Verfahrensfairness ohne Menschen? Evidenz aus Human- und LLM-Experimenten

Yoan Hermstrüwer & Alexander Stremitzer

1. Einleitung[1]

Algorithmische Entscheidungssysteme gewinnen in der Rechtsprechung zunehmend an Bedeutung (Stevenson & Doleac 2019; Kleinberg et al. 2018; Kehl et al. 2017; Barocas et al. 2019; Pasquale 2015). Trotz ihres Potenzials zur Verbesserung richterlicher Entscheidungen bleibt ihr Einsatz hochumstritten (Mayson 2018; Završnik 2020). Kritiker verweisen auf das Risiko, dass Algorithmen aufgrund geschützter Merkmale diskriminieren könnten (Dressel & Farid 2018; Corbett-Davies et al. 2017), und mahnen mangelnde Verfahrensfairness an (Re & Solow-Niederman 2019; Binns 2022). Diese Einwände sind sowohl für die Rechtsanwendung als auch für die Gewährleistung des Gleichbehandlungsgrundsatzes relevant (Hermstrüwer 2020), denn Verfahrensfairness fördert nachweislich die Gesetzesbefolgung (Tyler 1997; Tyler & Huo 2002).

Bereits heute kann KI in bestimmten juristischen Aufgaben mit menschlichen Experten mithalten oder sie übertreffen (Jung et al. 2020; Aletras et al. 2016; Katz et al. 2017). Gleichwohl zeigen Studien, dass KI-basierte Entscheidungen oft als weniger fair empfunden werden (Burton et al. 2020; Bigman & Gray 2018). Dieses Phänomen – die sogenannte Mensch-KI-Fairness-Lücke – entsteht im direkten Vergleich rein menschlicher mit rein algorithmischer Entscheidungsfindung (Chen et al. 2025; Dietvorst et al. 2015; Mahmud et al. 2022).

In der Praxis agieren Algorithmen allerdings selten ohne menschliche Aufsicht (Selbst 2021; Engstrom et al. 2020). Bestimmungen wie Art. 22 DSGVO und Art. 14(1) der EU-KI-Verordnung schreiben eine menschliche Beteiligung oder Überwachung von Hochrisiko-KI-Systemen vor (Wachter et al. 2017; Veale & Borgesius 2021). Die menschliche

1 Für hervorragende Assistenz bei der Durchführung des LLM-Experiments danken wir Cara Cerny.

Aufsicht dient dabei mehreren Zwecken: der Korrektur algorithmischer Fehler (Lehr & Ohm 2017), der Gewährleistung der Integrität des Begründungsprozesses (Liu & Li 2025) sowie der Wahrung von Autonomie und Würde der Betroffenen (Danaher 2016; Wischmeyer 2020).

Empirische Befunde zeigen, dass die Bereitstellung von Anhörungsrechten und Begründungen die Fairness-Lücke verringern kann (Chen et al. 2022; Henning & Langenbach 2024). Eine weitere Studie legt nahe, dass hybride Systeme, die algorithmische Unterstützung mit menschlicher Bewertung kombinieren, als fairer wahrgenommen werden als rein menschliche Entscheidungen (Hermstrüwer & Langenbach 2023). Unklar bleibt jedoch, wie intensiv die menschliche Überprüfung sein muss, um diesen Effekt zu erzielen.

In diesem Beitrag kombinieren wir die Befunde eines mit Menschen durchgeführten Vignetten-Experiments (Chen et al. 2025) mit zwei Replikationsstudien, in denen wir große Sprachmodelle (LLMs) als Probanden einsetzen. Im Humanexperiment untersuchen wir mit 4.817 US-Teilnehmern die Robustheit der Mensch-KI-Fairness-Lücke und analysieren, wie unterschiedliche Intensitäten menschlicher Aufsicht die Fairnesswahrnehmungen beeinflussen. In den LLM-Experimenten replizieren wir die Studie mit vier Modellen: Claude Sonnet 4, DeepSeek-V3, Gemini 2.5-Flash und GPT-4o.

Die Ergebnisse unseres Humanexperiments bestätigen die Existenz der Mensch-KI-Fairness-Lücke: Rein algorithmische Entscheidungen werden als weniger fair wahrgenommen als rein menschliche Verfahren. Zugleich stellen wir fest, dass hybride Prozesse – selbst bei minimaler menschlicher Beteiligung – die Lücke vollständig beseitigen können. Sie werden als ebenso fair angesehen wie rein menschliche Prozesse. Überraschenderweise erweist sich die bloße Tatsache menschlicher Überprüfung als wichtiger als deren Intensität.

Die Ergebnisse unserer beiden LLM-Experimente scheinen diesen Befund auf den ersten Blick zu bestätigen. Aggregiert man die Daten über alle Modelle und Entscheidungssituationen, ergibt sich qualitativ ein ähnliches Gesamtbild. Bei genauerer Betrachtung treten jedoch fundamentale Unterschiede zutage, die bedeutende methodische Schwierigkeiten offenbaren.

