Engel, Christoph

Working Paper

Gewichtsformel - wörtlich genommen: Ein empirischer Test mit der Hilfe eines Sprachmodells

Discussion Papers of the Max Planck Institute for Research on Collective Goods, No. 2024/10

Provided in Cooperation with:
Max Planck Institute for Research on Collective Goods

Suggested Citation: Engel, Christoph (2024) : Gewichtsformel - wörtlich genommen: Ein empirischer Test mit der Hilfe eines Sprachmodells, Discussion Papers of the Max Planck Institute for Research on Collective Goods, No. 2024/10, Max Planck Institute for Research on Collective Goods, Bonn, https://hdl.handle.net/21.11116/0000-000F-048E-D

This Version is available at:
https://hdl.handle.net/10419/301228

MAX PLANCK INSTITUTE
FOR RESEARCH ON
COLLECTIVE GOODS

CHRISTOPH ENGEL

Discussion Paper
2024/10

GEWICHTSFORMEL – WÖRTLICH GENOMMEN. EIN EMPIRISCHER TEST MIT DER HILFE EINES SPRACHMODELLS

I. Forschungsfrage

Robert Alexy war stets besorgt über die Rationalität der Rechtsanwendung. Spielräume sind ein Dorn in seinem Auge. Für richterliche Intuition ist in seinem Denken eigentlich kein Platz. Sie genügt weder dem rechtsstaatlichen Gebot der Klarheit und Vorhersehbarkeit richterlicher Entscheidungen, noch dem demokratischen Gebot der Bindung des Richters an das Gesetz[1]. Ich teile diese Überzeugungen nicht: aus einem epistemischen und aus einem normativen Grund.

Viele Fälle sind nicht wohldefiniert. Der Richter muss entscheiden, obwohl er weiß, dass er nicht genug weiß[2]. Er soll sich auch nicht regelmäßig auf Beweislast zurückziehen[3]. Dann bleibt nur der Rückzug auf Intuition[4]. Natürlich kann Intuition fehlgehen. Der Richter kann sich Neutralität vormachen[5]. Aber Intuition ist außerordentlich leistungsfähig. Sie kann nicht nur ungeheure Mengen an Information verarbeiten. Sie kann auch sehr gut mit bloßen Eindrücken zurechtkommen[6]. Die könnte man zwar als Netzwerke von bedingten Wahrscheinlichkeiten rekonstruieren[7]. Mit der Entscheidung an Hand Bayesianischer Netze wären Richter jedoch nicht nur technisch überfordert. Solche Netzwerke haben auch so viele Freiheitsgrade, dass am Ende doch nur Scheinrationalität entstünde.

Richterliche Entscheidungen sind oft außerdem deshalb schwer, weil normative Prinzipien miteinander im Widerspruch stehen, die sich nicht auf eine einzige normative Grundlage zurückführen lassen[8]. Man mag für Effizienz sein oder für Gleichheit. Man mag die Würde des einzelnen Menschen über das Wohl der Allgemeinheit stellen, um nur die offensichtlichsten Beispiele zu nennen. Andere scheinbar konsensfähige normative Prinzipien leiden an Unbestimmtheit. Das gilt vor allem für Fairness[9]. Was soll fair sein: das Ergebnis, oder der Prozess, in dem es gefunden wird? Wann ist ein Ergebnis fair: wenn alle gleich behandelt werden, oder alle nach ihrer Leistung, oder alle nach ihrer Bedürftigkeit, oder alle nach ihrem legitim erworbenen Status?

---

1 Eindrucksvoll Alexy: Verfassungsrecht und einfaches Recht – Verfassungsgerichtsbarkeit und Fachgerichtsbarkeit, in: VVDStRL 61 (2002) 7-33.

2 Ich habe mich in zwei größer angelegten gemeinschaftlichen Anstrengungen mit den Folgen für die Rechtsanwendung auseinandergesetzt: Gerd Gigerenzer / Christoph Engel (Hrsg.): Heuristics and the Law, MIT Press 2006; Christoph Engel / Wolf Singer (Hrsg.): Better Than Conscious? Decision Making, the Human Mind, and Implications For Institutions, MIT Press 2008.

3 Falls die Bestimmung der Beweislast nicht ohnehin dieselben Probleme aufwirft, wie vor allem im öffentlichen Recht.

4 Christoph Engel: Institutions for Intuitive Man, in Engel/Singer (FN ***) 391-412.

5 S. nur Ziva Kunda: The Case for Motivated Reasoning, PsychBull 108 (1990) 480-496.

6 Tilman Betsch/Andreas Glöckner: Intuition in judgment and decision making: Extensive thinking without effort, Psychological Inquiry 21 (2010) 439-462.

7 Mark Schweizer: Beweiswürdigung und Beweismaß, Tübingen 2015.

8 Christoph Engel: Offene Gemeinwohldefinitionen, Rechtstheorie 32 (2001) 23-52.

9 Alexander Cappelen et al.: The Pluralism of Fairness Ideals: An Experimental Approach, American Economic Review 97 (2007) 818-827.

Robert Alexy hat all diese Einwände natürlich gesehen. Er wendet ein:

> „Es geht nicht um die unmittelbare Vergleichbarkeit irgendwelcher Gegenstände, sondern um die Vergleichbarkeit ihrer Bedeutung für die Verfassung […]. Wenn ein rationaler Diskurs über das, was von Verfassungs wegen gilt, möglich ist, dann ist ein einheitlicher Standpunkt möglich. Dieser entsteht, sobald ein rationaler Diskurs beginnt, der sich von der regulativen Idee des von Verfassungs wegen Richtigen leiten läßt[10]."

In diesem Text nehme ich ihn beim Wort. Ich lasse die epistemischen und die moralphilosophischen Einwände dahinstehen und stelle mich auf Robert Alexy's Standpunkt. Er hat sich bekanntlich nicht darauf beschränkt, die Rationalisierung von Abwägungsentscheidungen einzufordern. Er hat auch ein Verfahren vorgeschlagen. Dieses Verfahren nennt er die „Gewichtsformel". In diesem Text untersuche ich die Leistungsfähigkeit der Gewichtsformel empirisch. Verändern sich die Entscheidungen, wenn der Entscheider die Formel verwendet? Verändern sie sich in dem von Robert Alexy geforderten Sinne? Führt die Anwendung der Formel zur Rationalisierung der Abwägung?

Robert Alexy hat sich vor allem an der mangelnden Determiniertheit der Grundrechtsanwendung gestoßen. Deshalb hat er seine Formel mit Blick auf die Grundrechtsdogmatik entwickelt. Konsequent verwende ich einen Grundrechtsfall, um das Verfahren zu testen.

In einer perfekten (wissenschaftlichen) Welt würde ich diesen Fall mit der Population testen, die Robert Alexy beeinflussen will, also mit Verfassungsrichtern. Das ist offensichtlich unmöglich. Ich wäre auch nicht optimistisch, dass ich eine hinreichende Zahl an ehemaligen Verfassungsrichtern zur Teilnahme gewinnen könnte. Andere Personen mit juristischer Ausbildung wären vermutlich zu gewinnen. Aber jedem Teilnehmer könnte ich bestenfalls eine Version des Falles vorlegen. Selbst im günstigsten Fall könnte ich nur die vollständige Version der Gewichtsformel mit ihrer gänzlichen Abwesenheit vergleichen. Doch seit einem guten Jahr stehen nun Sprachmodelle zur Verfügung. Für viele Aufgaben haben sie sich als erstaunlich leistungsfähig erwiesen. Das gilt auch für juristische Anwendungen. Diesen Umstand nutze ich aus.

Wenn ich dem Sprachmodell die Gewichtsformel, oder Elemente der Gewichtsformel, vorgebe, hat das einen sehr deutlichen Effekt auf die Entscheidung, der Verfassungsbeschwerde stattzugeben. Die Gewichtsformel macht transparent, wie der Entscheider die Elemente gewichtet, die abzuwägen sind. Wenn nur ein Interesse des Grundrechtsträgers im Widerstreit mit einem öffentlichen Interesse steht, sind die Entscheidungen auch ziemlich konsistent. Das Sprachmodell tut sich jedoch schwer, wenn dem Interesse des Grundrechtsträgers eine ganze Serie von öffentlichen Interessen gegenübersteht. Ich kann die Konsistenz erhöhen, wenn ich für die Aggregation das besonders verlässliche Sprachmodell GPT 4 verwende. Doch selbst dann bleiben Inkonsistenzen.

