Treu und Glauben aus Sicht der Rechts- und Verhaltensökonomie

Treu und Glauben: Frag GPT

Christoph Engel

1. Treu und Glauben als Einfallstor für Fairnessnormen

Als Studenten hat man uns eingebläut: wenn Ihr den Fall mit Treu und Glauben löst, gibt das höchstens ausreichend. Unsere akademischen Lehrer wollten damit vor allem auf die rechtserzeugende Funktion des Grundsatzes von Treu und Glauben hinweisen. Treu und Glauben ist im juristischen Normalbetrieb keine Norm, unter die man den Sachverhalt subsumiert. Weil das zu Lösungen führt, die im Einklang mit Treu und Glauben stehen, hat die Rechtsordnung viel spezifischere Regeln entwickelt. Wenn unklar ist, ob der konkrete Fall unter solch eine spezifische Regel gebracht werden kann, dann soll sich die juristische Anstrengung darauf konzentrieren, diese Norm besser auszuleuchten. Demgegenüber erscheint der direkte Rückgriff auf die Generalklausel von Treu und Glauben wie der Beweis geistiger Faulheit, ja beinahe wie eine dogmatische Konkurserklärung.

Aber auch in der juristischen Wissenschaft gilt: quod licet Iovi non licet bovi. Was Studenten aus pädagogischen Gründen verboten ist, kann zu einer anregenden Herausforderung für gestandene juristische Wissenschaftler werden. Diese Überzeugung prägt jedenfalls ganz offensichtlich die Architektur dieses Symposiums. Wir werden viel darüber lernen, auf welchen Wegen der Geist von Treu und Glauben Eingang in die unterschiedlichsten Rechtsgebiete gefunden hat. Jeder Autor wird in seinem Rechtsgebiet mehr als nur Spuren dieses Geistes finden. Wie sich ein Rechtsgebiet den Gedanken von Treu und Glauben anverwandelt hat, und wie offen es diese Wirkung ausweist, wird spannende Vergleiche zwischen den Teilgebieten des Rechts ermöglichen.

Die Herausgeber haben diesen Beitrag vor die dogmatische Klammer gezogen. Sie sind neugierig, was die Rechts– und Verhaltensökonomie zu dem Thema zu sagen hat. Wenn man lang genug sucht, findet man einschlägig klingende Texte [1]. Aber diese Texte stammen sicher nicht aus dem Kern der Debatten in der Rechtsökonomie oder der Verhaltensökonomie. Doch wie so oft verhandeln unterschiedliche Disziplinen eng verwandte Gegenstände in ihrer je eigenen Begrifflichkeit. So liegt es auch hier, oder vorsichtiger: das ist die zentrale These dieses Beitrags. Ich möchte eine Brücke schlagen von einem Konzept, mit dem sich die Rechts– und Verhaltensökonomie in den letzten zwei Jahrzehnten ganz außerordentlich intensiv beschäftigt hat, und dem Jahrtausende alten Interesse der Juristen an Treu und Glauben. Meine These lautet: das Konzept von Treu und Glauben ist eine juristische Ausprägung von Fairness.

Wenn Rechtsdogmatiker sagen, dass ein Verhalten im Widerspruch zu Treu und Glauben steht, dann meinen sie damit: das Verhalten ist so offensichtlich unfair, dass die Rechtsordnung diesen Verstoß gegen die normative Erwartung fairen Verhaltens nicht ignorieren kann. Und wenn Rechtsdogmatiker sagen, dass eine richterliche Entscheidung von Treu und Glauben gefordert ist, dann meinen sie damit: würde die Rechtsordnung anders entscheiden, wäre die Entscheidung unerträglich unfair. In beiden Deutungen, der negativen wie der positiven, dem Urteil über das Verhalten der Parteien und der Einschätzung der richterlichen Intervention, habe ich einen Schwellenwert eingefügt. Nicht jede unfaire Handlung und nicht jede unfaire richterliche Entscheidung ist deshalb schon rechtswidrig. Die Rechtsordnung würde sich verheben, würde sie perfekte Fairness anstreben. Aber wenn der Handlung oder der Entscheidung die Unfairness ins Gesicht geschrieben steht, dann wird das auch rechtlich bedeutsam.

Ich hoffe, ich kann Sie im Folgenden überzeugen, dass die Deutung von Treu und Glauben als die Sorge um Fairness fruchtbar ist. Das Tor von der Rechtsdogmatik zur Rechts– und Verhaltensökonomie ist geöffnet. Durch dieses Tor öffnet sich aber kein gelobtes Land. Die Verhaltensökonomen haben zwar intensiv über Fairness nachgedacht und sehr viel (vor allem experimentelle) Evidenz angehäuft. Vor allem hat diese Grundlagenforschung aber eines deutlich gemacht: Fairness ist nicht nur ungeheuer wirkmächtig, sondern auch ungeheuer facettenreich. Was fair ist, steht nicht ein für alle Mal fest. Der Kontext spielt eine große Rolle. Es kommt darauf an, wie die Beteiligten diesen Kontext wahrnehmen. Wenn es mehr als eine Möglichkeit gibt, faires Verhalten zu bestimmen, dann neigen viele Menschen (wohl eher unbewusst) dazu, gerade den Aspekt von Fairness herauszustellen, der ihnen persönlich nutzt.

Wer sich als Jurist auf die Fairnessforschung einlässt, der erfährt etwas über mögliche Dimensionen. Er erweitert seinen juristischen Argumentationshaushalt. Aber er kann die intrikaten juristischen Wertungsprobleme nicht

[1] S. etwa Marinova, D. (2021). Good Faith In The Contract Formation. In THE LAW AND THE BUSINESS IN THE CONTEMPORARY SOCIETY (pp. 88-114). Chirico, F. (2010). The economic function of good faith in European contract law. Economic Analysis of the Draft Common Frame of Reference, 31. Duke, A. (2007). A universal duty of good faith: an economic perspective. Monash University Law Review, 33(1), 182-202. Mackaay, E. (2012). Good faith in civil law systems: A legal-economic analysis. Revista chilena de derecho privado(18), 149-177. Mackaay, E., & Leblanc, V. (2003). The law and economics of good faith in the civil law of contract. European Association of Law and Economics. Conference,

einfach an die Ökonomen delegieren. Er muss am Ende doch selbst entscheiden. Die wichtigste Botschaft ist deshalb keine konzeptionelle, sondern eine methodische. Juristen sind daran gewohnt, Wertungsfragen im Diskurs zu lösen. Das entspricht nicht nur Jahrtausende alter Tradition. Je mehr die Wertung durch juristische Dogmatik gelenkt wird, desto mehr dient die Professionalisierung des Diskurses auch als Gewähr der Richtigkeit. Nur wer in all den geronnenen Wertentscheidungen geschult ist, die Eingang in die Dogmatik gefunden haben, der darf die Hoffnung haben, zu einer wohlabgewogenen Entscheidung zu finden. Aber genau an diesem festen Grund fehlt es, wenn sich der Rechtsanwender ausnahmsweise doch einmal direkt auf Treu und Glauben beruft. Dann kann Professionalisierung nicht mehr im selben Maße für Richtigkeit sorgen.

