Datenschutzrechts gerecht werden. Betrachtet wurden hier Muster für

[54] Diese Altersgrenzen werden in Art. 8 DSGVO genannt. Zur Einwilligungsfähigkeit von Minderjährigen: Klement, in: Simitis/Hornung/Spiecker gen. Döhmann, Datenschutzrecht, Art. 8 Rn. 1 ff.

[55] https://de.statista.com/statistik/daten/studie/180570/umfrage/meistbesuchte-websites-in-deutschland-nach-anzahl-der-besucher/ (aufgerufen am 30. Oktober 2023).

Websites,[56] mobile Apps,[57] Videoüberwachung[58] und eine Social-Media-Präsenz.[59] Wird die Lesbarkeit der Texte[60] analysiert, so ergeben sich tatsächlich bessere Werte als im Datensatz A. Die Verbesserungen können bspw. darauf zurückgeführt werden, dass die Texte insgesamt kompakter sind und kürzere Sätze enthalten. Doch auch in diesem Datensatz kann noch nicht von einer guten Lesbarkeit die Rede sein (LIX: 51, G-SMOG: 10, WSF: 11).

Im Folgenden werden zwei verschiedene Sprachmodelle unterschiedlicher Komplexität für die Vereinfachung dieser Datenschutzerklärungen betrachtet. Zum einen wurde dazu GPT-3.5 Turbo (im Folgenden nur GPT) von OpenAI genutzt.[61] Mit mehr als 175 Mrd. Parametern[62] kann dieses Modell als komplexes Modell angesehen werden. Zum anderen wurde ein LLama-basiertes Modell genutzt, welches für deutsche Texte optimiert wurde[63] (aus der "EM German"-Reihe, im Folgenden nur EM). Mit nur sieben Mrd. Parametern kann dieses Modell als vergleichsweise weniger komplex angesehen werden. Während GPT in der Analyse bereits im Zero-Shot-Verfahren[64] vielversprechende Ergebnisse liefert, generiert das EM-Modell nur in der Few-Shot-Variante annehmbare Ergebnisse.[65] Die Auswertung sowohl der originalen als auch der generierten Texte mithilfe der Lesbarkeitsmasse ist in Abbildung 1 zu sehen. Beide Modelle wurden mit demselben Prompt instruiert ('Fasse den folgenden Text für Kinder zusammen. {text}').

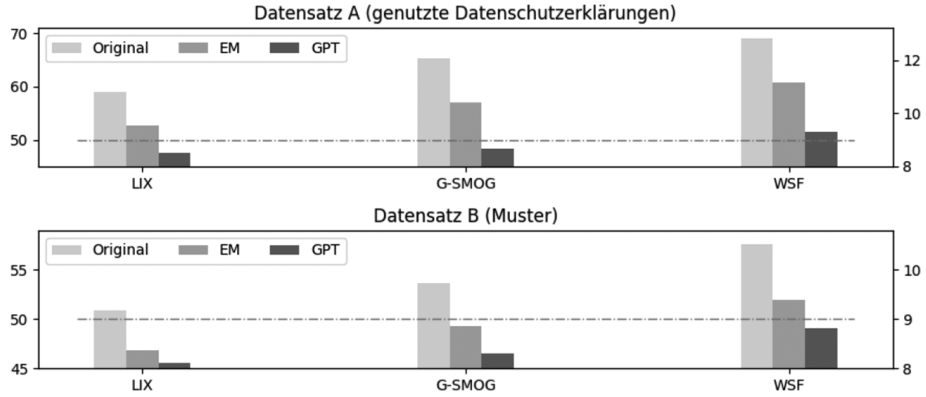

Abbildung 1: Ergebnisse der Lesbarkeitsmasse

Es zeigt sich, dass beide Modelle eine Verbesserung der Lesbarkeit bewirken, wobei das komplexere Modell (GPT) laut den Lesbarkeitsmassen die besseren Ergebnisse liefert. Auf den komplexeren Texten von Datensatz A fällt es beiden Modellen schwerer, vereinfachte Textversionen zu erstellen, als auf den einfacheren Tex-

[56] Lachenmann, in: Koreng/Lachenmann, Formularhandbuch Datenschutzrecht, F. I. 1; Missling, in: Weitnauer/Mueller-Stöfen, Beck'sches Formularbuch IT-Recht, H, 1.