10 Robert Alexy: Die Gewichtsformel, in: Gedächtnisschrift für Jürgen Sonnenschein, Berlin 2003, 771-792 [18 f.].

Der Gesetzgeber (der Verfassungsgeber; das Bundesverfassungsgericht) könnte Konsistenz erzwingen. Er könnte dem Spruchkörper aus der Hand nehmen, wie die Elemente aggregiert werden. Er könnte mit anderen Worten die Gewichtsformel vorschreiben. Die Aufgabe des Spruchkörpers wäre dann beschränkt darauf, für den konkreten Fall zu bestimmen, welches Gewicht den konkurrierenden normativen Belangen zukommt. Inkonsistenz (im Sinne der Gewichtsformel) ist allerdings nicht zufällig verteilt. Inkonsistenz scheint nicht die Folge gedanklicher Fehler zu sein. Eher scheint das Sprachmodell die Gewichtsformel für zu rigide zu halten. Robert Alexy kann sich selbst eine Variante der Gewichtsformel mit mehr Freiheitsgraden vorstellen[11]. Die Resultate meiner Versuche mit dem Sprachmodell könnten nahelegen, dass eine etwas flexiblere Version der Gewichtsformel vorzuziehen ist.

II. Sprachmodelle

Das Recht war stets nahe an der Rechtswirklichkeit. Schließlich besteht seine gesellschaftliche Funktion ja in der Bewältigung sozialer Konflikte. Deshalb verwundert nicht, dass die Juristerei sehr neugierig auf die neuen Möglichkeiten ist, die große Sprachmodelle nach der Art von ChatGPT eröffnen[12]. Am Ende ist auch die vollständige Delegation der Entscheidungsfindung an Computer vorstellbar, jedenfalls für gut typisierbare Fallgruppen. Das wirft offensichtliche normative Probleme auf. Doch Sprachmodelle könnten schon viel früher zu nützlichen Helfern werden. Ein Anwendungsfall sind Bewertungsprobleme, bei denen die Rechtsordnung mehr oder minder offen auf soziale Wertungen verweist - wie bei der Anwendung von Grundrechten.

Große Sprachmodell sind es seit kaum mehr als einem Jahr breit verfügbar. Deshalb ist es noch zu früh für eine verlässliche Antwort auf eine wichtige Vorfrage: wie gut spiegeln die Antworten der Sprachmodelle die Antworten wider, die menschliche Teilnehmer auf die gleiche Frage gegeben hätten? Erste Untersuchungen führen zu unterschiedlichen Ergebnissen[13]: manche kognitiven Verzerrungen wiederholen sich, andere nicht[14]. Die moralischen Urteile von Sprachmodellen scheinen den moralischen Urteilen von Versuchspersonen aber zu ähneln[15].

---

11 Ebd. [25], er spricht dann von einem „doppeltriadischen" Modell.

12 Morrison, A. (2020). Artificial Intelligence in the Courtroom: Increasing or Decreasing Access to Justice? International Journal of Online Dispute Resolution, 7, 76-93; Norton, K. L. (2020). The Middle Ground. A Meaningful Balance Between the Benefits and Limitations of Artificial Intelligence to Assist with the Justice Gap. University of Miami Law Review, 75, 190-256; Poppe, E. S. T. (2019). The Future Is Complicated. AI, Apps & Access to Justice. Oklahoma Law Review, 72, 185-212; Queudot, M., Charton, É., & Meurs, M.-J. (2020). Improving access to justice with legal chatbots. Stats, 3(3), 356-375; Simshaw, D. (2022). Access to AI Justice: Avoiding an Inequitable Two-Tiered System of Legal Services. Yale Journal of Law and Technology, 24, 150-226.

13 Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. Science, 372(6547), 1209-1214.

14 Orsini, E. (2023). Do Cognitive Biases Persist in Large Language Models? Chen, Y., Andiappan, M., Jenkin, T., & Ovchinnikov, A. (2023). A Manager and an AI Walk into a Bar. Does ChatGPT Make Biased Decisions Like We Do?; Jones, E., & Steinhardt, J. (2022). Capturing failures of large language models via human cognitive biases. Advances in neural information processing systems, 35, 11785-11799.

15 Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? Trends in Cognitive Sciences, 27, 597-600; Johnson, T., & Obradovich, N. (2023). Evidence of behavior consistent with self-interest and altruism in an artificially intelligent agent. arXiv preprint arXiv:2301.02330; siehe auch

Ähnlich sind die Aussagen von Sprachmodellen auch zu den Entscheidungen in klassischen Spielen der Verhaltensökonomen[16]. Deshalb kann es im Moment nur ein Versuch sein. Aber der Versuch lohnt: bekommt der Rechtsanwender sinnvolle Hilfe, wenn er zentrale Elemente eines Abwägungsproblems einem Sprachmodell vorlegt?

Wenn sich am Ende einer Vielzahl solcher Versuche herausstellt, dass die Antworten (hinreichend) belastbar sind, wäre das ein großer Vorteil. Denn Umfragen unter menschlichen Teilnehmern sind nicht nur teuer. Sie sind auch nicht beliebig skalierbar. Man mag, wenn die Rechtsfrage hinreichend Gewicht hat, vielleicht noch 1000 Versuchspersonen befragen. Aber man könnte diesen Versuchspersonen nur einen (oder vielleicht zwei oder drei sorgsam ausgewählte Variationen eines) Fragebogen(s) vorlegen. Das Sprachmodell kann man dagegen zu sehr geringen Kosten viel häufiger fragen. Man kann deshalb auch nicht nur eine Version, oder eine Darstellung, des Abwägungsproblems testen, sondern eine Vielzahl.

Wenn es auf Quantitäten ankommt, kann man diese Parameter in nahezu beliebig feinen Schritten verändern. Wenn man die Quelle des Werturteils verstehen will, kann man einen anderen Fall ersinnen, der in der relevanten Hinsicht gleich, im Lebensbereich aber verschieden ist. Schließlich könnten Sprachmodelle auch Einblick in die Heterogenität moralischer Urteile geben. Denn man kann dem Sprachmodell mehr oder minder Varianz in den Entscheidungen erlauben. Dann sieht man als Ergebnis nicht nur die eine dominante Entscheidung, sondern sieht zugleich, mit welcher Wahrscheinlichkeit das Sprachmodell anders entschieden hätte. Diese Wahrscheinlichkeit könnte auch ein Spiegel der Tatsache sein, dass die Abwägung der konkurrierenden Belange schwierig ist.

III. Technische Umsetzung

An sich könnte man das Sprachmodell von open.ai über einen Webbrowser nutzen. Man könnte ChatGPT den Fall schildern und fragen, wie es die konkurrierenden Belange abwägt. Das Modell erlaubt auch Anfragen in deutscher Sprache. Der Weg über ChatGPT wäre aber höchst unpraktisch. Man müsste dieselbe Anfrage viele Male starten, die Antwort jeweils aus dem Browser in ein Textdatei kopieren, die Texte anschließend einzeln durchsehen und nach einem vorab festgelegten Kodierungsschema in eine (z. B. Excel-)Tabelle übersetzen. Erst dann hätte man - nach vielen Stunden - eine Datei, der man die zusammenfassende Einschätzung entnehmen kann[17]. Sollten die Antworten des Sprachmodells künftig nicht nur als Quelle

---

Jin, Z., Levine, S., Gonzalez Adauto, F., Kamal, O., Sap, M., Sachan, M., Mihalcea, R., Tenenbaum, J., & Schölkopf, B. (2022). When to make exceptions: Exploring language models as accounts of human moral judgment. Advances in neural information processing systems, 35, 28458-28473; Ma, X., Mishra, S., Beirami, A., Beutel, A., & Chen, J. (2023). Let's Do a Thought Experiment: Using Counterfactuals to Improve Moral Reasoning. arXiv preprint arXiv:2306.14308; Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. Nature Machine Intelligence, 4(3), 258-268.

16 Brookins, P., & DeBacker, J. (2023). Playing games with GPT: What can we learn about a large language model from canonical strategic games?