Der Gang der Untersuchung ist folgender: Zunächst beschreiben wir das experimentelle Design unseres Humanexperiments und unserer LLM-Experimente (Abschnitt 2). Es folgen eine Auswertung und Diskussion unserer Ergebnisse (Abschnitt 3). In der Schlussbetrachtung diskutieren

wir die rechtlichen und methodischen Implikationen unserer Befunde (Abschnitt 4).

2. Experimentelles Design

Wir führen drei Vignetten-Experimente durch: eines mit Menschen und zwei mit LLMs. Alle drei Experimente zielen darauf ab, die wahrgenommene Verfahrensfairness in drei verschiedenen Szenarien zu untersuchen.

Humanexperiment. Die Teilnehmer wurden über Prolific in den USA (N = 4.817) rekrutiert und zufällig einer von fünf Bedingungen zugeordnet, die mit einem von drei Szenarien gekreuzt wurden (5x3 between-subjects design). Aufbauend auf einer unserer Vorgängerstudien (Chen et al. 2022) las jeder Teilnehmer eine stilisierte Vignette über ein Szenario betreffend die Verhängung des Strafmaßes, den Erlass einer strafprozessualen Haftentscheidung oder einen Fall zur Verbraucherschiedsgerichtsbarkeit. Im Strafmaßszenario bewertet ein Richter die Schuld des Angeklagten John und entscheidet, ob eine härtere oder mildere Strafe zu verhängen ist. Im Haftszenario schätzt ein Richter die Flucht- und Wiederholungsgefahr des Beschuldigten John ein und entscheidet über die Freilassung gegen Kaution. Im Verbraucherschiedsgerichtsszenario stellt der Schiedsrichter fest, ob es einen Schmierfleck auf der Linse einer Kamera gibt, die der Verbraucher John auf einer Online-Plattform gekauft hat, und entscheidet, ob der Kaufpreis zu erstatten ist.

Wir testen fünf Bedingungen. In den beiden Bedingungen mit rein menschlicher Entscheidungsfindung, Human High und Human Low, wird die Entscheidung immer ausschließlich von einem Menschen getroffen. Die Bedingungen Human High und Human Low unterscheiden sich in der Intensität, mit welcher der menschliche Entscheidungsträger den Fall prüft und die Beweise sichtet. Diese Treatmentvariation berücksichtigt die begrenzte Zeit, die (Schieds-)Richter haben, um die Beweise zu bewerten. Asylanhörungen zum Beispiel dauern in der Regel nicht länger als einige Minuten (Political Asylum/Immigration Representation Project 2016); dasselbe gilt für Anhörungen im Haftverfahren (Stevenson 2018). Während der Mensch in der Human High-Bedingung immer „eine gründliche Prüfung des Falles" und eine „gründliche" Überprüfung der Beweise durchführt, führt er in der Human Low-Bedingung nur „eine kurze Prüfung des Falles" durch und überprüft die Beweise nur „kurz".

Am entgegengesetzten Ende des Mensch-KI-Spektrums implementieren wir eine Robot-Bedingung, in der die Entscheidung immer von einem Algorithmus getroffen wird. In dieser Bedingung führt ein Algorithmus eine „gründliche Prüfung des Falles" und eine „gründliche" Überprüfung der Beweise durch. In den zwei hybriden Bedingungen, Hybrid High und Hybrid Low, interagiert ein menschlicher Entscheidungsträger mit einem algorithmischen Entscheidungsassistenzsystem, das eine Empfehlung abgibt. Auch hier besteht der Unterschied zwischen den Hybrid High- und den Hybrid Low-Bedingungen in der Intensität, mit der der Mensch den Fall prüft und die Beweise sichtet (gründlich oder kurz). In jedem Szenario und jeder Bedingung messen wir die wahrgenommene Verfahrensfairness auf einer 7-Punkte-Likert-Skala mit sieben Labels („stimme überhaupt nicht zu" = 1, „stimme voll und ganz zu" = 7). Zusätzlich erheben wir Daten zur wahrgenommenen Ergebnisfairness, Gründlichkeit, Transparenz, Empathie, Genauigkeit und Effizienz.

LLM-Experimente. Können große Sprachmodelle menschliches Verhalten vorhersagen? Und inwiefern lassen sich große Sprachmodelle als Substitut menschlicher Probanden einsetzen? Zur Beantwortung dieser beiden Fragen replizieren wir unser Vignetten-Experiment mit mehreren LLMs (Claude Sonnet 4, DeepSeek-V3, Gemini 2.5-Flash und Open AI GPT-4o).[2] Über die jeweilige Schnittstelle (API) werden jede Bedingung und jedes Szenario in je 25 Iterationen in jedes der vier Modelle eingespeist. Alle Iterationen werden, soweit verfügbar, bei einer Temperatur von 0 und ohne festen Seed implementiert. Pro Experiment erheben wir 375 individuelle Beobachtungen je LLM; insgesamt umfasst unsere Stichprobe damit über beide Experimente hinweg 3.000 individuelle Beobachtungen.