Wenn Juristen ungefiltert mit Treu und Glauben argumentieren, dann hantieren sie im Ergebnis direkt mit Fairnessnormen und Fairnessüberzeugungen. Sie werden das für gewöhnlich sorgsam und gewissenhaft tun. Aber sie können weniger als sonst darauf vertrauen, in ihren Werturteilen von der Dogmatik, und damit der kollektiven Erfahrung ihrer Disziplin, gehalten zu sein. Dieser Unterschied könnte nahelegen, dass Rechtsanwender mehr Gewicht darauf legen, wie die Bevölkerung das Fairnessproblem bewertet. An mehreren Beispielen zeige ich, dass die methodischen Hürden für solche auf den konkreten Konflikt bezogene empirische Arbeit mittlerweile deutlich niedriger geworden sind. Diese Möglichkeit ergibt sich aus der Verfügbarkeit von Sprachmodellen. Man kann mit vertretbarem Aufwand ermitteln, ob das populäre Sprachmodell GPT im zur Entscheidung stehenden Fall einen Verstoß gegen Treu und Glauben feststellt. Diese Einschätzung sollte nicht an die Stelle des richterlichen Urteils treten. Aber sie könnte dem Richter (und den Parteien) Anlass zum Nachdenken geben.

2. Konkurrierende Fairnessnormen und ihre Bedeutung für das Recht

Die Verhaltensökonomen sind fasziniert von Fairness. Das Fach ist ja dem methodologischen Individualismus verpflichtet. Es erklärt nicht nur wirtschaftliches, sondern potentiell alles menschliche Verhalten mit der Verfolgung des eigenen Nutzens. In erster Näherung besteht der Nutzen aus dem Einkommen. Wenn ein Individuum sein Einkommen maximiert, hat es keinen Grund, auf einen höheren Gewinn zu verzichten, nur weil ihm das einer anderen Person gegenüber unfair erscheint. Umgekehrt hat das Individuum auch keinen Grund, sich gegen das Verhalten Dritter zu wehren, weil es sich unfair behandelt fühlt. Solch ein Individuum versteht, dass andere die Chance auf einen höheren Gewinn ergreifen, auch wenn das zu Lasten des ersten Individuums geht.

Vor diesem Hintergrund sind zwei klassische experimentelle Resultate überraschend. Im sogenannten Diktatorspiel werden zwei Versuchspersonen zufällig einander zugeordnet. Sie interagieren anonym miteinander, sodass soziale Sanktionen im Anschluss an das Experiment ausgeschlossen sind. Der Diktator erhält einen Geldbetrag zur freien Verfügung. Er weiß, dass sein Spielpartner keinen Geldbetrag erhalten hat. Der Diktator erhält die Möglichkeit, einen Teil seines Geldbetrags an den Partner abzugeben. Wenn ein Diktator seinen Gewinn maximiert, wird er das nicht tun. Tatsächlich geben viele Versuchspersonen im Experiment aber einen beträchtlichen Teil des Geldbetrags weiter[2]. Es gibt verschiedene Möglichkeiten, dieses robuste Resultat zu erklären. Eine konsistente Erklärung ist Ungleichheitsaversion[3]. Der Diktator fühlt sich schlecht, wenn er nichts abgibt. Das ist ein Fairnessargument. Es geht um die faire Allokation knapper Ressourcen. Man kann das auch eine faire Verteilung nennen.

Das zweite überraschende Resultat stammt aus dem so genannten Ultimatumspiel[4]. Auch in diesem Spiel werden zwei Personen zufällig einander zugeordnet und interagieren anonym. Wiederum erhält eine Versuchsperson einen Geldbetrag zu ihrer Verfügung. Sie kann diesen Betrag beliebig zwischen sich selbst und der zweiten Versuchsperson aufteilen. Jetzt ist die zweite Person aber nicht mehr passiv. Sie kann die vorgeschlagene Aufteilung annehmen oder ablehnen. Lehnt sie ab, erhält keiner der beiden Teilnehmer etwas. Falls beide Teilnehmer ihr Einkommen maximieren, ist die Vorhersage auch in diesem Spiel eindeutig,. Der erste Teilnehmer weiß dann nämlich, dass der zweite Teilnehmer einen positiven Betrag besser findet, als nichts zu erhalten. Das

[2] Engel, C. (2011). Dictator Games. A Meta-Study. Experimental Economics, 14, 583-610: Kahneman, D., Knetsch, J. L., & Thaler, R. (1986). Fairness and the Assumptions of Economics. Journal of Business, 59, S285-S300.

[3] Bolton, G. E., & Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity and Competition. American Economic Review, 90, 166-193; Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. Quarterly Journal of Economics, 114, 817-868.

[4] Güth, W., Schmittberger, R., & Schwarze, B. (1982). An Experimental Analysis of Ultimatum Bargaining. Journal of Economic Behavior and Organization, 3, 367-388; Cooper, D. J., & Dutcher, E. G. (2011). The Dynamics of Responder Behavior in Ultimatum Games. A Meta-study. Experimental Economics, 14(4), 519-546.

antizipiert der erste Teilnehmer [5]. Er überlässt dem zweiten Teilnehmer nur einen minimalen Anteil und rechnet fest damit, dass der zweite Teilnehmer dieses Angebot annehmen wird. Tatsächlich lehnen viele Teilnehmer deutlich asymmetrische Verteilungen aber ab. Sie verzichten lieber selbst, wenn sie dadurch verhindern können, dass der erste Teilnehmer sie ausbeutet[6]. Auch dieses Resultat lässt sich mit Ungleichheitsaversion erklären.

Ungleichheitsaversion denkt vom Ergebnis her. Die Verteilung ist unfair, weil eine Person am Ende weniger verdient als eine andere, ohne dass für diese Ungleichheit eine Rechtfertigung zu erkennen wäre. Eine Variante des Ultimatumspiels zeigt, dass das noch nicht die ganze Geschichte ist. Wenn die Triebkraft ausschließlich der Wunsch nach Gleichverteilung ist, dann müsste die zweite Versuchsperson auch dann ablehnen, wenn die Ungleichheit vom Experimentator vorgegeben ist. Tatsächlich werden solche Verteilungen aber viel seltener abgelehnt[7]. Das zeigt, dass Fairness nicht nur vom Ergebnis her gedacht werden kann, sondern auch von den erkennbaren Intentionen.[8]

Beides sind materiale Fairnesskonzepte. Andere Experimente zeigen, dass Versuchspersonen auch eine Präferenz für prozedurale Fairness haben[9]. Wenn sie zwischen zwei Umgebungen wählen können, ziehen sie die Umgebung vor, in der sie einem Dritten erklären können, warum er zu ihren Gunsten entscheiden sollte. Interessanterweise nützt ihnen diese Möglichkeit, ihren Standpunkt zu vertreten, aber gar nicht. Im Ergebnis stellen sie sich sogar schlechter. Denn wenn der Entscheider ihnen noch mehr hätte entgegenkommen wollen, dann nimmt er seine Entscheidung auf das zurück, was die Versuchsperson gefordert hat. Wenn sie dagegen mehr fordert, als der Entscheider geben wollte, dann lässt er sich davon kaum beeindrucken[10].