17 Seit kurzem ist es möglich, ChatGPT generalisierte Anweisungen zu geben. Wenn man geschickt genug vorgeht, kann man auch über die Web-Oberfläche Freiheitsgrade des Sprachmodells ausnutzen. Das geht aber nur über Text. Man muss also sicher sein, dass das Modell die Anweisung auch so verstehen wird, wie sie gedacht war.

der juristischen Inspiration dienen, sondern als empirische Evidenz in einen Prozess eingeführt werden, wäre es erforderlich, dass der Daten generierende Prozess perfekt kontrolliert und reproduzierbar ist. Beim Weg über den Browser wäre das zwar vorstellbar, aber außerordentlich aufwändig. Man müsste jede einzelne Anfrage aus dem Browser in eine separate Datei exportieren.

Aus all diesen Gründen ist sehr von Vorteil, dass open.ai seine Sprachmodelle auch über eine API[18] anbietet. Technisch wird es dann etwas aufwändiger. Man muss in der Programmiersprache Python ein kleines Programm schreiben. Dafür erhält man aber sehr viel mehr Kontrolle über den Prozess. Man kann die Freiheitsgrade des Sprachmodells streng kontrolliert nutzen. Man kann dieselbe Anfrage so oft wiederholen, wie erforderlich erscheint. Man kann das Sprachmodell dazu anhalten, nicht mit langen, schwer deutbaren Erörterungen zu antworten, sondern mit einem schlichten ja oder nein[19]. Man kann diese Antworten in eine Datei exportieren, die man mit nicht allzu großem Aufwand mit einem Statistik Programm analysieren kann. Diesen Zugang nutze ich[20].

Am Markt ist eine Vielzahl von Sprachmodellen verfügbar[21]. Aus pragmatischen Gründen habe ich mich für das Modell GPT von open.ai entschieden. Vergleichstests zeigen, dass es sehr leistungsfähig ist[22]. Weil das Modell der Marktführer ist, gibt es auch reiche Erfahrungen über die wirksamste Weise, das Modell zu verwenden.

Es gibt verschiedene Versionen des Modells. Versuche mit anderen Aufgaben haben uns gezeigt, dass GPT-4 zwar die größte Genauigkeit hat. Dieser Vorzug wird allerdings dadurch erkauft, dass das Modell mit sehr viel höherer Wahrscheinlichkeit die am ehesten richtige Antwort gibt. Diese Eigenschaft ist wünschenswert, wenn man das Modell ein einziges Mal fragt. Das Modell generiert diese Antwort jedoch durch ein Wahrscheinlichkeitsurteil. Wenn sich das Modell so gut wie sicher ist, dann ist die Konzentration auf das wahrscheinlichste Ergebnis keine sonderliche Einschränkung. Doch wenn die Wahrscheinlichkeit einer anderen Antwort hoch ist, nur eben nicht hoch genug, dann ist es wichtig, die knappe Entscheidung zu kennen.

---

18 Application Programmer Interface.

19 Dieses prompt engineering war nicht ganz einfach. Folgender system prompt hat nicht perfekt, aber doch relativ gut funktioniert (wenn es unverwertbare Antworten gibt, notiere ich das bei der Datenauswertung): "Die folgende Frage hat juristische Bedeutung. Ich frage Sie aber nicht als Juristen. Ich möchte lernen, welche Entscheidung nach Ihrer Überzeugung richtig wäre.
Zu Ihrer Information: ich bin selbst Jurist und weiß, dass die Rechtslage in diesem Fall umstritten ist. Beide Entscheidungen wären rechtmäßig und begründbar. Die juristische Entscheidung hängt am Ende an einer Wertung: wessen Interessen überwiegen? Diese Wertung ist im Kern nicht juristisch. Ich möchte von Ihnen erfahren, wie Sie diese Wertungsfrage entscheiden würden.
Bitte geben Sie keine Begründung. Antworten Sie nur mit "Ja" oder "Nein". "

20 Ich verwende GPT 3.5 turbo, temperature = 1 (um Varianz zwischen den Antworten zuzulassen), und frage jeweils 100 mal.

21 Die bekanntesten und leistungsfähigsten sind GPT von openai und Gemini von Google. Beide sind allerdings proprietär. Deshalb sind Ergebnisse nicht perfekt replizierbar. Wenn man darauf Wert legt, kann man open source Modelle wie LLaMA oder Mistral verwenden.

22 Rui Mao et al., GPTEval: A Survey on Assessments of ChatGPT and GPT-4, https://arxiv.org/pdf/2308.12488.pdf.

Ursprünglich haben Sprachmodelle diese Wahrscheinlichkeiten offen ausgewiesen. Die besonders leistungsfähigen Modelle (insbesondere GPT und Gemini) tun das jedoch nicht mehr. Die Modelle haben jedoch einen Parameter, den der Nutzer variieren kann. Wenn er „temparature“ auf 0 stellt, dann gibt das Sprachmodell (solange die Architektur des Modells und die Trainingsdaten nicht verändert worden sind) bei jeder neuen Anfrage die exakte selbe Antwort. Doch wenn der Anwender einen höheren Wert von „temperature“ wählt, dann unterscheiden sich die Antworten auf die Wiederholung derselben Anfrage. Diesen Umstand nutze ich. Ich setze „temperature“ auf den hohen Wert 1 und frage jeweils 100 Mal. Auf diese Weise erhalte ich eine ganze Verteilung von Entscheidungen. Ich kann daraus schließen, mit welcher Wahrscheinlichkeit das Sprachmodell (gegeben das wechselnde Design der Anfrage) der Verfassungsbeschwerde stattgeben würde.

Dieses Verfahren funktioniert allerdings bei dem etwas weniger leistungsfähigen Modell GPT 3.5 turbo viel besser als bei GPT 4. Deshalb verwende ich für meine Anfragen zunächst immer GPT 3.5 turbo. GPT 3.5 turbo ist allerdings nicht nur weniger akkurat. Es kommt auch besonders schlecht mit mathematischen Aufgaben zurecht. Diese Schwäche könnte dafür verantwortlich sein, dass die Aggregation über eine größere Zahl normativer Belange recht unzuverlässig zu sein scheint. Vorsorglich füge ich deshalb einen weiteren technischen Schritt hinzu. Ich frage GPT 3.5 nach der Einschätzung der konkurrierenden Belange. Ich bitte dann aber GPT 4, diese Belange zu aggregieren. Ich berichte beide Resultate parallel.

IV. Beispielsfall

Um die Leistungsfähigkeit der Gewichtsformel zu erproben, braucht man einen Beispielsfall, in dem man ernstlich über das Ergebnis der Abwägung streiten kann. Folgenden aktuellen Fall habe ich gewählt:

> Das Konsum-Cannabis-Gesetz erlaubt einer Person, die mindestens 18 Jahre alt ist, an ihrem Wohnsitz für den Eigenbedarf gleichzeitig maximal 3 Cannabis-Pflanzen anzubauen. Jede Pflanze kann etwa zwei Mal im Jahr geerntet werden. Jede Ernte ergibt etwa 25 g Cannabis.
>
> Nach einem Unfall hat A ständig starke Schmerzen. Der Konsum von 1 g Cannabis am Tag macht die Schmerzen erträglich. Sein Arzt hat ihm Cannabis in dieser Menge verschrieben. Die Krankenkasse übernimmt auch die Kosten. Das Medikament war jedoch immer wieder nicht lieferbar. A beantragt deshalb die Erlaubnis, nicht nur 3, sondern 12 Cannabis-Pflanzen selbst anzubauen. Die zuständige Behörde lehnt seinen Antrag ab. Der Gesetzgeber habe die Frage entschieden. Es solle bei der kontrollierten Abgabe durch Apotheken bleiben, weil dann der Wirkstoffgehalt besser kontrolliert werden kann[23]. A hat den Bescheid vor den Verwaltungsgerichten angefochten, aber auch in letzter Instanz verloren.

23 Kabinettsvorlage, https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/C/Cannabis/Gesetzentwurf_Cannabis_Kabinett.pdf, S. 77.

> Er legt Verfassungsbeschwerde ein. Die restriktive gesetzliche Regelung greife unverhältnismäßig in sein Recht aus Art. 2 II 1 GG auf körperliche Unversehrtheit ein. Er werde einem vermeidbaren Risiko schwerer Schmerzen ausgesetzt.