Das erste LLM-Experiment zielt auf die Vorhersage (prediction) menschlicher Entscheidungen. Der entsprechende Prompt enthält neben der Aufforderung zur Prognose auch Informationen zur soziodemographischen Struktur des Probandenpools aus unserem Humanexperiment. Das zweite LLM-Experiment ist auf die Generierung eigener Fairnessbewertungen (decision) der unterschiedlichen Verfahren durch das jeweilige Modell gerichtet. Wir führen also ein Prediction-Experiment und ein Decision-Experiment durch.

2 Die Prompts und Python-Skripten stellen wir auf Anfrage gerne zur Verfügung.

3. Ergebnisse

Treatment-Effekte. Die Treatment-Effekte unseres Humanexperiments bekräftigen unseren früheren Befund: Roboter-Richter werden als weniger fair wahrgenommen als menschliche Richter (Abb. 1a). Es zeigt sich eine deutliche KI-Mensch-Fairness-Lücke. Paarweise Vergleiche der Bedingungen bestätigen, dass die Roboter-Bedingung zu signifikant niedrigeren Fairness-Bewertungen führt als die menschlichen Bedingungen (Wilcoxon-Rangsummentests, $p < 0.001$).

Kann diese Mensch-KI-Fairness-Lücke im richterlichen Kontext durch die Einführung menschlicher Aufsicht über den Roboter-Richter gemindert werden? Zunächst stellen wir fest, dass menschliche Beteiligung an der algorithmischen Entscheidungsfindung für die Verfahrensfairness wichtig ist. Diese ist in beiden Hybrid-Bedingungen höher als in der Robot-Bedingung (Wilcoxon-Rangsummentests, $p < 0.001$). Wir beobachten indes kaum Unterschiede in den Fairnessbewertungen zwischen den menschlichen und hybriden Bedingungen. Dies stimmt mit der Annahme überein, dass die Menschen empfindlich darauf reagieren, ob die richterliche Entscheidung ausschließlich von einem Algorithmus getroffen wird, nicht aber auf unterschiedliche Intensitäten menschlicher Beteiligung. Bereits ein hybrides Verfahren – also ein Verfahren, in dem