[57] Lachenmann, in: Koreng/Lachenmann, Formularhandbuch Datenschutzrecht, F. I. 2.

[58] Thüsing/Pötter, in: Thüsing, Beschäftigtendatenschutz und Compliance, § 11, IV.

[59] Schubert, in: Beck´sche Online-Formulare IT- und Datenrecht, 2. 17.

[60] Einige Texte beinhalteten Platzhalter, etwa für den Namen des Unternehmens. Diese Platzhalter wurden vor der Analyse mit Beispieldaten gefüllt.

[61] https://platform.openai.com/docs/models/gpt-3-5 (aufgerufen am 30. Oktober 2023).

[62] Brown/Mann/Ryder/Subbiah et. al, Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Curran Associates Inc., Red Hook, 2020, Article 159, S. 1877–1901.

[63] https://huggingface.co/jphme/em_german_13b_v01_gptq (aufgerufen am 30. Oktober 2023).

[64] Zero-Shot bedeutet, dass das Modell nicht auf das Problem angepasst wurde und keine Beispieldaten für das konkrete Problem erhalten hat.

[65] Few-Shot bedeutet, dass dem Modell zur Lösung des Problems Beispiele von Anfragen und korrekten Lösungen mitgegeben werden. So lernt das Modell, wie das Problem zu lösen ist. In unserem Fall wurden dem Modell die Artikel 4, 5, 25 und 30 der UN-Kinderrechtskonvention als Beispiele mitgegeben (LIX: 35, G-SMOG: 7, WSF: 7).

ten von Datensatz B. EM bleibt dabei stets über der Grenze von für Kinder gut lesbaren Texten (gestrichelte Linie in der Abbildung). Das bedeutet, dass sie nicht so gut lesbar sind, wie es für Kinder zu empfehlen ist. Für Datensatz B schaffen es dagegen beide Modelle, Texte im Bereich der guten Lesbarkeit zu generieren, wobei die Texte von EM nahe an der Schwelle zur guten Lesbarkeit liegen.

Werden die generierten Texte stichprobenartig betrachtet, zeigen sich jedoch einige Probleme, die in den folgenden drei Beispielen exemplarisch aufgezeigt werden.

	Rechtsgrundlage für die Verarbeitung der Daten nach Anmeldung zum Aboservice durch den Nutzer ist Art. 6 Abs. 1 Satz 1 lit. a und b DSGVO bzw. § 25 Abs. 1 TTDSG. (LIX: 23, G-SMOG: 4, WSF: 5)
GPT	Der Text sagt uns, dass es Regeln gibt, denen wir folgen müssen, wenn wir uns für einen Aboservice anmelden. Es gibt Gesetze, die festlegen, wie unsere Daten verwendet werden dürfen. Es ist wichtig, dass wir diese Regeln einhalten, um unsere Daten zu schützen. (LIX: 33, G-SMOG: 7, WSF: 7)
EM	Nach der Anmeldung zum Aboservice verarbeitet der Dienst Ihre Daten. Die Grundlage dafür ist Art. 6 DSGVO. Das bedeutet, dass Sie die Daten freiwillig zur Verfügung gestellt haben. Ausserdem wird die Datenverarbeitung auch durch das Telemediengesetz geregelt. (LIX: 42, G-SMOG: 7, WSF: 9)

Tabelle 2: Beispiel zur Rechtsgrundlage der Verarbeitung.[66]

Die in Tabelle 2 dargestellten Texte erscheinen zunächst gut verständlich. Näher betrachtet stimmt der Inhalt jedoch nicht mit dem Ursprungstext überein und ist rechtlich fehlerhaft. Insbesondere die inhaltliche Bedeutung der Rechtsnormen, also die Verarbeitung aufgrund der Einwilligung (Art. 6 Abs. 1 Satz 1 lit. a) und "zur Erfüllung eines Vertrages" (Art. 6 Abs. 1 Satz 1 lit. b), wird nur oberflächlich ohne die Kernaussagen bzw. verzerrt wiedergegeben. Zudem ist laut den Lesbarkeitsmassen die Lesbarkeit der generierten Texte in diesem Beispiel ausnahmsweise schlechter als die des Originals, obgleich sie immer noch im Bereich der gut lesbaren Texte liegt.