In solchen sehr einfachen Experimenten ist es offensichtlich, welche Handlung fair wäre. In der Lebenswirklichkeit liegt der Fall häufig komplizierter. Soll derjenige mehr erhalten, der sich besonders stark angestrengt hat? Soll der mehr erhalten, der besonders fähig ist? Oder soll es auf den Status ankommen, so dass zum Beispiel Kinder und Frauen besser behandelt werden als Männer? Soll derjenige einen Ausgleich erhalten, der bei früherer Gelegenheit nicht gut weggekommen ist? Experimente zeigen nicht nur, dass es für alle diese Fairnessnormen Unterstützung gibt[11]. Vielmehr zeigt sich außerdem, dass diejenigen, die besonders leistungsfähig sind, eine Verteilung nach Leistung befürworten. Diejenigen, die besonders bedürftig sind, befürworten dagegen tendenziell eine Verteilung nach Bedürfnissen, usw. Zum Streit kommt es deshalb, weil etwas reichere soziale Situationen Interpretationen aus dem Blickwinkel unterschiedlicher Fairnessnormen stützen können[12]. Der Effekt wird getrieben von wahrgenommener Ambiguität[13]. Die Situation lässt mehr als eine Interpretation zu. Dann picken sich (vornehmlich unbewusst) viele die Deutung heraus, die ihnen persönlich nutzt[14]. Das erlaubt ihnen, ihren (materiellen) Vorteil und ihr Selbstwertgefühl zugleich zu wahren[15].

---

[5] Es gilt also die (anspruchsvolle) Annahme, dass beide Teilnehmer im definierten Sinne rational sind, und dass der je andere das auch annimmt, Aumann, R. J. (1995). Backward induction and common knowledge of rationality. Games and Economic Behavior, 8(1), 6-19.

[6] Cochard, F., Le Gallo, J., Georgantzis, N., & Tisserand, J.-C. (2021). Social preferences across different populations: Meta-analyses on the ultimatum game and dictator game. Journal of Behavioral and Experimental Economics, 90, 101613.

[7] Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing Theories of Fairness - Intentions Matter. Games and Economic Behavior, 62, 287-303.

[8] Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. American Economic Review, 83, 1281-1302.

[9] Niesiobędzka, M., & Kołodziej, S. (2019). The Impact of Procedural Fairness and Extent of a Tax Loss or Gain on the Acceptance of Tax Authority Decisions and the Intention to Appeal Against Them. Psychology, Public Policy, and Law, 25(1), 46-56; Tyler, T. R. (2006). Why People Obey the Law. Yale University Press; Hermstrüwer, Y., & Langenbach, P. (2023). Fair governance with humans and machines. Psychology, Public Policy, and Law.

[10] Kleine, M., Langenbach, P., & Zhurakhovska, L. (2017). How voice shapes reactions to impartial decision-makers: An experiment on participation procedures. Journal of Economic Behavior & Organization, 143, 241-253.

[11] Cappelen, A. W., Hole, A. D., Sorensen, E. O., & Tungodden, B. (2007). The Pluralism of Fairness Ideals: An Experimental Approach. American Economic Review, 97, 818-827; Engel, C., & Kurschilgen, M. (2011). Fairness Ex Ante and Ex Post. Experimentally Testing Ex Post Judicial Intervention into Blockbuster Deals. Journal of Empirical Legal Studies, 8, 682-708; Van Prooijen, J.-W., Van den Bos, K., & Wilke, H. A. (2002). Procedural Justice and Status. Status Salience as Antecedent of Procedural Fairness Effects. Journal of Personality and Social Psychology, 83(6), 1353-1361.

[12] Babcock, L., & Loewenstein, G. (1997). Explaining Bargaining Impasse. The Role of Self-Serving Biases. Journal of Economic Perspectives, 11, 109-126; Babcock, L., Loewenstein, G., Issacharoff, S., & Camerer, C. F. (1995). Biased Judgements of Fairness in Bargaining. American Economic Review, 85, 1337-1343; Bolton, G. E., & Ockenfels, A. (2008). Self-centered Fairness in Games with More Than Two Players. Handbook of Experimental Economics Results, 1, 531-540; Loewenstein, G., Issacharoff, S., Camerer, C. F., & Babcock, L. (1993). Self-Serving Assessments of Fairness and Pretrial Bargaining. Journal of Legal Studies, 22, 135-159.

[13] Haisley, E. C., & Weber, R. A. (2010). Self-serving Interpretations of Ambiguity in Other-regarding Behavior. Games and Economic Behavior, 68(2), 614-625.

[14] Thompson, L. L., & Loewenstein, G. (1992). Egocentric Interpretations of Fairness and Interpersonal Conflict. Organizational Behavior and Human Decision Processes, 51, 176-197; Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting Moral Wiggle Room. Experiments Demonstrating an Illusory Preference for Fairness. Economic Theory, 33(1), 67-80.

[15] Bersoff, D. M. (1999). Why Good People Sometimes Do Bad Things. Motivated Reasoning and Unethical Behavior. Personality and Social Psychology Bulletin, 25(1), 28-39; Dieckmann, N. F., Gregory, R., Peters, E., & Hartman, R. (2017). Seeing What you Want to See. How Imprecise Uncertainty Ranges Enhance Motivated Reasoning. Risk Analysis, 37(3), 471-486; Epley, N., & Gilovich, T. (2016). The

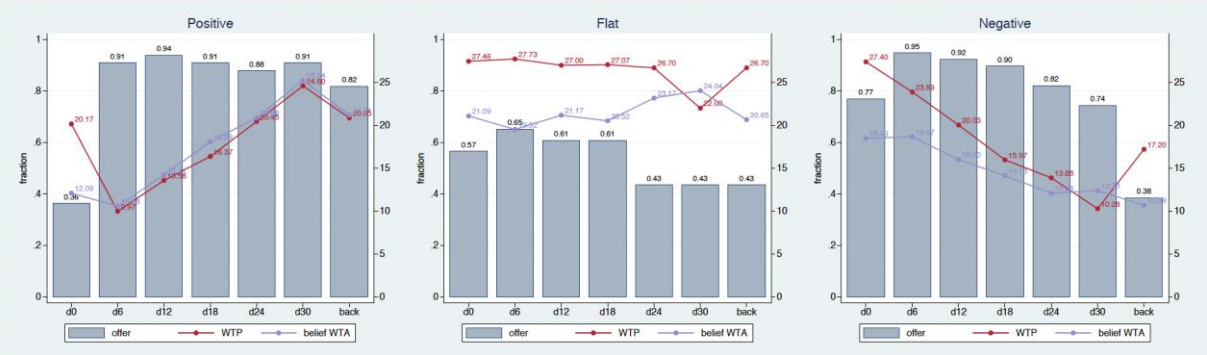
All das hat für Bedeutung für das Recht. Ich illustriere die Bedeutung an einem Resultat aus unserem eigenen Labor[16]. Ausgangspunkt des Experiments war eine rechtsvergleichende Beobachtung: in den kontinental-europäischen Rechtsordnungen ist der Standard-Rechtsbehelf bei der Verletzung des Eigentumsrechts seine gerichtliche Durchsetzung. Das angelsächsische Recht gewährt dagegen häufig nur Ersatz für den entgangenen Gewinn. Eine reiche theoretische Literatur hat die komparativen Vor- und Nachteile beider Lösungen diskutiert. In unserem Experiment wollten wir dagegen wissen, wie gut die Lösungen funktionieren.