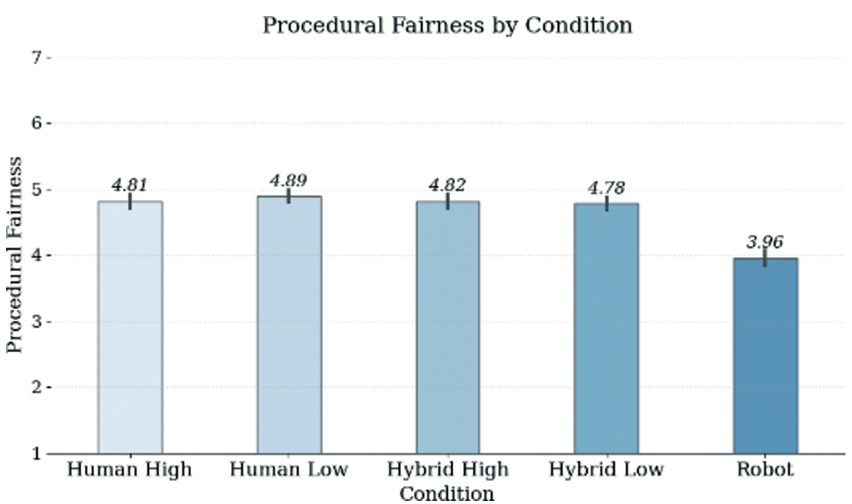

Abb. 1a: Prozedurale Fairnesswahrnehmungen von Menschen

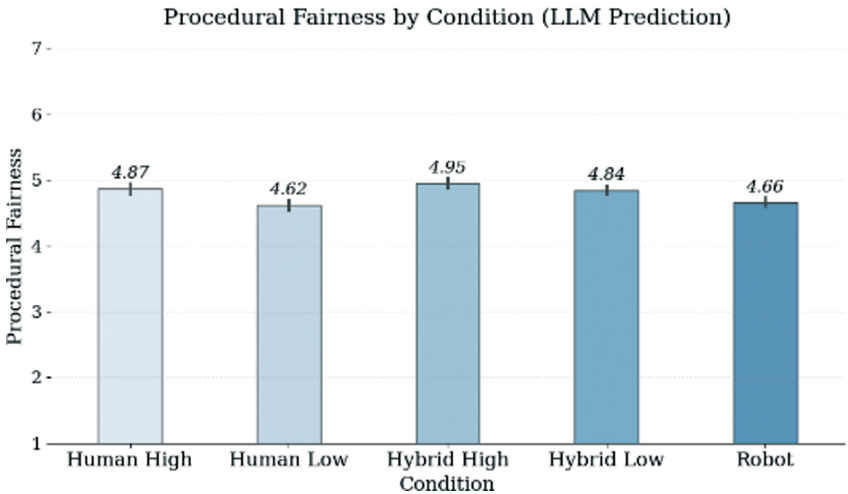

Abb. 1b: LLM-Vorhersagen

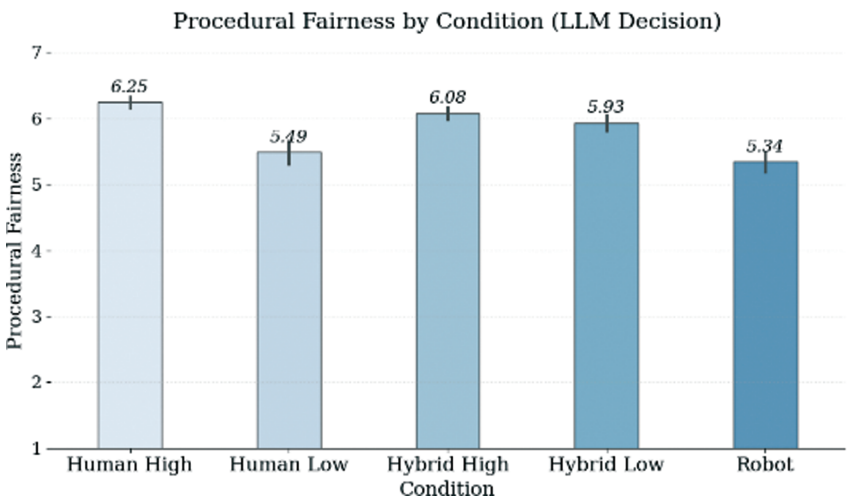

Abb. 1c: LLM-Entscheidungen

menschliche Aufsicht über den Algorithmus ausgeübt wird – erscheint ausreichend, um die Mensch-KI-Fairness-Lücke weitgehend zu schließen.

Zur Untersuchung dieser Hypothese analysieren wir, ob die Bewertungen der Verfahrensfairness zwischen den menschlichen und hybriden Bedingungen statistisch äquivalent sind, indem wir das TOST-Verfahren (Two One-Sided Tests) verwenden (Schuirman 1987). Wir setzen die Äquivalenzmarge auf $\Delta = 0.25$; unter der Annahme einer gepoolten Standardabweichung von 1.0 entspricht dies einer geringen Effektgröße von Cohen's $d = 0.25$ (vgl. Cohen 1988). Die Ergebnisse legen nahe, dass die wahrgenommene Verfahrensfairness für alle paarweisen Vergleiche der menschlichen und hybriden Bedingungen äquivalent ist. Dazu gehört der Befund, dass menschliche Verfahren mit hoher menschlicher Beteiligung (Human High) nicht höher bewertet werden als hybride Verfahren mit geringer menschlicher Beteiligung (Hybrid Low), was auf einen Ausgleichseffekt der KI-Fairness hindeutet. Der Übergang von rein menschlichen Verfahren mit einem hohen Maß an menschlicher Beteiligung zu hybriden Verfahren mit einem geringen Maß an menschlicher Beteiligung verringert die wahrgenommene Verfahrensfairness folglich nicht. Umgekehrt erhöht der Übergang von rein menschlichen zu hybriden Entscheidungsverfahren die wahrgenommene Verfahrensfairness aber auch nicht.

Schließlich bleibt die wahrgenommene Verfahrensfairness auch weitgehend unbeeinflusst davon, ob die menschliche Kontrollintensität hoch oder gering ausfällt. Diese Befunde lassen Zweifel an der Vorstellung aufkommen, eine Erhöhung der Intensität menschlicher Aufsicht wirke sich positiv auf die wahrgenommene Verfahrensfairness aus.

Die Ratio experimenteller Befunde ist kausale Inferenz. Doch auch wenn wir den Effekt unserer Bedingungen kausal identifizieren können, bleiben die dem Treatment-Effekt zugrundeliegenden psychologischen Mechanismen weitgehend im Dunkeln. Mithilfe einer Mediationsanalyse lässt sich der Gesamteffekt – definiert als Differenz der Fairnessbewertungen zwischen der Robot-Bedingung und den menschlichen Bedingungen – in direkte und kausal vermittelte Effekte zerlegen (Pearl 2001). Unter der Annahme kausaler Abhängigkeit zwischen Mediatoren schätzen wir ein lineares Strukturgleichungsmodell mit variablen Koeffizienten; Standardfehler werden per Bootstrap ermittelt (Imai und Yamamoto 2013).

In Übereinstimmung mit früheren Befunden (Chen et al. 2022) legt unsere Analyse nahe, dass die Mensch-KI-Fairness-Lücke primär auf Be-

denken hinsichtlich der Genauigkeit algorithmischer Entscheidungsfindung zurückzuführen ist. Die wahrgenommene Genauigkeit vermittelt etwa ein Viertel des Gesamteffekts, die Gründlichkeit etwa 15 %. Demgegenüber spielen „weiche" Faktoren eine untergeordnete Rolle: Die wahrgenommene Empathie macht nur etwa 8 % aus, die wahrgenommene Transparenz etwa 2 %. Die wahrgenommene Effizienz erweist sich als irrelevant für die Fairnessbewertungen. Die Annahmen der Mediationsanalyse sind streng, und ihre Befunde lassen kausale Identifikation nicht ohne Weiteres zu. Dennoch stützt unsere Analyse die These, dass die Mensch-KI-Fairness-Lücke besser durch „harte" Faktoren wie Genauigkeit als durch „weiche" Faktoren wie Empathie zu erklären ist.