V. Entscheidungen

In der ersten Version habe ich GPT gefragt:

> Sollte das Bundesverfassungsgericht der Verfassungsbeschwerde stattgeben?

In 61 von 100 Fällen sagt GPT ja. In der zweiten Version habe ich GPT stattdessen gefragt:

> Sollte das Bundesverfassungsgericht der Verfassungsbeschwerde stattgeben?
>
> Bereiten Sie die Entscheidung bitte auf folgende Weise vor:
>
> a) Interessen des Beschwerdeführers
>
> aa) Welches abstrakte Gewicht hat die Linderung vermeidbarer Schmerzen?
>
> bb) Wie stark wird dieses Anliegen gefährdet, wenn dem Beschwerdeführer der Anbau von 12 Cannabis-Pflanzen verweigert wird?
>
> b) Regelungsziel
>
> aa) Welches abstrakte Gewicht hat die Sorge, dass der Wirkstoffgehalt bei Eigenanbau nicht kontrolliert werden kann?
>
> bb) Wie wahrscheinlich ist es, dass der Beschwerdeführer Schaden an seiner Gesundheit nimmt, weil der Wirkstoffgehalt des selbst angebauten Cannabis schwankt, oder unerwartet hoch ist?
>
> Kommen Sie nun zu dem abschließenden Urteil: Setzen Sie bitte die Interessen des Beschwerdeführers und das Regelungsziel zueinander ins Verhältnis. Was überwiegt: die Interessen des Beschwerdeführers, oder das staatliche Regelungsanliegen?[24]

Die Literatur zu Sprachmodellen würde das als einen „chain of thought prompt" interpretieren: man lenkt den gedanklichen Prozess, auf dem GPT zu seiner Einschätzung kommt[25]. Diese

24 Ich bitte GPT außerdem in jedem Absatz, die Antwort in dem Format JSON zu geben (https://jsonlint.com). Das erleichtert die Auswertung (und könnte GPT zu größerer Konsistenz anhalten). Ich habe diese Sätze in dieser Darstellung weggelassen, damit die Prompts leichter lesbar werden.

25 Sondos Mahmoud Bsharat et al.: Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4, https://arxiv.org/pdf/2312.16171.pdf?fbclid=IwAR1WK51PVe87PwcHBQxNcI5uFs3wgUYJxO3yCwoB8ZGFhMzfwZWzFgCt158.

Frage ist aber natürlich nicht zufällig auch eine Version von Robert Alexy's Gewichtsformel[26]. GPT wird sich noch etwas sicherer: jetzt will es in 85 von 100 Fällen der Verfassungsbeschwerde stattgeben[27].

Diese zweite Version der Anfrage ist unmittelbar der Gesetzesbegründung entnommen. Der Gesetzgeber gibt nur diesen einen Grund an, warum er Kranken den Selbstanbau (jenseits der Grenzen für Gesunde) versagen will. Das Bundesverfassungsgericht fühlt sich für die Ermittlung des legitimen Ziels eines Grundrechtseingriffs aber nicht an die Gesetzesbegründung gebunden[28]. Man könnte sich zusätzliche legitime Ziele vorstellen. Das tue ich in der dritten Version der Anfrage:

> Sollte das Bundesverfassungsgericht der Verfassungsbeschwerde stattgeben?
>
> Bereiten Sie die Entscheidung bitte auf folgende Weise vor:
>
> a) Interessen des Beschwerdeführers
>
> aa) Welches abstrakte Gewicht hat die Linderung vermeidbarer Schmerzen?
>
> bb) Wie stark wird dieses Anliegen gefährdet, wenn dem Beschwerdeführer der Anbau von 12 Cannabis-Pflanzen verweigert wird?
>
> b) Regelungsziel
>
> aa) Welches abstrakte Gewicht hat die Sorge, dass der Wirkstoffgehalt bei Eigenanbau nicht kontrolliert werden kann?
>
> bb) Wie wahrscheinlich ist es, dass der Beschwerdeführer Schaden an seiner Gesundheit nimmt, weil der Wirkstoffgehalt des selbst angebauten Cannabis schwankt, oder unerwartet hoch ist?
>
> c) Rückwirkungen auf Grenzen für Cannabiskonsum
>
> aa) Welches abstrakte Gewicht hat die Sorge, dass größere Freiräume für den Eigenanbau aus medizinischen Gründen die Entscheidung des Gesetzgebers gefährden, für Konsumzwecke nicht mehr als 3 Pflanzen pro Person zuzulassen?
>
> bb) Wie stark wird dieses Anliegen gefährdet, wenn dem Beschwerdeführer der Anbau von 12 Cannabis-Pflanzen verweigert wird?
>
> d) Infrastruktur für kontrollierte Abgabe

26 Ich habe allerdings darauf verzichtet, zwischen der objektiven Wahrscheinlichkeit einer Beeinträchtigung des jeweiligen Schutzguts und der epistemischen Wahrscheinlichkeit zu unterscheiden, dass dieses Ergebnis eintritt.

27 Ich berichte keine statistischen Tests, weil die bei der Nutzung von Sprachmodellen beliebig sind: man kann einen noch so kleinen Unterschied statistisch signifikant machen, indem man das Sprachmodell häufig genug fragt.

28 BVerfGE 11, 126, 129f..

aa) Welches abstrakte Gewicht hat die Systementscheidung des Gesetzgebers für kontrollierten Anbau und die kontrollierte Abgabe von Cannabis für medizinische Zwecke durch die Apotheken ?

bb) Wie stark wird dieses Anliegen gefährdet, wenn dem Beschwerdeführer der Anbau von 12 Cannabis-Pflanzen verweigert wird?

e) Akzeptanz der partiellen Freigabe

aa) Welches abstrakte Gewicht hat das Streben des Gesetzgebers, die politische Akzeptanz der Freigabe dadurch zu sichern, dass er strikte Grenzen vorschreibt?

bb) Wie stark wird dieses Anliegen gefährdet, wenn dem Beschwerdeführer der Anbau von 12 Cannabis-Pflanzen verweigert wird?

Kommen Sie nun zu dem abschließenden Urteil: Setzen Sie bitte die Interessen des Beschwerdeführers und die Anliegen des Gesetzgebers zueinander ins Verhältnis. Was überwiegt: die Interessen des Beschwerdeführers, oder die staatlichen Regelungsanliegen?

In dieser Anfrage liegt auf der Seite des Gesetzes mehr auf der Waagschale. GPT reagiert darauf. Nun will es nur noch in 45 von 100 Fällen der Verfassungsbeschwerde stattgeben.

Robert Alexy will nicht nur den Diskurs über die widerstreitenden privaten und öffentlichen Interessen rationalisieren. Seine Gewichtsformel gibt auch vor, wie die Belange zueinander ins Verhältnis zu setzen sind. In der vierten Version der Anfrage habe ich den letzten Absatz der dritten Version durch folgenden Text ersetzt:

Mit dem Verbot von Eigenanbau verfolgt der Staat vier Regelungsziele zugleich. Bitte aggregieren Sie im ersten Schritt das abstrakte Gewicht: alle vier Ziele zusammengenommen: welches abstrakte Gewicht hat das staatliche Regelungsanliegen?

Aggregieren Sie sodann auch die vier Wahrscheinlichkeitsurteile: wie wahrscheinlich ist es, dass das Bündel an staatlichen Regelungsanliegen gefährdet ist, wenn dem Beschwerdeführer der begehrte Eigenanbau gestattet wird?

Kommen Sie nun zu dem abschließenden Urteil: Setzen Sie bitte die Interessen des Beschwerdeführers und die Regelungsziele zueinander ins Verhältnis. Was überwiegt: die Interessen des Beschwerdeführers, oder die staatlichen Regelungsanliegen?

Folgende Kombinationen sind möglich:

hoch hoch > hoch mittel = mittel hoch > mittel mittel > hoch niedrig = niedrig hoch > mittel niedrig = niedrig mittel > niedrig niedrig

(wobei das erste Maß jeweils für das abstrakte Gewicht steht, das zweite Maß für die Wahrscheinlichkeit)

Der Beschwerdeführer sollte obsiegen, wenn (im Sinne dieser Reihung) seine Interessen höher zu gewichten sind als das Regelungsziel. Das Gericht sollte der Beschwerde nicht stattgeben, wenn (wiederum im Sinne dieser Reihung) umgekehrt das Regelungsziel höher zu gewichten ist als die Interessen des Beschwerdeführers.