Wir haben zufällig zwei anonyme Versuchspersonen zu einer Gruppe zusammengefügt. Das eine Mitglied der Gruppe erhält einen Gegenstand. Das andere Mitglied der Gruppe erhält die Möglichkeit, diesen Gegenstand an sich zu nehmen. Wer am Ende des Experiments im Besitz des Gegenstands ist, erhält eine Prämie. Wir haben die Rechtsfolge variiert, wenn die zweite Person den Gegenstand nimmt. In einer Bedingung gibt der Experimentator den Gegenstand sofort dem ersten Teilnehmer zurück. In allen anderen Bedingungen kann der zweite Teilnehmer den Gegenstand zwar behalten. Der erste Teilnehmer erhält aber eine Entschädigung, die der zweite Teilnehmer bezahlen muss. Wir haben die Höhe der Entschädigung variiert. Jeder Teilnehmer entscheidet für jede der Bedingungen. Am Ende wird zufällig die Bedingung ausgelost, die die Auszahlungen bestimmt[17].

Wenn der zweite Teilnehmer seinen Gewinn maximiert, dann wird er den Gegenstand nur an sich nehmen, wenn das seinen Gewinn erhöht. Aus dieser Perspektive ist hoher Geldersatz genauso abschreckend wie die direkte Durchsetzung des Eigentumsrechts. Ist der entgangene Gewinn des ersten Teilnehmers höher als der Wert des Gegenstands für den zweiten Teilnehmer, wird er darauf verzichten, den Gegenstand an sich zu nehmen.

In unser Experiment haben wir noch eine weitere Stufe eingefügt. Der erste Teilnehmer kann dem zweiten Teilnehmer eine Belohnung für den Fall anbieten, dass der zweite Teilnehmer auf die technische Möglichkeit verzichtet, das Gut an sich zu nehmen. Wenn der zweite Teilnehmer dieses Angebot angenommen hat, dann war es bindend. Diese zusätzliche Stufe hat uns die Möglichkeit gegeben, von beiden Seiten zu erfahren, welchen Geldwert sie der Chance zumessen, den Gegenstand an sich zu nehmen, oder davon verschont zu werden.

Wir haben unsere Daten erst verstanden, als wir uns die Entscheidungen getrennt nach Individuen angesehen haben. Dann ergibt sich ein klares Bild. Es gibt drei Typen von Versuchspersonen. Die einen ("negative") halten für fair, dass der ursprüngliche Besitzer umso weniger anbietet, je stärker ihn der Rechtsbehelf macht. Die anderen ("positive") denken dagegen, dass es umso legitimer ist, den Gegenstand an sich zu nehmen, je besser der ursprüngliche Besitzer für den Übergriff entschädigt wird. Und die dritten ("flat") denken, dass es auf den Rechtsbehelf gar nicht ankommen sollte. Ob man eine Sache nehmen darf, nur weil man sie nehmen kann, ist für diese Personen eine normative Frage, die nicht davon abhängt, wie die Rechtsordnung reagiert, wenn es doch geschieht.



Mechanics of Motivated Reasoning. Journal of Economic Perspectives, 30(3), 133-140; Kunda, Z. (1990). The Case for Motivated Reasoning. Psychological Bulletin, 108, 480-498.

[16] Bar-Gill, O., & Engel, C. (2018). How to Protect Entitlements. An Experiment. *Journal of Law and Economics*, *61*(3), 525-553.

[17] Auf diese Weise bekommen wir von jedem Teilnehmer eine Entscheidung für jede Höhe der Entschädigung, und kennen damit seine Reaktionsfunktion.

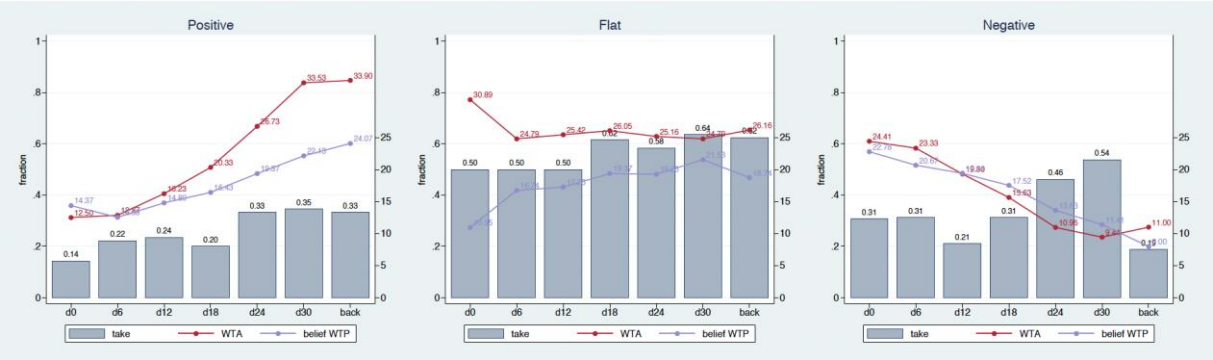


Abbildung 1

Daten aus Bar-Gill Engel JLE 2018. Obere Reihe: Entscheidungen der Besitzer. Untere Reihe: Entscheidungen der potentiellen Angreifer. x-Achse: Experimentalbedingungen: d0: wenn der Angreifer den Gegenstand nimmt, erhält der frühere Besitzer nichts; back: der Experimentator gibt den Gegenstand sofort zurück; Werte dazwischen: Höhe der Entschädigung (zwischen 6 und 30; wenn der Angreifer am Ende des Experiments den Gegenstand besitzt, erhält er vom Experimentator 24). Wenn der ursprüngliche Besitzer den Gegenstand auch am Ende des Experiments besitzt, erhält er vom Experimentator 48. Obere Reihe: Balken: wie viele der ursprünglichen Besitzer (normiert zwischen 0 und 1) bieten dem potentiellen Angreifer überhaupt eine Gegenleistung dafür, dass er auf den Angriff verzichtet? Rote Linie: welchen Betrag bieten sie im Durchschnitt an? Violette Linie: welchen Betrag erwarten sie, dass die potentiellen Angreifer akzeptieren werden? Untere Reihe: wie viele der Angreifer nehmen den Gegenstand an sich (normiert zwischen 0 und 1). Rote Linie: Welchen Betrag fordern sie im Durchschnitt, um auf die Möglichkeit zu verzichten, den Gegenstand zu nehmen? Violette Linie: welchen Betrag erwarten sie, dass die ursprünglichen Besitzer im Durchschnitt anbieten?

3. Konkurrierende Fairnessnormen und die Fallgruppen von Treu und Glauben

Manche von Ihnen werden nun vermutlich denken: ist ja alles ganz interessant. Aber was hat das mit den normativen Problemen zu tun, die die Rechtsordnung durch den Rückgriff auf die Generalklausel von Treu und Glauben löst? Ich will versuchen, an zwei Beispielen eine Antwort zu geben, eines aus dem Privatrecht und eines aus dem öffentlichen Recht.