Anders liegen die Dinge in unseren beiden LLM-Experimenten (Abb. 1b und 1c). Zwar beobachten wir auch hier die niedrigsten Fairnessbewertungen in der Robot-Bedingung. Allerdings erweist sich die Mensch-KI-Fairness-Lücke als deutlich geringer – und das Muster unterscheidet sich quantitativ und qualitativ von den Humandaten.

Im Prediction-Experiment (Abb. 1b) zeigen die LLMs ein bemerkenswert flaches Antwortprofil: Die Mittelwerte variieren lediglich zwischen 4.62 (Human Low) und 4.95 (Hybrid High) auf der 7-Punkte-Skala. Die Robot-Bedingung (M = 4.66) wird dabei sogar geringfügig fairer bewertet als Human Low (M = 4.62). Die höchste Fairnessbewertung erhält nicht etwa die Human High-Bedingung (M = 4.87), sondern Hybrid High (M = 4.95). Insgesamt scheinen LLMs bei der Vorhersage menschlicher Urteile die Intensität der Mensch-KI-Fairnesslücke systematisch zu unterschätzen.

Im Decision-Experiment (Abb. 1c) zeigt sich ein differenzierteres Bild. Hier bewerten die LLMs rein menschliche Verfahren mit hoher Prüfungsintensität am fairsten (Human High: M = 6.25), gefolgt von den hybriden Bedingungen (Hybrid High: M = 6.08; Hybrid Low: M = 5.93). Auffällig ist die Nähe zwischen Human Low (M = 5.49) und Robot (M = 5.34): Die Differenz beträgt lediglich 0.15 Skalenpunkte. Dieser Befund deckt sich mit den Ergebnissen einer unserer Vorgängerstudien, in der vollautomatisierte Verwaltungsverfahren nahezu gleichermaßen fair bewertet wurden wie Verfahren mit nur geringer menschlicher Beteiligung (Hermstrüwer & Langenbach 2023). Aus Sicht der LLMs scheint nicht die bloße Anwesenheit eines Menschen entscheidend, sondern die Intensität seiner Beteiligung – ein Muster, das sich in den Humandaten gerade nicht zeigt.

Zusammenfassend lässt sich festhalten: Während menschliche Pro-

banden bereits durch minimale menschliche Beteiligung (Hybrid Low) vollständig zufriedengestellt werden und nicht zwischen unterschiedlichen Intensitätsstufen differenzieren, reagieren LLMs sensibler auf Variationen in der Prüfungsintensität, unterschätzen aber zugleich die kategoriale Bedeutung, die Menschen der Mensch-Maschine-Unterscheidung beimessen.

Modellkonsistenz. Rechtsprobleme sind selten wohldefiniert (Engel 2001). Insbesondere Entscheidungen über das Strafmaß zeichnen sich durch hohe Komplexität und erhebliche Unsicherheit über die normativ angemessene Strafe aus. Vor diesem Hintergrund überrascht es nicht, dass sich Unterschiede zwischen den Entscheidungsverfahren gerade dort manifestieren, wo subjektive Bewertungen rechtlich relevanter Tatsachen eine zentrale Rolle spielen. Die modellspezifische Analyse offenbart allerdings erhebliche Heterogenität zwischen den untersuchten LLMs (Abb. 2a und 2b).

Claude Sonnet 4 zeigt im Prediction-Experiment (Abb. 2a) ein nahezu flaches Profil: Die Werte variieren lediglich zwischen 3.78 (Robot, Strafmaß) und 5.12 (Human High, Schiedsgericht). Das Modell prognostiziert nur minimale Treatment-Effekte – die vorhergesagte Mensch-KI-Fairness-Lücke bleibt marginal. Im Decision-Experiment (Abb. 2b) differenziert Claude ebenfalls kaum: Im Schiedsgerichts- und Haftszenario bewertet das Modell alle Bedingungen identisch mit 6.0. Treatment-Effekte zeigen sich ausschließlich im Strafmaßszenario, wo Human High und die hybriden Bedingungen Werte von 5.0–6.0 erzielen, während Human Low (M = 4.0) und Robot (M = 4.0) deutlich abfallen. Dieses Muster entspricht den in Chen et al. (2022) präregistrierten Hypothesen, die sich allerdings im damaligen Humanexperiment nicht bestätigt hatten.

DeepSeek-V3 erzeugt das differenzierteste Antwortmuster – und zugleich den auffälligsten Einzelbefund der gesamten Studie. Im Prediction-Experiment liegen die Werte durchweg im mittleren Bereich (ca. 3.5–5.0), wobei das Modell im Strafmaßszenario die niedrigsten Werte prognostiziert (Human High: 3.54, Robot: 3.90). Im Decision-Experiment zeigt sich eine ausgeprägte Mensch-KI-Fairness-Lücke: Im Schiedsgerichtsszenario werden Human High, Human Low und Hybrid High jeweils mit 7.0 bewertet, während Robot auf 5.0 abfällt. Im Haftszenario ergibt sich ein ähnliches Bild mit einer Lücke von etwa einem Skalenpunkt zwischen den menschlichen/hybriden Bedingungen und der Robot-Bedingung.