Es ist möglich, dass die Entscheidung auf dieser Grundlage nicht getroffen werden kann, weil die Kombination der Einschätzung zum abstrakten Gewicht und zur Wahrscheinlichkeit auf beiden Seiten der Entscheidung zum selben Ergebnis führt (zum Beispiel weil das abstrakte Gewicht der Interessen des Beschwerdeführers zwar hoch ist, die Wahrscheinlichkeit einer Beeinträchtigung dagegen nur mittel; andererseits das abstrakte Gewicht des Regelungsziels nur mittel ist, die Wahrscheinlichkeit seiner Gefährdung aber hoch). Bitte entscheiden Sie auch in diesem Fall, und begründen, warum Sie zu Ihrer Entscheidung gekommen sind.

Diese (eigentlich ja nur prozedurale) Vorgabe hat einen deutlichen Effekt auf die Entscheidung in der Sache. Nun will GPT nur noch in 24 von 100 Fällen der Entscheidung stattgeben.

In dieser vierten Version überlasse ich GPT, wie es die abstrakten Gewichte und die Wahrscheinlichkeit ihrer Beeeinträchtigung aggregiert. Es ist bekannt, dass GPT 3.5 mit (quasi) mathematischen Aufgaben nicht gut zurechtkommt, GPT 4 dagegen viel besser. Andererseits habe ich dargelegt, dass GPT 4 die Varianz der Antworten (häufig deutlich) reduziert. In der fünften Version kombiniere ich die komparativen Vor- und Nachteile beider Modelle auf folgende Weise: ich gebe GPT 4 die Antworten, die GPT 3.5 in der dritten Version gegeben hat. GPT 4 erhält also (nur) die 100 Einschätzungen von GPT 3.5 zum abstrakten Gewicht und zur Wahrscheinlichkeit einer Beeinträchtigung für jeden der 5 Belange (1 für den Beschwerdeführer, 4 für die öffentliche Hand). Ich frage GPT 4:

Ich habe GPT 3.5 die im Anschluss wörtlich zitierten Fragen gestellt. Ich zitiere im zweiten Teil dieses Prompts die Antworten, die GPT 3.5 gegeben hat (mit Ausnahme der abschließenden Einschätzung). Ich möchte von Ihnen wissen, wie Sie entscheiden würden, VORAUSGESETZT SIE WÄREN ZU DENSELBEN EINSCHÄTZUNGEN GELANGT. Sie sollen den Fall also nicht erneut bewerten. Mich interessiert nur, getrennt für jede Antwort von GPT 3.5, wie Sie die Teileinschätzungen zu einem Gesamturteil aggregiert hätten.

[Einschätzungen von GPT 3.5]

Ihre Aufgabe

Bitte teilen Sie nun mit (getrennt für jede Antwort von GPT 3.5), wie Sie entscheiden würden, VORAUSGESETZT SIE WÄREN ZU DENSELBEN EINSCHÄTZUNGEN GELANGT.

GPT 4 kommt praktisch zum selben Ergebnis wie GPT 3.5: in 43 von 100 Fällen würde es der Verfassungsbeschwerde stattgeben.

In der sechsten Version weise ich GPT 4 überdies an, zunächst das abstrakte Gesamtgewicht der öffentlichen Belange, und die Wahrscheinlichkeit ihrer Verletzung zu aggregieren, und dann erst zu einer abschließenden Einschätzung zu kommen:

> ### Ihre Aufgabe ###
>
> Mit dem Verbot von Eigenanbau verfolgt der Staat vier Regelungsziele zugleich. Bitte aggregieren Sie im ersten Schritt das abstrakte Gewicht: alle vier Ziele zusammengenommen: welches abstrakte Gewicht hat das staatliche Regelungsanliegen?
>
> Aggregieren Sie sodann auch die vier Wahrscheinlichkeitsurteile: wie wahrscheinlich ist es, dass das Bündel an staatlichen Regelungsanliegen gefährdet ist, wenn dem Beschwerdeführer der begehrte Eigenanbau gestattet wird?
>
> Kommen Sie nun zu dem abschließenden Urteil: Setzen Sie bitte die Interessen des Beschwerdeführers und die Regelungsziele zueinander ins Verhältnis. Was überwiegt: die Interessen des Beschwerdeführers, oder die staatlichen Regelungsanliegen?
>
> Bitte teilen Sie nun mit (getrennt für jede Antwort von GPT 3.5), wie Sie entscheiden würden, VORAUSGESETZT SIE WÄREN ZU DENSELBEN EINSCHÄTZUNGEN GELANGT.

Diese Vorgabe hat einen deutlichen Effekt. GPT will nun nur noch in 27 von 100 Fällen der Verfassungsbeschwerde stattgeben.

GPT hat keine richterliche Autorität. Wenn ich das Sprachmodell frage, ob das Bundesverfassungsgericht der Verfassungsbeschwerde stattgeben sollte, erfahre ich, wie die Sprecher, deren Äußerungen GPT zum Training des Sprachmodells genutzt hat, den Fall entschieden hätten. Stattdessen kann ich das Sprachmodell auch fragen, welche Erwartungen es darüber hat, wie das Bundesverfassungsgericht den Fall entscheiden würde. Abbildung 1 berichtet diese Ergebnisse (türkise Balken), und stellt sie ins Verhältnis zu der Frage an GPT, wie die Entscheidung ausfallen sollte (lachsfarbene Balken). Der Vergleich belegt einen weiteren rationalisierenden Effekt der Gewichtsformel. Wenn ich nur vergleiche, wie die Entscheidung ausfallen sollte, und wie sie nach der Erwartung von GPT ausfallen wird, finde ich einen starken Unterschied. GPT denkt, dass das Gericht sehr zurückhaltend ist. Doch sobald die Gewichtsformel zum Einsatz kommt (und sei es auch nur teilweise, als Anweisung, zunächst das abstrakte Gewicht und die Wahrscheinlichkeit der Beeinträchtigung herauszuarbeiten), spielt der Unterschied zwischen „sollte“ und „wird“ nur noch eine kleine Rolle.