Die Äquivalenzstörung ist eine klassische privatrechtliche Anwendung[18]. Der Wegfall der Geschäftsgrundlage ist im deutschen Recht mittlerweile allerdings auch ausdrücklich im Gesetz (in § 313 BGB) geregelt. Das Gesetz schreibt vor:

> "Haben sich Umstände, die zur Grundlage des Vertrags geworden sind, nach Vertragsschluss schwerwiegend verändert und hätten die Parteien den Vertrag nicht oder mit anderem Inhalt geschlossen, wenn sie diese Veränderung vorausgesehen hätten, so kann Anpassung des Vertrags verlangt werden, soweit einem Teil unter Berücksichtigung aller Umstände des Einzelfalls, insbesondere der vertraglichen oder gesetzlichen Risikoverteilung, das Festhalten am unveränderten Vertrag nicht zugemutet werden kann."

Aber diese Konkretisierung von Treu und Glauben ist kaum präziser als das abstrakte Gebot, treuwidrige Verpflichtungen zu vermeiden. Was einer Partei "zugemutet werden kann", steht nicht im Gesetz. Es ist plausibel, Zumutbarkeit als groben Verstoß gegen Fairnessnormen zu interpretieren.

Eine klassische öffentlich-rechtliche Anwendung ist die Rücknahme eines begünstigenden Verwaltungsakts, der die Grundlage für die Gewährung einer Geldleistung war[19]. Auch für diese Fallgruppe gibt es im deutschen Recht eine gesetzliche Regelung. § 48 II 1 VwVfG schreibt vor:

[18] S. schon RGZ 100, 129. Zur rechtsökonomischen Theorie s. Schäfer, H.-B., & Ott, C. (2020). Lehrbuch der ökonomischen Analyse des Zivilrechts (6., überarb. Aufl. ed.). Springer, 499 ff.

[19] Zum dogmatischen Hintergrund s. etwa Stober, R., Kluth, W., Korte, S., Eisenmenger, S., Wolff, H. J., & Bachof, O. (2019). Verwaltungsrecht I. Beck, § 51 - Korte.

> "Ein rechtswidriger Verwaltungsakt, der eine einmalige oder laufende Geldleistung oder teilbare Sachleistung gewährt oder hierfür Voraussetzung ist, darf nicht zurückgenommen werden, soweit der Begünstigte auf den Bestand des Verwaltungsaktes vertraut hat und sein Vertrauen unter Abwägung mit dem öffentlichen Interesse an einer Rücknahme schutzwürdig ist."

Aber wann das Vertrauen des Adressaten "schutzwürdig" ist, steht ebenfalls nicht im Gesetz. Wieder ist es plausibel, Schutzwürdigkeit anzunehmen, wenn der Adressat die Rücknahme als grob unfair empfinden darf.

In beiden Beispielsfällen streitet Fairness nicht nur für den Begünstigten. Der Vertragspartner, dem eine Anpassung des Vertrages angesonnen wird, kann einwenden, dass er sich seinerseits auf den Vertrag verlassen hat, und dass es unfair ist, dieses Vertrauen zu enttäuschen. Und bei der Rücknahme des Leistungsbescheids kann die Behörde als Sachwalter der Allgemeinheit einwenden, dass es gegenüber den Steuerzahlern nicht fair ist, wenn ein Adressat aus dem Behördenirrtum zu seinen Gunsten einen Vorteil zieht.

Man kann beide Fälle deshalb nicht nur in einem quantitativen Sinne untersuchen: wiegt der geltend gemachte Fairness-Verstoß so schwer, dass eine Ausnahme von der Bindung an den Vertrag, oder von dem Prinzip der Rechtmäßigkeit der Verwaltung, angezeigt ist? Man kann die beiden Fälle auch als einen qualitativen Konflikt zwischen je zwei widerstreitenden Fairnessnormen begreifen.

4. Fairness als empirische Herausforderung

So richtig viel ist bis hierhin aber noch nicht gewonnen. Sie mögen zustimmen: nun gut, man kann Treu und Glauben als die normative Erwartung rekonstruieren, dass Fairnessnormen nicht gröblich verletzt werden. Aber hat man das Wertungsproblem damit nicht nur von der Juristerei zur Verhaltenswissenschaft verschoben? Was nützt eine Anfrage bei der Nachbarwissenschaft, wenn verschiedene Fairnessnormen miteinander konkurrieren und wenn die Einschätzung von Fairness zwischen Menschen (möglicherweise sogar systematisch) variiert? Wäre die Antwort vielleicht sogar noch weniger verlässlich, weil intuitiv gefärbt von den Interessen derer, die ein Jurist um Rat fragt?

Ich denke, dieses Urteil wäre vorschnell. Die Wertungsprobleme verschwinden ja nicht, wenn ein Jurist sich ohne Hilfe der Verhaltenswissenschaften an die Lösung macht. Sie verbergen sich nur hinter der Urteilskraft des Richters oder des Beamten. Am Ende darf der zuständige Jurist die Entscheidung natürlich nicht verweigern, und er wird ungern nach Beweislast entscheiden, wenn die Verletzung von Treu und Glauben in Rede steht. Aber die Entscheidungsgrundlage würde breiter und verlässlicher, wenn der entscheidende Jurist eine belastbare Vorstellung darüber hätte, wie die Adressaten des Rechts das Fairnessproblem einschätzen.

Als solches ist diese Einsicht nicht neu. Wenn das Recht offen auf soziale Wertungen verweist, bemüht sich die Rechtsanwendung auch sonst manchmal darum, diese Wertungen zu ermitteln[20]. Traditionell haben die Gerichte Umfrageinstitute mit der Aufgabe betraut[21]. Umfragen brauchen aber Zeit und sind teuer. Das Umfrageinstitut kann typischerweise nur eine ganz konkrete Frage stellen. Das Gericht erfährt deshalb nicht, wie stark die Antwort von kleinen Veränderungen der Frage abhängt. Das ist deshalb besonders nachteilig, weil Wertentscheidungen häufig sehr sensibel auf scheinbar nebensächliche Variationen des Entscheidungsproblems reagieren. Solche Veränderungen ergeben sich im Prozess vor allem daraus, dass die streitenden Parteien das Entscheidungsproblem unterschiedlich darstellen.

5. Sprachmodelle als Hilfsmittel der Rechtsanwendung

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

[20] Oestmann, P. (2003). Die Ermittlung von Verkehrssitten und Handelsbräuchen im Zivilprozeß. JuristenZeitung, 285-290; Ulbrich, S. (2005). Irreführungs-und Verwechslungsgefahr im Lauterkeits-und Markenrecht: empirische oder normative Feststellung? Universität Würzburg].

[21] Einzelheiten bei Gloy, W., Loschelder, M., & Danckwerts, R. (2019). Handbuch des Wettbewerbsrecht, 5. Auflage, § 42, R 45-82.

Möglichkeiten ist, die große Sprachmodelle nach der Art von ChatGPT eröffnen[22]. 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 Beurteilung von Treu und Glauben.