Besonders markant ist das Strafmaßszenario: Human High erzielt hier 6.64, während Robot bei 3.0 liegt – eine Differenz von über 3.5 Ska-

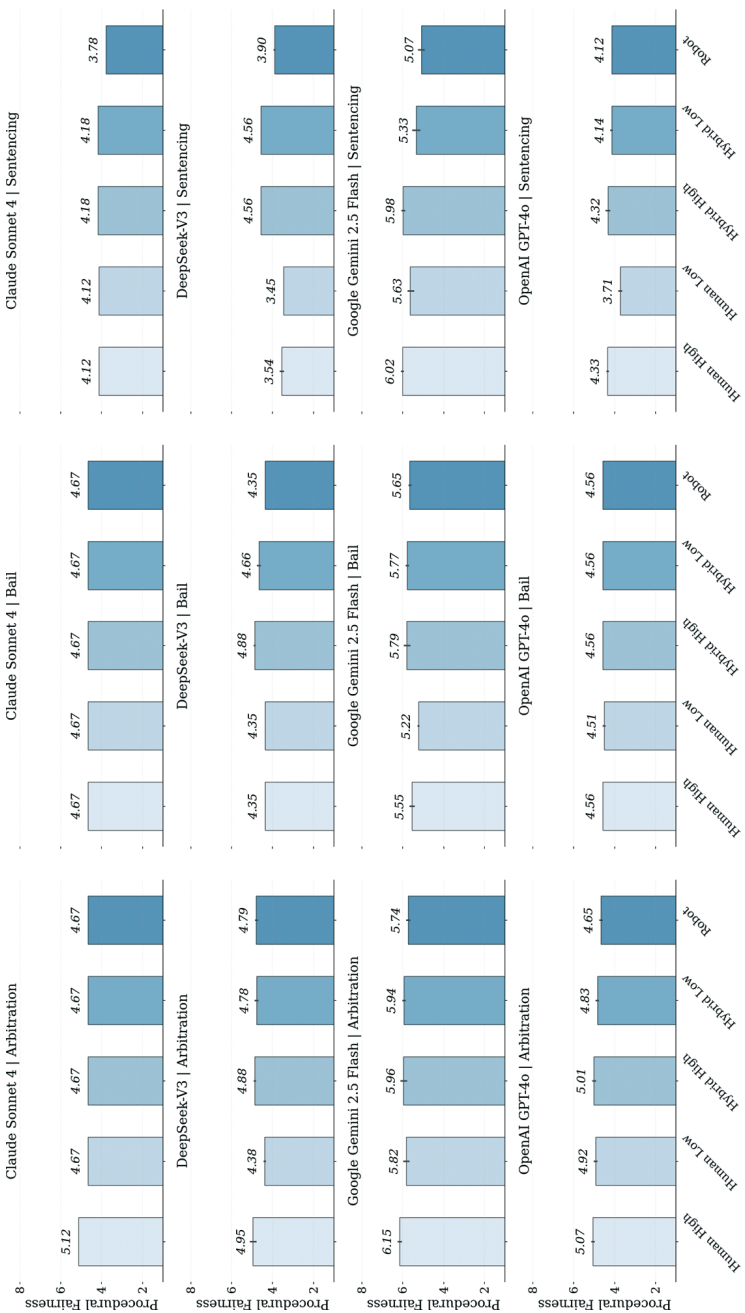

Abb. 2a: LLM-Vorhersagen

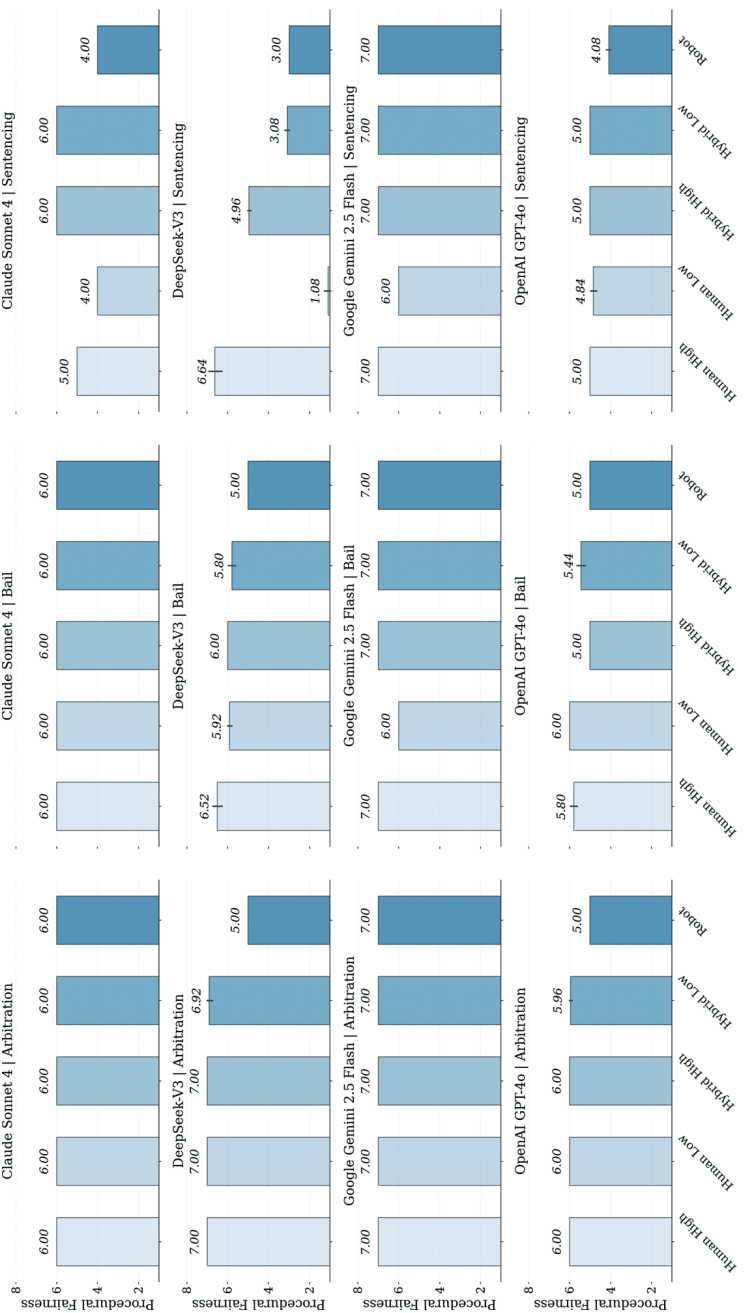

Abb. 2b: LLM-Entscheidungen

lenpunkten. Der extremste Befund betrifft jedoch Human Low: Mit einem Wert von nur 1.08 erreicht diese Bedingung den niedrigsten Wert aller Modell-Szenario-Kombinationen in der gesamten Studie. DeepSeek bewertet demnach eine flüchtige menschliche Prüfung bei schwerwiegenden Entscheidungen als nahezu maximal unfair – unfairer noch als eine rein algorithmische Entscheidung. Dies legt nahe, dass das Modell nicht die bloße Anwesenheit eines Menschen als fairnessrelevant erachtet, sondern die Ernsthaftigkeit seiner Befassung. Eine kursorische menschliche Beteiligung erscheint aus dieser Perspektive als Scheinlegitimation, die das Verfahren stärker delegitimiert als vollständige Automatisierung.

Bemerkenswerterweise spiegelt DeepSeek die zentralen Prinzipien des europäischen Regulierungsrahmens damit qualitativ wider. Die risikobasierte Klassifikation ebenso wie das Human-in-the-Loop-Prinzip fungieren als normative Referenzpunkte, die sich in den Fairnessurteilen des Modells niederschlagen. Dabei wird deutlich, dass nicht jede Form menschlicher Beteiligung genügt, sondern vielmehr eine substanzielle menschliche Aufsicht (meaningful human oversight) erforderlich ist.