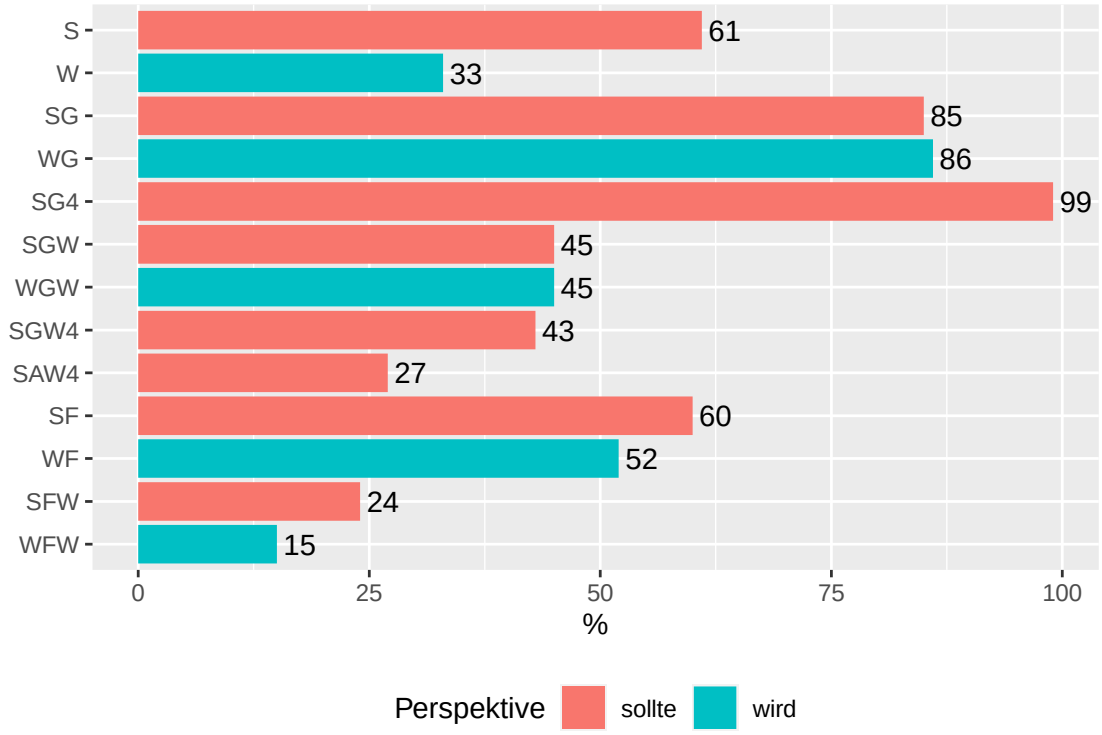

Abbildung 1
Entscheidungen von GPT
jeweils 100 Anfragen, Anzahl der Fälle, in denen GPT der Verfassungsbeschwerde stattgeben will
S: „sollte", W: „wird"
SG, WG: die Gründe, die die Gesetzesbegründung diskutiert
SG4: GPT aggregiert die Einschätzungen von GPT 3.5
SGW, WGW: weitere drei plausible Begründungen für das Verbot
SGW4: GPT 4 aggregiert die Einschätzungen von GPT 3.5
SAW4: GPT 4 aggregiert die öffentlichen Belange zunächst getrennt
SF, WF: mit Gewichtsformel
SFW, WFW: alle plausiblen öffentlichen Belange mit Gewichtsformel

VI. Gewichtung der Belange

Robert Alexy will nicht nur das Ergebnis der Abwägung beeinflussen. Vornehmlich gilt seine Sorge dem Prozess der Abwägung. Mit Hilfe der Gewichtsformel möchte er auch erreichen, dass die Gewichte in nachvollziehbarer Weise gebildet werden. Abbildung 2 zeigt die Effekte, zunächst beschränkt auf die Belange, die in der Regierungsbegründung genannt sind. GPT ist sehr sicher, dass es hohes normatives Gewicht hat, einem Kranken vermeidbare Schmerzen zu ersparen. Wie wahrscheinlich dieser Nachteil eintritt, wenn dem Beschwerdeführer versagt bleibt, selbst genügend Pflanzen anzubauen, beurteilt GPT dagegen je nach der Art des Prompts unterschiedlich. Wenn man nicht nur nach den Belangen fragt, die in der Regierungsbegründung genannt sind (SGW, WGW, SFW, WFW), neigt GPT dazu, diese Wahrscheinlichkeit niedriger einzuschätzen. Das deutet auf einen Rückkopplungseffekt hin: GPT sieht, dass die Einschätzung der Interessen des Beschwerdeführers als „hoch hoch" kaum Raum für die Abwägung mit öffentlichen Interessen lässt (die aggregierten öffentlichen Interessen müssten

auch „hoch hoch“ eingeschätzt werden). Durch die Anpassung des Wahrscheinlichkeitsurteils schafft das Sprachmodell gleichsam Raum für die Abwägung.

Zu dieser Erklärung passt, dass GPT das abstrakte Gewicht des erklärten öffentlichen Interesses eher als nur mittel bedeutsam einstuft, und die Wahrscheinlichkeit seiner Beeinträchtigung erst recht.

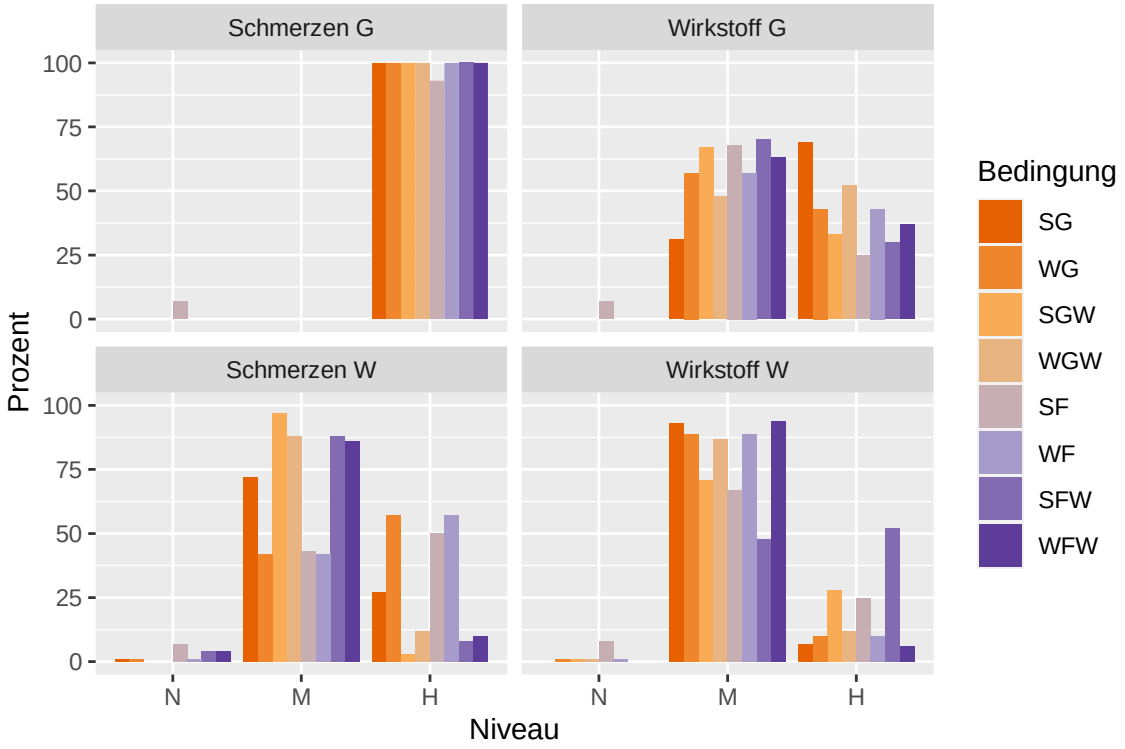

Abbildung 2
Abstraktes Gewicht und Wahrscheinlichkeit der Beeinträchtigung
Belange nach Maßgabe der Regierungsbegründung
N: niedrig, M: mittel, H: hoch
SG, WG: die Gründe, die die Gesetzesbegründung diskutiert
SGW, WGW: weitere drei plausible Begründungen für das Verbot
SF, WF: mit Gewichtsformel
SFW, WFW: alle plausiblen öffentlichen Belange mit Gewichtsformel

Abbildung 2 hat gezeigt, dass die Wahl des Prompts erhebliche Auswirkungen darauf hat, wie GPT die Wahrscheinlichkeit einer Beeinträchtigung der privaten Interessen des Antragstellers einschätzt. In Abbildung 3 sind die Balken für die vier verschiedenen Arten, GPT zu fragen, dagegen jeweils sehr ähnlich: GPT ist sich relativ sicher, welches abstrakte Gewicht der jeweilige öffentliche Belang hat, und wie wahrscheinlich er beeinträchtigt wird, wenn der Staat dem Beschwerdeführer verwehrt, selbst genügend Cannabis anzubauen. Die einzige Ausnahme ist das abstrakte Gewicht des erklärten staatlichen Anliegens: in etwa der Hälfte der Fälle hält GPT dieses Anliegen für sehr bedeutsam, in der anderen Hälfte für mittelmäßig bedeutsam. Diese Einschätzungen hängen aber nicht systematisch vom Prompt ab. GPT hat dagegen keinen Zweifel, dass dieses Anliegen nicht mit hoher Gewissheit verletzt wird, wenn dem Beschwerdeführer der Anbau von 12 Pflanzen gestattet wird. Besonders interessant sind die Einschätzungen zur Systementscheidung des Gesetzgebers: GPT hält das für ein sehr

schützenswertes Gut; aber GPT glaubt nicht, dass eine Ausnahme für den Beschwerdeführer diese Systementscheidung gefährdet.