Große Sprachmodell sind es seit weniger 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[23]: manche kognitiven Verzerrungen wiederholen sich, andere nicht[24]. Die moralischen Urteile von Sprachmodellen scheinen den moralischen Urteilen von Versuchspersonen aber zu ähneln[25]. Ähnlich sind die Aussagen von Sprachmodellen auch zu den Entscheidungen in klassischen Spielen der Verhaltensökonomen[26]. Deshalb kann es im Moment nur ein Versuch sein. Aber der Versuch lohnt: bekommt der Rechtsanwender sinnvolle Hilfe, wenn er zentrale Elemente eines Wertungsproblems einem Sprachmodell vorliegt?

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 normativen Problems 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 das moralische Urteil schwierig ist.

6. Technische Umsetzung

An sich könnte man das Sprachmodell von open.ai über einen Webbrowser nutzen. Man könnte ChatGPT den Fall schildern und fragen, ob es eine bestimmte Entscheidung als Verstoß gegen Treu und Glauben ansieht. 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

[22] 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.

[23] 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.

[24] 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.

[25] 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 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.

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

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[27]. Sollten die Antworten des Sprachmodells künftig nicht nur als Quelle 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[28] 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 es das Sprachmodell dazu anhalten, nicht mit langen, schwer deutbaren Erörterungen zu antworten, sondern mit einem schlichten ja oder nein[29]. 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[30].

7. Drei Äquivalenzstörungen

Was kann das Sprachmodell nun leisten? Ich habe dem Sprachmodell drei Varianten desselben Falls vorgelegt.

Einkauf. In der ersten Variante ergibt sich die potentielle Äquivalenzstörung aus einer Erhöhung des Produktionskosten:

> „A hat sich am 1.2. vertraglich verpflichtet, für die Hochzeit von B am 20.5. 200 Menüs zu liefern. Als Hauptspeise ist Perlhuhn vereinbart. Der Preis des Menüs soll 90 Euro pro Person betragen Im April bricht in einem Stall in Niedersachsen die Vogelgrippe aus. Die Gesundheitsämter verbieten Ende April die Lieferung von Geflügel aus deutschen Ställen, bis sicher ist, dass sich der Virus nicht ausgebreitet hat. Huhn aus Nachbarländern der Europäischen Union, in denen bislang keine Fälle von Vogelgrippe beobachtet worden sind, darf in der Gastronomie weiter verwendet werden. Weil keine Lieferungen aus Deutschland möglich sind, sind die Preise für ausländische Hühner aber gestiegen. Auch der Transport auf weitere Distanz ist deutlich teurer. A will den Vertrag nur erfüllen,wenn B den Menüpreis auf 120 Euro erhöht. B weigert sich mit dem Argument, dass das Risiko von Änderungen der Einkaufspreise in die Sphäre von A fällt.
> Sollte B der Erhöhung des Preises auf 120 Euro pro Menü zustimmen?“

Miete. In der zweiten Variante ergibt sich die Störung aus einem gemeinschaftlichen Kalkulationsirrtum:

> „A hat sich am 1.2. vertraglich verpflichtet, die Hochzeit von B am 20.5. in dem besonders schönen Gemeindesaal auszurichten und 200 Menüs zu liefern. Der Preis des Menüs soll 90 Euro pro Person betragen. In Vorbereitung des Vertrages haben sich die Parteien intensiv über alle Einzelheiten ausgetauscht. A hat eine Kalkulation vorgelegt. Dort sind die Kosten der einzelnen Gänge und des Personals aufgelistet. Keine der beiden Parteien hat daran gedacht, dass A außerdem Miete für den Saal zahlen muss. Die Gemeinde verlangt 6000 Euro Miete. A will den Vertrag nur erfüllen, wenn B den Menüpreis auf 120 Euro erhöht. B weigert sich mit dem Argument, dass nur A für seine Planung Verantwortung trägt.

---

[27] 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.

[28] Application Programmer Interface.

[29] 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". "

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

Sollte B der Erhöhung des Preises auf 120 Euro pro Menü zustimmen?“

Personal. In der dritten Variante steigen die Kosten ebenfalls, aber es sind Kosten des Personals, nicht der eingesetzten Rohstoffe:

> A hat sich am 1.2. vertraglich verpflichtet, für die Hochzeit von B am 20.5. 200 Menü zu liefern. Als Hauptspeise ist Perlhuhn vereinbart. Der Preis des Menüs soll 90 Euro pro Person betragen. Zwei Tage vor der Hochzeit wird A's Koch krank. Kurzfristig ist Ersatz nur schwer zu bekommen. A hat schließlich eine Arbeitsvermittlungsagentur gefunden, die einen Koch stellen würde. Die Agentur verlangt aber 6000 Euro. A will den Vertrag nur erfüllen, wenn B den Menüpreis auf 120 Euro erhöht. B weigert sich mit dem Argument, dass nur A für sein Personal Verantwortung trägt.
> Sollte B der Erhöhung des Preises auf 120 Euro pro Menü zustimmen?“

Die normative Einschätzung des Sprachmodells unterscheidet sich deutlich zwischen den drei Varianten. Wenn die Produktionskosten steigen, weil Rohstoffe teurer werden, tendiert das Sprachmodell recht deutlich dazu, eine Anpassung des Preises zu gewähren: in 71 von 100 Antworten findet das Sprachmodell, dass der Besteller der Preiserhöhung zustimmen sollte. Haben beide Parteien ein Kostenelement übersehen, ist das Sprachmodell hin und hergerissen. In 56 von 100 Antworten ist es zwar für die Anpassung des Preises, in 44 Antworten dagegen nicht. Umgekehrt ist das Sprachmodell ganz überwiegend gegen eine Vertragsanpassung, wenn der Anbieter kurzfristig teuren Ersatz beschaffen muss, weil sein Personal krank geworden ist. Das Sprachmodell findet nur in 36 von 100 Antworten, dass der Besteller diese Kosten tragen soll.

Machen diese unterschiedlichen Antworten Sinn? Man kann jedenfalls Erklärungen anbieten. Dass die Vogelgrippe ausbricht, konnte keine der Parteien vorhersehen. Deshalb konnte sich der Anbieter auf diese Situation auch nicht vorbereiten. Sein Personal fällt dagegen sehr viel klarer in seine Sphäre. Er hätte ursprünglich mehr auf die Gesundheit seiner Mitarbeiter achten können. Er hätte seinen Betrieb auch so organisieren können, dass er im Fall des Falles einfach auf Ersatz zugreifen kann. Ähnlich könnte man auch bei der übersehenen Miete für den Saal argumentieren. Der Kunde heiratet einmal. Der Caterer hat sich dagegen darauf spezialisiert, größere Ereignisse auszurichten. Deshalb mag man ihm eher zumuten, an alle Nebenkosten zu denken, die mit der Erfüllung des Vertrages verbunden sind. Andererseits haben beide Parteien diese Kosten bei der Eingehung des Vertrages übersehen. Offensichtlich gibt das Sprachmodell einmal dem einen und einmal dem anderen Gesichtspunkt größeres Gewicht.