Google Gemini 2.5 Flash weicht grundlegend von den übrigen Modellen ab. Im Prediction-Experiment zeigt Gemini moderate Differenzierung mit Werten zwischen etwa 4.4 und 6.0, wobei im Strafmaßszenario die niedrigsten Werte auftreten (Human High: 6.02, Robot: 5.07). Im Decision-Experiment hingegen tritt ein Ceiling-Effekt auf: Das Modell bewertet nahezu alle Bedingungen in allen Szenarien mit dem Maximalwert 7.0 – selbst die Robot-Bedingung im Strafmaßszenario. Die einzige Ausnahme bildet Human Low im Haftszenario (M = 6.0). Die ansonsten robuste Mensch-KI-Fairness-Lücke lässt sich mit diesem Modell nicht reproduzieren. Methodisch wirft dieser Befund die Frage auf, ob Gemini bei normativen Bewertungsaufgaben zu affirmativen Antworten neigt oder ob das Modell genuine Indifferenz gegenüber der Mensch-Maschine-Unterscheidung aufweist.

OpenAI GPT-4o reproduziert die Mensch-KI-Fairness-Lücke in moderater, aber konsistenter Form. Im Prediction-Experiment zeigt das Modell ein relativ flaches Profil mit Werten zwischen 3.71 (Human Low, Strafmaß) und 5.07 (Human High, Schiedsgericht). Die vorhergesagten Treatment-Effekte fallen gering aus, wobei die Robot-Bedingung tendenziell etwas niedriger bewertet wird als die menschlichen Bedingungen. Im Decision-Experiment differenziert GPT-4o stärker: Im Strafmaßszenario zeigt sich eine klare Abstufung mit Human High und den hybriden Bedingungen bei etwa 5.0 und der Robot-Bedingung bei 4.08. Im Schiedsgerichtsszenario

liegen Human High, Human Low und Hybrid High bei 6.0, während Robot auf 5.0 abfällt. Im Haftszenario fallen die Unterschiede am geringsten aus. Insgesamt zeigt GPT-4o ein Muster, das dem des Humanexperiments am nächsten kommt, wenngleich die absoluten Niveaus variieren.