	Die Daten werden gelöscht, sobald sie für die Erreichung des Zweckes ihrer Erhebung nicht mehr erforderlich sind. Im Falle der Erfassung der Daten zur Bereitstellung der Website ist dies der Fall, wenn die jeweilige Sitzung beendet ist. Im Falle der Speicherung der Daten in Logfiles ist dies nach spätestens 30 Tagen der Fall. Eine darüberhinausgehende Speicherung ist möglich. In diesem Fall werden die IP-Adressen der Nutzer gelöscht oder verfremdet, sodass eine Zuordnung des aufrufenden Clients nicht mehr möglich ist. (LIX: 42, G-SMOG: 7, WSF: 8)
GPT	Der Text sagt, dass die Daten gelöscht werden, sobald sie nicht mehr gebraucht werden. Zum Beispiel werden die Daten einer Internetseite gelöscht, wenn man die Seite nicht mehr benutzt. Wenn Daten in Logfiles gespeichert werden, werden sie spätestens nach 30 Tagen gelöscht. Es kann aber sein, dass die Daten noch länger gespeichert werden. In diesem Fall werden die IP-Adressen der Nutzer gelöscht oder verändert, sodass man nicht mehr sehen kann, wer die Seite besucht hat. (LIX: 46, G-SMOG: 9, WSF: 9)
EM	Die Daten werden nur gespeichert, wenn sie für den Betrieb der Webseite benötigt werden. Die Daten werden nach spätestens 30 Tagen gelöscht. Eine längere Speicherung ist möglich, dann werden die IP-Adressen der Nutzer aber direkt anonymisiert, sodass sie nicht mehr zurückverfolgt werden können. (LIX: 47, G-SMOG: 7, WSF: 8)

Tabelle 3: Beispiel zur Dauer der Speicherung.[67]

66 Entnommen aus: https://www.dsgv.de/datenschutz.html (aufgerufen am 30. Oktober 2023).
67 Entnommen aus: https://www.dsgv.de/datenschutz.html (aufgerufen am 30. Oktober 2023).

Die Umformulierungen in Tabelle 3 sind ein positives Beispiel. Der Text der beiden Modelle stimmt inhaltlich weitgehend mit dem Ursprungstext überein. Allerdings ist der zweite Abschnitt im EM-Modell zu stark vereinfacht und die Löschung inhaltlich komplett weggefallen.

	Sie haben unter den in Art. 20 Abs. 1 DSGVO genannten Voraussetzungen das Recht, die Sie betreffenden personenbezogenen Daten, die Sie uns bereitgestellt haben, in einem strukturierten, gängigen und maschinenlesbaren Format zu erhalten, und das Recht, diese Daten einem anderen Verantwortlichen ohne Behinderung durch uns zu übermitteln. Bei der Ausübung des Rechts auf Datenübertragbarkeit haben Sie das Recht, zu erwirken, dass die personenbezogenen Daten direkt von uns an eine andere verantwortliche Stelle übermittelt werden, soweit dies technisch machbar ist. (LIX: 53, G-SMOG: 12, WSF: 12)
GPT	Wenn du bestimmte Informationen, die du uns gegeben hast, gerne an eine andere Person oder Firma übertragen möchtest, gibt es bestimmte Bedingungen, unter denen du das Recht dazu hast. In diesem Fall können die Informationen in einem Format übertragen werden, das von Maschinen leicht gelesen werden kann. Wir müssen dir dabei helfen und deine Anfrage so gut wie möglich erfüllen, wenn es technisch möglich ist. (LIX: 46, G-SMOG: 9, WSF: 10)
EM	Jeder hat das Recht, seine persönlichen Daten in einem Format zu bekommen, das Computer verstehen. Es ist wichtig, dass diese Daten nicht einfach an andere weitergegeben werden können. (LIX: 39, G-SMOG: 7, WSF: 8)

Tabelle 4: Beispiel zur Datenübertragbarkeit.[68]

Auch die Ergebnisse in Tabelle 4 sind auf den ersten Blick verständlicher als der Ursprungstext. Insbesondere die Umsetzung des in Art. 20 DSGVO enthaltenen Inhaltes findet jedoch gar nicht, bzw. nur unzureichend statt. Zudem werden hier Nebeninformationen wie die Info, "dass diese Daten nicht einfach an andere weitergegeben werden können" hinzugefügt, die kaum etwas mit dem ursprünglichen Text gemeinsam haben.