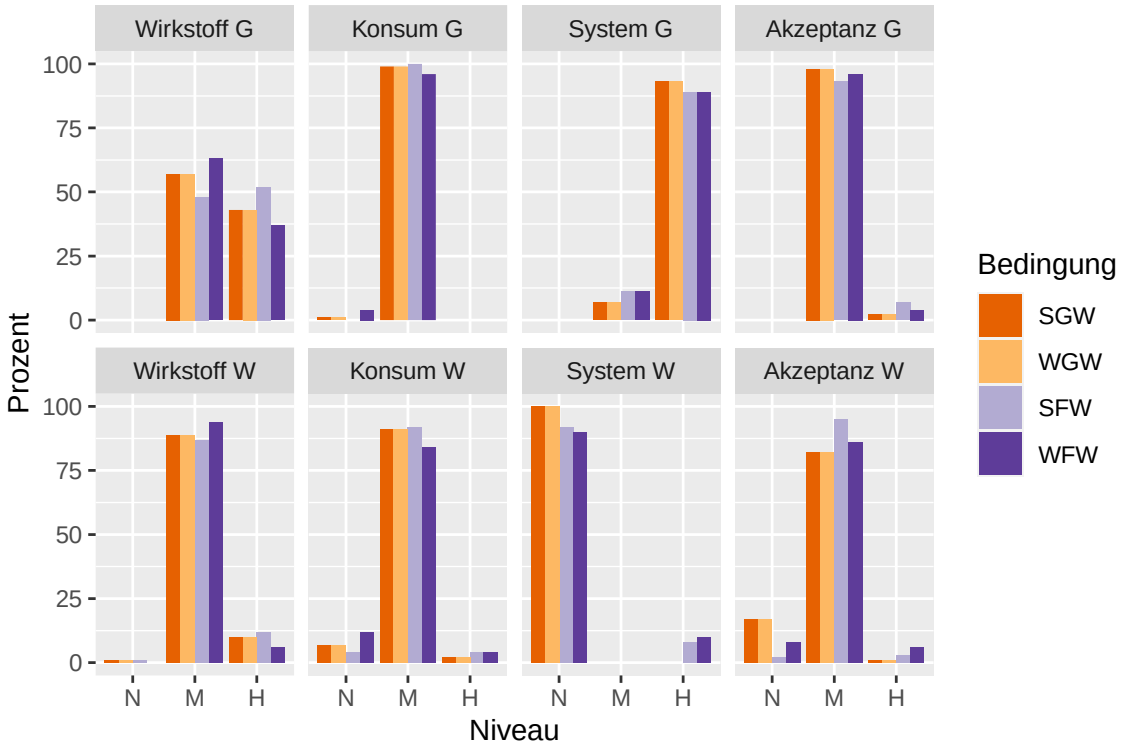

Abbildung 3
Abstraktes Gewicht und Wahrscheinlichkeit der Beeinträchtigung öffentlicher Belange
sämtliche normativen Belange
N: niedrig, M: mittel, H: hoch
SGW, WGW: alle plausiblen öffentlichen Belange
SFW, WFW: alle plausiblen öffentlichen Belange mit Gewichtsformel

Abbildung 4 zeigt schließlich, dass es einen beträchtlichen Unterschied macht, ob GPT 3.5 die öffentlichen Belange aggregiert, oder ob GPT 4 das tut. GPT 4 kommt sehr viel häufiger zu der Einschätzung, dass die Belange zusammen genommen ein hohes abstraktes Gewicht haben. Andererseits hält es eine Beeinträchtigung der Gesamtheit der öffentlichen Anliegen häufiger für wenig wahrscheinlich.

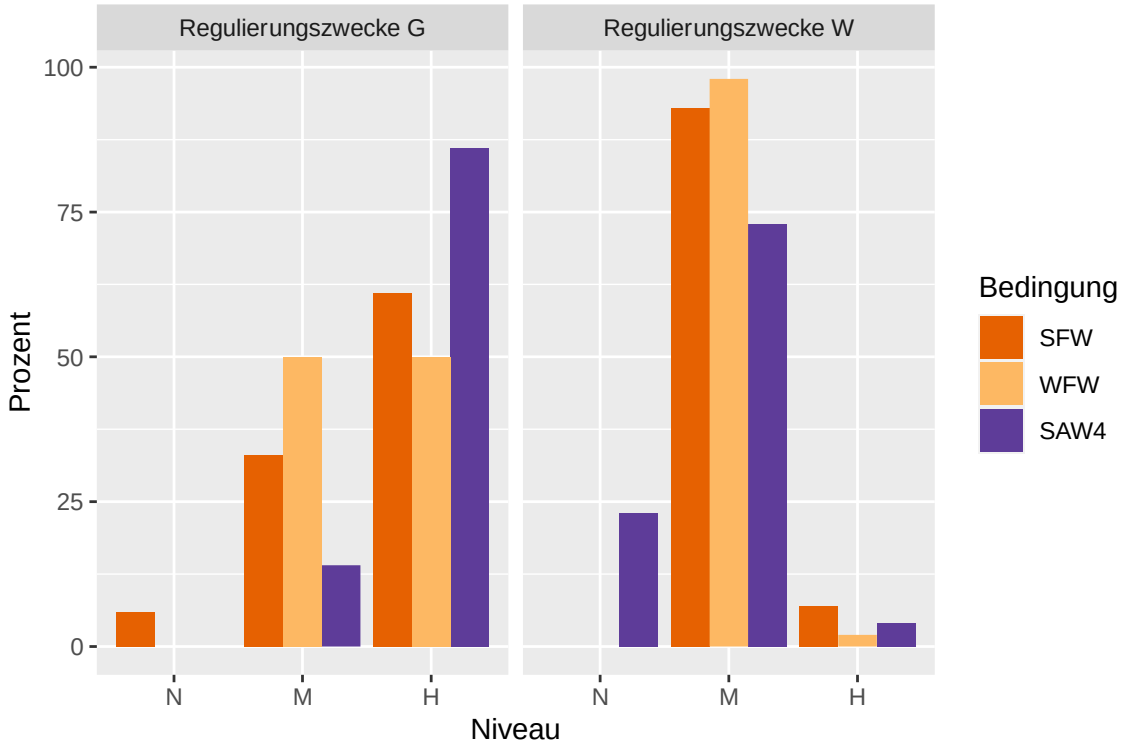

Abbildung 4
Explizite Aggregation der Regulierungszwecke mit Hilfe der Gewichtsformel
SFW, WFW: GPT 3.5
SAW4: GPT 4

VII. Konsistenz

Robert Alexy hat die Gewichtsformel mechanisch gedacht. Sobald die abstrakten Gewichte und die Wahrscheinlichkeiten festgelegt sind, ist die Abwägung eine reine Rechenoperation. Entscheidungsfreiraum bleibt nur, wenn gleich viel für und wider eine stattgebende Entscheidung spricht. Hält sich GPT daran? Um das prüfen zu können, übersetze ich ein hohes abstraktes Gewicht, und eine hohe Wahrscheinlichkeit der Verletzung eines normativen Belangs, in das Gewicht 0.9. Ein mittleres Gewicht, oder eine mittelhohe Wahrscheinlichkeit der Verletzung, übersetze ich in das Gewicht 0.5. Ein niedriges Gewicht, oder die niedrige Wahrscheinlichkeit der Verletzung, übersetze ich in das Gewicht 0.1[29]. Ich multipliziere jeweils den Wert für das abstrakte Gewicht mit dem Wert für die Wahrscheinlichkeit. Der höchste denkbare Wert ist also 0.9 * 0.9 = 0.81. Der niedrigste denkbare Wert ist 0.1 * 0.1 = 0.01. Die von der Gewichtsformel geforderte Entscheidung ergibt sich aus der Differenz zwischen dem aggregierte Gewicht der Belange, die für den Beschwerdeführer streiten, und dem aggregierten Gewicht der Belange, die für das staatliche Verbot streiten. Wenn diese Differenz negativ ist, dann sollte die Verfassungsbeschwerde abgewiesen werden. Wenn sie positiv ist, sollte das Gericht der Beschwerde stattgeben.

29 Auf die konkreten Zahlen kommt es nicht an. Ich brauche sie nur für eine einfache graphische Darstellung.

Abbildung 5 zeigt, dass GPT 3.5 mit dieser Aggregation überfordert ist. Vor allem die ablehnenden Entscheidungen sind nahezu symmetrisch über die Gewichte verteilt: selbst wenn aus den Gewichten ganz eindeutig folgt, dass die Verfassungsbeschwerde Erfolg haben sollte, lehnt GPT 3.5 sie oft ab. GPT 4 ist sehr viel verlässlicher. Wenn die Gewichte für den Erfolg der Beschwerde sprechen, spricht das Sprachmodell das auch stets aus. Es gibt aber eine beträchtliche Anzahl von Entscheidungen, bei denen GPT 4 die Verfassungsbeschwerde ablehnt, obwohl es nach der Gewichtsformel hätte zusprechen müssen.