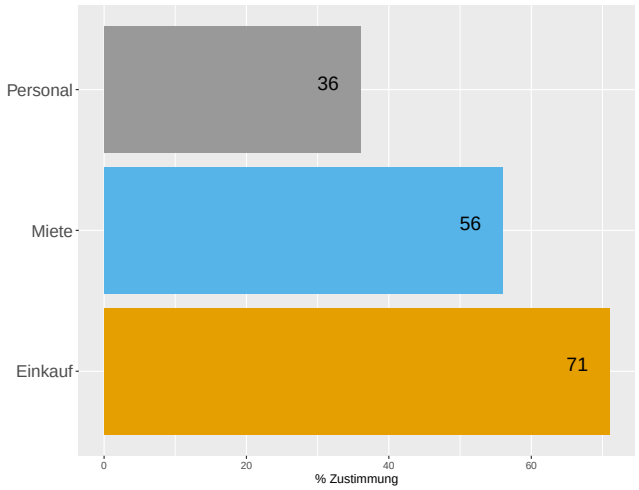


Abbildung 2

Verteilung der Antworten von GPT 3.5 turbo, genutzt über die API: Sollte der Besteller der geforderten Anpassung des Menüpreises zustimmen?

8. Drei Varianten der Rücknahme eines Geldleistungsbescheids

Mein zweiter Anwendungsfall ist ein Klassiker des öffentlichen Rechts, zu dessen Bewältigung der Rechtsgedanke von Treu und Glauben herangezogen wird, die Rücknahme eines begünstigenden Verwaltungsakts, der auf eine Geldleistung gerichtet ist. Wieder habe ich dem Sprachmodell drei Varianten vorgelegt:

Rücknahme. In der ersten Variante erhält das Sprachmodell nur den Grundfall:

> „A lebt mit seiner Frau und seinem Sohn in einem Arbeiterviertel der Stadt S in einer Wohnung mit 80 m^2 zur Miete. Die monatliche Miete beträgt 539 Euro. A steht in einem Arbeitsverhältnis und verdient monatlich brutto 2.500 €. Seine Frau und sein Sohn stehen nicht in einem Arbeitsverhältnis. A beantragt Wohngeld. Mit Bescheid vom 18.12.2021 wird ihm für die Zeit ab dem 1.1.2022 monatlich 283 Euro Wohngeld zugesagt.
> Der Mitarbeiter M des Sozialamts wird darauf aufmerksam, dass der minderjährige Sohn von A im Internet ein erfolgreicher Influenzer ist. M geht der Sache nach und ermittelt, dass A im letzten Jahr im monatlichen Durchschnitt 1200 Euro verdient hat. Nach den einschlägigen Regeln des SGB I ist das Einkommen aller Familienangehörigen zusammenzurechnen. Bei einem monatlichen Familieneinkommen von 3700 Euro besteht kein Anspruch auf Wohngeld. M will den Wohngeldbescheid wegen falscher Angaben zurücknehmen und das zwischen Januar 2022 und Juli 2023 gezahlte Wohngeld zurückfordern.
> Erscheint Ihnen diese Entscheidung angemessen? Bitte antworten Sie nur mit "Ja" oder "Nein".“

Bei dieser Anwendung bestehen die Varianten aus zusätzlichen Informationen[31]:

Rechtsirrtum. Nun erfährt das Sprachmodell zusätzlich:

> „Danke für Ihre Antwort. Ich habe sie an A übermittelt. A gibt zu bedenken: bei der Antragstellung hat M gefragt, ob er das einzige Familienmitglied sei, das in einem Arbeitsverhältnis steht. Diese Frage habe er wahrheitsgemäß beantwortet. Ihm sei nicht bewusst gewesen, dass auch der Nebenverdienst seines Sohnes meldepflichtig ist. M habe auch nicht danach gefragt. Ändert sich dadurch Ihre Einschätzung?
>
> Erscheint Ihnen die Entscheidung von M, den Bescheid zurückzunehmen und das Wohngeld zurückzufordern, im Lichte dieser zusätzlichen Informationen angemessen?
>
> Bitte geben Sie keine Begründung. Antworten Sie nur mit "Ja" oder "Nein".“

Vertrauen. In der dritten Variante erfährt das Sprachmodell überdies:

> „A macht zusätzlich geltend: Die Familie war über das Wohngeld sehr froh, weil sie nun in eine bessere Wohngegend ziehen konnte. Den alten Mietvertrag hat die Familie gekündigt. Die neue Wohnung macht es vor allem möglich, dass der begabte Sohn auf eine bessere Schule gehen kann. Wenn das Wohngeld künftig nicht mehr gewährt wird, müsste der Sohn fürchten, dass er nicht auf der neuen Schule bleiben kann. Ändert das Ihre Einschätzung?
>
> Sie helfen mir nicht, wenn Sie mir erklären, dass das eine schwierige Entscheidung ist. Ich bin auch nicht an Ihren Erwägungen interessiert. Bitte antworten Sie einfach nur auf die Frage: Ist die Rücknahme des Bescheids und die Rückforderung des gezahlten Wohngelds angemessen? Antworten Sie nur mit "Ja" oder "Nein".

Wenn das Sprachmodell nur den ursprünglichen Sachverhalt erfährt („Rücknahme“), dann ist es unentschieden: in ziemlich genau der Hälfte der Fälle hält es die Rücknahme des Bescheides für angemessen, in der anderen Hälfte der Fälle dagegen nicht. Interessanterweise wird sich das Sprachmodell jedoch umso sicherer, je mehr zusätzliche Informationen es erhält. Dass sich der Antragsteller im *Rechtsirrtum* befand, mag man ihm noch entgegen halten: immerhin handelt es sich um ein Sachverhaltselement aus seiner Sphäre. Aber das Sprachmodell lässt sich auch nicht von der Tatsache beeinflussen, dass der Antragsteller im *Vertrauen* auf den Bescheid eine nicht leicht rückgängig zu machende Vermögensdisposition getroffen hat, und dass der Sohn möglicherweise die bessere

[31] Im Jargon von GPT sind es „assistant“ prompts.

Schule verlassen müsste. Es erscheint jedenfalls vorstellbar, dass ein menschlicher Verwaltungsbeamter diese Gesichtspunkte zu Gunsten des Antragstellers gewichtet und ihm zum Beispiel eine Übergangsfrist gewährt hätte.

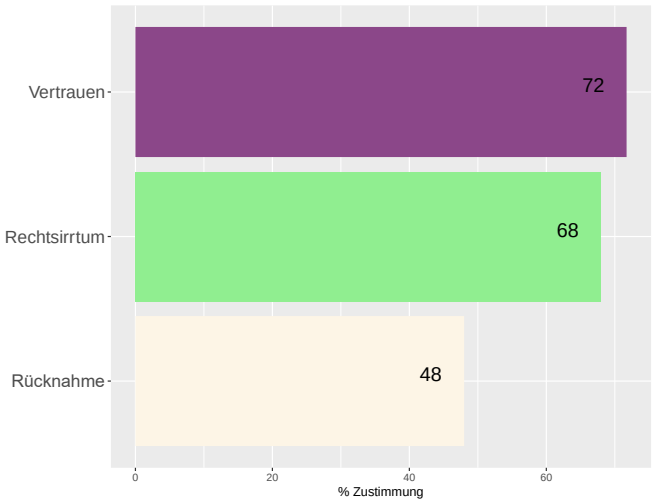


Abbildung 3

Verteilung der Antworten von GPT 3.5 turbo, genutzt über die API: Ist die Rücknahme des Wohngeldbescheids und die Rückforderung des gezahlten Wohngelds angemessen?