Nach der Betrachtung einer Reihe weiterer zufälliger Beispiele bleibt festzuhalten, dass beide Modelle die Texte oftmals derart vereinfachen, dass einige Informationen komplett entfernt werden. Vor allem die Rechtsnormen finden selten ihren Weg in die generierten Texte. Naheliegend wäre, dass dies eine Folge des Prompts sein könnte ('Fasse [...] zusammen'), andere Prompts etwa mit Verben wie 'erläutere', 'erkläre' oder 'vereinfache' lieferten jedoch keine besseren Ergebnisse. Zudem kann es vorkommen, dass kleinere Nebeninformationen hinzugefügt wurden, die nicht im Originaltext stehen. Die beiden Modelle unterscheiden sich darin, dass GPT tendenziell dazu neigt, von den Originaltexten zu abstrahieren. EM dagegen führt eher ein Vereinfachen des gegebenen Textes durch; auch die Normen wurden hier oftmals in den generierten Text übernommen. Hier treten jedoch häufiger gravierende Fehler auf: manchmal haben die generierten Texte keinen Bezug zum Original oder es werden verhältnismässig viele neue Informationen hinzugefügt oder gar eine falsche Information wiedergegeben.

5. Fazit

In der kritischen Auseinandersetzung mit den Möglichkeiten von grossen Sprachmodellen für die Erstellung verständlicher Datenschutzerklärungen wurden einerseits die rechtlichen Rahmenbedingungen aufgezeigt und andererseits zwei Sprachmodelle exemplarisch auf die Lösung dieser Aufgabe überprüft. Beide Modelle zeigen jedoch, dass die generierten Texte, auch wenn sie auf den ersten Blick vielversprechend erscheinen, gravierende inhaltliche Schwächen haben. So abstrahieren die Modelle oftmals so stark, dass wichtige Detailinformationen verloren gehen, anstatt simplifizierte Erläuterungen zu geben. Für eine sinnvolle Anwendung

[68] Entnommen aus: https://www.burda.com/de/datenschutz/ (aufgerufen am 30. Oktober 2023).

von grossen Sprachmodellen zur Erstellung von verständlichen Datenschutzerklärung ist daher zu empfehlen, die generierten Texte stets mit dem Originaltext abzugleichen und zu prüfen, ob die erstellten Texte den gewünschten Ergebnissen entsprechen.

Während diese Studie einen ersten Einblick in die Möglichkeiten von Sprachmodellen in der Erstellung von verständlichen Datenschutzerklärungen gibt, sind noch viele Forschungsfragen offen. Zum einen betrachtet diese Studie nur eine sehr begrenzte Menge an Texten. In zukünftigen Arbeiten könnte daher ein grösserer Textkorpus betrachtet werden. Zudem könnten anstatt der Beispiele aus der Konvention für Kinderrechte massgeschneiderte Beispiele für Datenschutzerklärungen erstellt werden, die den Modellen im Few-Shot-Verfahren helfen können. Auch das Erstellen eines für dieses konkrete Problem spezialisierten Sprachmodells ist denkbar. Ein Aspekt der Verständlichkeit, der in diesem Beitrag nicht betrachtet wurde, ist die visuelle Darstellung der Datenschutzerklärungen. Viele Datenschutzerklärungen enthalten Bilder oder Videos, welche die Inhalte des Textes verständlicher machen sollen.[69] Ein interessanter Aspekt für zukünftige Arbeiten wäre daher die Erfassung der Verständlichkeit derartiger Darstellung und die Erforschung der Möglichkeiten bildgenerierender Modelle wie etwa Midjourney.[70]

6. Danksagung

Diese Arbeit ist im Kontext der durch das BMBF geförderten Projekte A-DigiKomp (16SVS8544) und D'accord (16KIS1510) entstanden.

[69] Ein Beispiel ist etwa die Datenschutzerklärung von Google: https://policies.google.com/privacy (aufgerufen am 30. Oktober 2023).
[70] https://docs.midjourney.com/ (aufgerufen am 30. Oktober 2023).