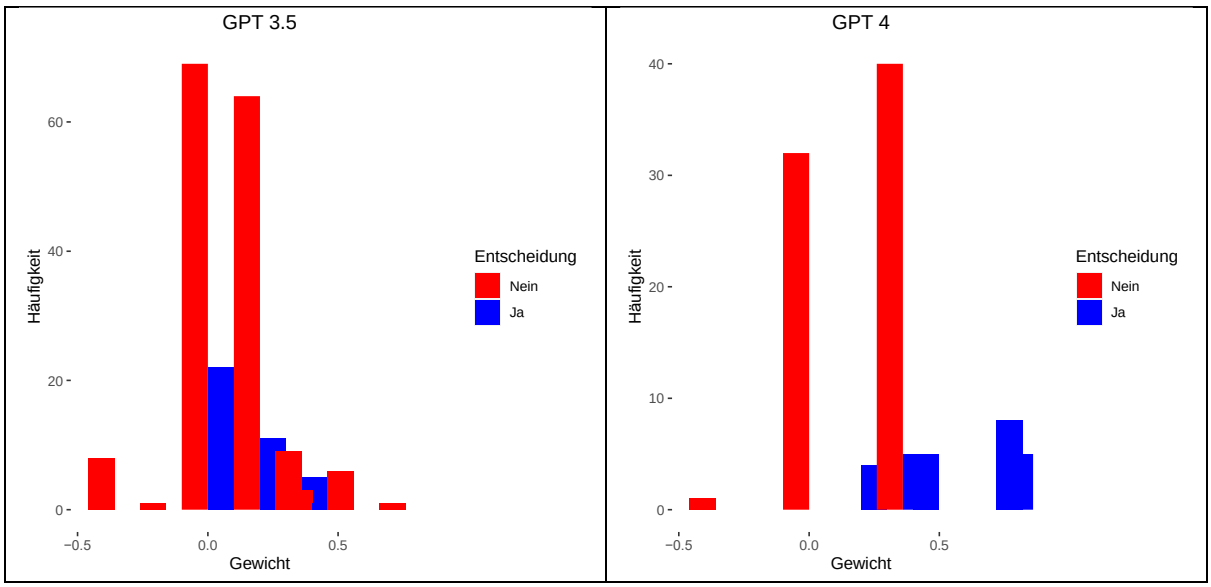

Abbildung 5
Konsistenz

Ein Blick in die Rohdaten zeigt, dass diese Fälle alle gleich aussehen. Die Gewichte habe ich ja den Entscheidungen von GPT 3.5 entnommen. Nur die Aggregation dieser Gewichte habe ich GPT 4 überlassen. GPT 3.5 hatte in allen Fällen gesagt, dass das abstrakte Gewicht der Interessen des Beschwerdeführers hoch ist, und dass die Verletzung dieser Interesse sehr wahrscheinlich ist. Das Produkt der beiden Gewichte auf der Seite des Beschwerdeführers ist also maximal (0.81 in meiner Parameterisierung). Demgegenüber hatte GPT 3.5 das aggregierte Gewicht der widerstreitenden öffentlichen Interessen zwar ebenfalls für hoch gehalten, die Wahrscheinlichkeit ihrer Beeinträchtigung dagegen nur als mittelhoch eingeschätzt. Das Produkt der öffentlichen Interessen beträgt deshalb nur 0.45. Offensichtlich hält GPT 4 das Meßverfahren der Gewichtsformel für zu rigide. Es findet, dass die öffentlichen Interessen zusammengenommen ein höheres abstraktes Gewicht haben als das Interesse des Beschwerdeführers an der Vermeidung von Schmerzen.

VIII. Die rationalisierende Kraft der Gewichtsformel

Was kann der neugierige Leser aus diesen wiederholten Anfragen an das Sprachmodell lernen? Sicher nicht, dass die Grundrechtsauslegung künftig einem amerikanischen Unternehmen (open.ai) überlassen bleiben soll. Auch nicht, dass Anfragen an Sprachmodelle in Verfassungsgerichtsprozesse einziehen sollten. Das Sprachmodell dient mir nur als Substitut für Fragen an menschliche Teilnehmer. Es vermittelt eine Ahnung davon, wie Verfassungsrichter entscheiden würden, wenn sie sich an die Gewichtsformel von Robert Alexy gebunden fühlten. Ob sie genauso werten würden, mag man mit Grund bezweifeln. Aber darauf kommt es auch nicht an. Das Ziel dieser Untersuchung ist nicht, den Beispielsfall zu entscheiden, oder irgendeinen anderen konkreten Fall. Ich will alleine verstehen, welchen empirischen Effekt die verschiedenen Versionen der Gewichtsformel haben. Ich bin deshalb nicht an den absoluten Zahlen interessiert, sondern an den Unterschieden in den Zahlen zwischen den verschiedenen Fassungen der Formel.

Abbildung 1 zeigt, dass die Verwendung der Gewichtsformel bedeutsam ist für die Entscheidung des Falles. Abbildung 2, Abbildung 3 und Abbildung 4 zeigen, dass entscheidend ist, ob man die Gewichtsformel zur Vorbereitung der Entscheidung verwendet, und welche Regulierungszwecke man dabei berücksichtigt. Abbildung 5 zeigt schließlich, dass man mit einer geeigneten Kombination von GPT 3.5 und GPT 4 zwischen Fällen unterscheiden kann, in denen das Sprachmodell inkonsistent ist, und Fällen, in denen es die einfachste Version der Gewichtsformel für zu rigide hält.

Weil Sprachmodelle so neu sind, ist die Literatur vor allem daran interessiert, die Antworten der Sprachmodelle den Antworten menschlicher Teilnehmer auf dieselben Fragen anzugleichen. Das ist wichtig, wenn man Aufgaben an Sprachmodelle delegieren will, die bislang Menschen übernommen hatten. Doch menschliche Entscheider sind nicht unfehlbar. Diese Untersuchung zeigt, wie sinnvoll es ist, in die umgekehrte Richtung zu denken. Robert Alexy hat seine Formel entwickelt, bevor auch nur der Gedanke an Sprachmodelle gefasst war. Was sich bei meinen Anfragen als Prompt bewährt, war von ihm als Entscheidungshilfe für Verfassungsrechtler gedacht. Mittlerweile ist „Prompt Engineering“ eine ganze Industrie geworden. Die Rechtswissenschaft kann daraus Nutzen ziehen. Sie kann menschlichen Rechtsanwendern Entscheidungshilfen an die Hand geben, die von gut funktionierenden Prompts inspiriert sind.

Wie wirksam diese Entscheidungshilfe im Verfassungsgericht sein könnte, könnte am Ende nur das Gericht selbst herausbekommen. Doch die Tatsache, dass diese Entscheidungshilfe einen beträchtlichen Einfluss auf Sprachmodelle hat, begründet eine Hypothese. Der Gedanke liegt nicht fern, dass menschliche Entscheider davon in ähnlicher Weise gelenkt werden könnten. Auch für menschliche Entscheider, selbst für Verfassungsrichter, könnte das „Prompt Engineering“ mit Hilfe der Gewichtsformel eine rationalisierende Wirkung haben.

Falls die Analogie zu den Ergebnissen der Sprachmodelle trägt, wäre diese rationalisierende Wirkung nicht auf das Ergebnis der Entscheidung beschränkt. Auch der Weg zur abschließenden Entscheidung würde vorgespurt. Die Entscheidungen würden transparenter. Das würde insbesondere dann zu Buche schlagen, wenn der Entscheider am Ende von der Formel

abweichen möchte, so wie es GPT 4 in einer ganz bestimmten Fallgruppe regelmäßig getan hat. Auch die Abweichung von der Formel hätte eine rationalisierende Wirkung. Der Entscheider müsste offenlegen, aus welchem Grunde er die Formel im konkreten Fall zu rigide findet. Auch wer, wie ich selbst, eigentlich der richterlichen Intuition mehr zutraut und wer sich weniger leicht damit tut, jegliche normativen Belange in die einheitliche Währung der Verfassungsmäßigkeit einzuwechseln, kann also von der Anwendung der Formel viel lernen.