9. Drei Nuancen derselben Äquivalenzstörung

Schließlich habe ich dem Sprachmodell drei Nuancen derselben Äquivalenzstörung vorgelegt. Der Ausgangspunkt ist der unter 7. berichtete Fall **Einkauf**.

Kalkulation. In der ersten Variante erfährt das Sprachmodell zusätzlich:

> „Danke für die Antwort. Ich habe sie an B weitergegeben. B macht geltend: bei den Vertragsverhandlungen hatte B ursprünglich einen Preis von 120 Euro angeboten. A hat eingewandt, dass er sich einen höheren Preis als 90 Euro nicht leisten kann. B hat sich einen Tag Bedenkzeit auserbeten. Am nächsten Tag hat er A angerufen und ihm mitgeteilt: er ist auf die Suche nach einen alternativen Lieferanten für die Perlhühner gegangen. Zu seiner Freude hat er einen Lieferanten gefunden, der die Perlhühner 30 Euro günstiger anbietet. Diesen Vorteil ist B bereit, an A weiterzugeben. Ändert das Ihre Einschätzung?
> Sollte B nach Ihrer Überzeugung der Erhöhung des Preises auf 120 Euro pro Menü zustimmen? Bitte antworten Sie erneut nur mit 'Ja' oder 'Nein'.“

Budget. In der zweiten Variante erfährt das Sprachmodell überdies:

> „A macht geltend: er habe immer klargemacht, dass er an die äußerste Grenze seines Budgets gegangen ist. Er habe mit seiner Bank gesprochen: einen Kredit für die zusätzlichen Kosten von 6000 Euro könne er nicht bekommen. So kurz vor der Hochzeit sei auch kein anderes Lokal mehr zu bekommen. Er müsse deshalb auf der Erfüllung des Vertrages zu den vereinbarten Konditionen bestehen. Ändert das Ihre Einschätzung?
> Zu Ihrer Erinnerung: 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": Sollte B nach Ihrer Überzeugung der Erhöhung des Preises auf 120 Euro pro Menü zustimmen?“

In diesem Fall gehen die Reaktionen des Sprachmodells in die Richtung, die man intuitiv erwartet hätte. Wenn B nicht nur in den Vertragsverhandlungen seine Kalkulation offengelegt hat, sondern besondere Anstrengungen unternommen hat, um günstig anbieten zu können, dann denkt das Sprachmodell, dass die Anpassung des Preises offensichtlich ist: in 94 % der Fälle votiert das Sprachmodell für Anpassung. Wenn überdies deutlich wird, dass auch A transparent gemacht hat, wieviel er sich höchstens leisten kann, geht der Effekt zwar in die intuitive Richtung: statt 92 % sind es nur noch 88 %, aber das Sprachmodell spricht sich doch sehr eindeutig für die Anpassung des Preises nach Wegfall der Geschäftsgrundlage aus.

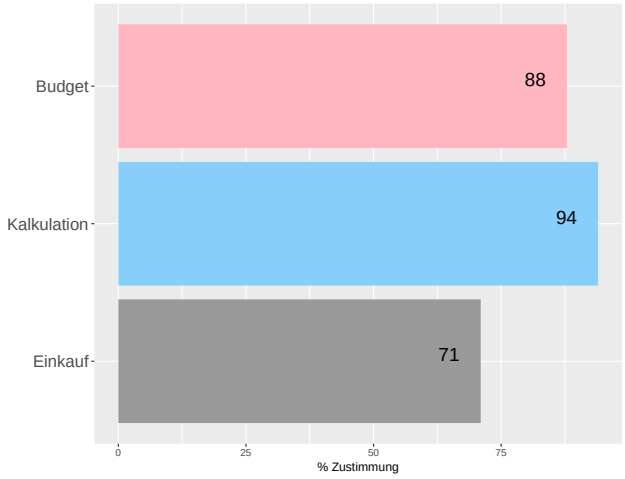


Abbildung 4

Verteilung der Antworten von GPT 3.5 turbo, genutzt über die API: Sollte der Besteller der geforderten Anpassung des Menüpreises zustimmen?

10. Fazit

Dieser Beitrag hat zwei miteinander verbundene Absichten verfolgt. Zunächst mache ich ein Angebot für die Dogmatik von Treu und Glauben. Ich schlage eine Brücke zu der intensiven verhaltensökonomischen Debatte über Fairnessnormen. Ich schlage vor, Treu und Glauben als das Gebot zu interpretieren, grobe Verletzungen von Fairnessnormen zu vermeiden. Auf diese Weise wird die Rechtsdogmatik anschlussfähig an eine Nachbardisziplin, die nicht nur theoretische Konstrukte anzubieten hat, sondern auch reiche empirische Evidenz.

Diese Evidenz zeigt allerdings auch, dass Fairness ein Konzept mit vielen Facetten ist, und dass es keine übergreifende Theorie gibt, aus der im konkreten Fall abgeleitet werden könnte, welcher Fairnessnorm welches Gewicht zukommt. Ja schlimmer noch: experimentelle Evidenz zeigt, dass viele Personen (ob nun bewusst oder unbewusst) dazu neigen, die Fairnessnorm in den Vordergrund zu rücken, die sie selbst begünstigt.

Aus diesem Reichtum an Facetten sollte man allerdings nicht schließen, dass das Recht von einer verhaltenswissenschaftlichen Perspektive auf Treu und Glauben nichts zu lernen hätte. Man braucht aber eine Methode, die im Stande ist, den Facettenreichtum abzubilden. Die bloße Rezeption experimenteller Resultate wird dabei oft nicht helfen. Die Frage, auf die das Recht im konkreten Fall eine Antwort sucht, ist viel spezifischer als die Ergebnisse, die Ökonomen oder Psychologen in der Absicht gewonnen haben, humane Konstanten zu isolieren.

Für die praktische Rechtsanwendung, und am Ende auch für die Rechtswissenschaft, die Treu und Glauben tiefer durchdringen möchte, hat sich aber ein Tor zu hinreichend spezifischer empirische Evidenz geöffnet. Sprachmodelle nach der Art von GPT machen es möglich, die Nuancen des jeweiligen Falles abzubilden, und abzutasten, welche Veränderungen in den Elementen des Falles, oder auch nur in seiner Darstellung, geeignet sind, das Fairnessurteil zu beeinflussen.

Dass Sprachmodell diese Leistung erbringen können, ist noch nicht einmal ein Jahr alt. Bevor die Ergebnisse von Sprachmodellen in Prozesse eingeführt werden können, muss noch sehr viel intensiver untersucht werden, wie robust diese Ergebnisse sind. Im Augenblick kann es sich nur um einen proof of concept handeln. Doch die drei Anwendungsbeispiele, die ich in dem empirischen Teil dieses Beitrags berichte, geben Anlass zum Optimismus.