Die vier Modelle lassen sich hinsichtlich ihrer Sensitivität für die Mensch-KI-Unterscheidung auf einem Spektrum anordnen: DeepSeek reagiert am stärksten auf Treatmentvariationen, gefolgt von GPT-4o und Claude, während Gemini kaum differenziert. Bemerkenswert ist zudem die Diskrepanz zwischen Prediction und Decision: Alle Modelle zeigen im Decision-Experiment stärkere Treatment-Effekte als im Prediction-Experiment – sie unterschätzen also systematisch die Bedeutung, die sie selbst der Mensch-KI-Unterscheidung beimessen, wenn sie menschliches Verhalten vorhersagen sollen.

4. Schlussbetrachtung

Für die rechts- und justizpolitische Diskussion ergeben sich mehrere Schlussfolgerungen. Erstens bestätigen unsere Humandaten, dass der Einsatz algorithmischer Entscheidungssysteme ohne menschliche Aufsicht unabhängig vom Entscheidungskontext mit erheblichen Fairnesseinbußen verbunden ist. Die Human-in-the-Loop-Anforderungen der DSGVO und der EU-KI-Verordnung finden damit empirische Unterstützung.

Zweitens ist der Befund, dass bereits minimale menschliche Beteiligung ausreicht, um die Mensch-KI-Fairness-Lücke zu schließen, ambivalent. Einerseits deutet er auf erhebliche Kosteneinsparungen hin, ohne Einbußen in der wahrgenommenen Fairness. Hybride Systeme könnten so Vorteile wie höhere Genauigkeit, größere Konsistenz oder eine Reduktion systematischer Verzerrungen bieten, ohne ihre gesellschaftliche Legitimität zu gefährden. Andererseits legt derselbe Befund das Risiko von Humanwashing nahe: Wenn rein formale, kursorische Prüfungen dieselbe Legitimationswirkung entfalten wie substanzielle menschliche Kontrolle, entsteht ein Anreiz für ressourcenschonende, aber faktisch wirkungslose Aufsichtsmechanismen. Die zentrale regulatorische Herausforderung besteht daher darin, substanzielle menschliche Aufsicht von bloß symbolischer Beteiligung zu unterscheiden – eine Differenzierung, die sich in den Fairnesswahrnehmungen der Rechtsunterworfenen offenbar nicht zuverlässig widerspiegelt.

Das zentrale methodische Ergebnis der Studie ist die ausgeprägte Heterogenität zwischen den untersuchten LLMs. Die Modelle weichen nicht nur deutlich von den Befunden der Humanexperimente ab; sie erzeugen auch untereinander stark divergierende und teils gegensätzliche Resultate, die je nach Modell unterschiedliche Narrative über Fairnesswahrnehmungen stützen. Während DeepSeek substanzielle menschliche Beteiligung schätzt, implizieren die Befunde von Gemini, dass die Unterscheidung zwischen menschlichen und maschinellen Entscheidern weitgehend irrelevant ist; GPT-4o reproduziert das Muster der Humanexperimente lediglich abgeschwächt.

Das zentrale Problem liegt dabei weniger in der Abweichung von den Urteilen menschlicher Probanden, die über gängige Rekrutierungsplattformen wie Prolific gewonnen wurden. Deren Stichproben sind nicht notwendig repräsentativ; entsprechend kann auch das Humanexperiment nicht ohne Weiteres als ground truth gelten.

Das eigentliche methodische Kernproblem beim Einsatz von LLMs

als Ergänzung oder Substitut zu Humanexperimenten besteht vielmehr in der mangelnden Modellstabilität. Wenn identische Stimuli zu qualitativ divergierenden Ergebnissen führen, ist zweifelhaft, inwiefern LLM-Experimente derzeit eine belastbare Grundlage für empirische oder normative Schlussfolgerungen bieten. Die Resultate hängen in hohem Maße vom gewählten Modell und – häufig implizit – von der Aggregationsmethode ab. Da sowohl die Modellauswahl als auch die Aggregationsstrategie erhebliche Freiheitsgrade bieten, lassen sich mit LLMs nahezu beliebige Ergebnisse generieren.

Auch wenn LLMs ein faszinierendes und hilfreiches Instrumentarium bieten, werfen unsere Befunde zumindest Zweifel daran auf, dass LLMs sich bereits heute als kostengünstige Alternative zu klassischen Humanexperimenten nutzen lassen. Human-in-the-Loop bleibt daher in doppelter Hinsicht zentral: zur Sicherung wahrgenommener Verfahrensfairness und zur Gewinnung belastbarer empirischer Erkenntnisse über die Wirkungen des Rechts. Auch wenn eine Konvergenz menschlicher und maschineller Bewertungen denkbar ist, lässt sich mit Blick auf die Generierung von Hypothesen und von Verhaltensbefunden folgern: Human-Experiment und LLM-Experimente sind keine Substitute, sondern Komplemente. Für die experimentelle und empirische Rechtsforschung ist das möglicherweise ein Vorteil. Das Arsenal von Methoden im Antlitz fundamentaler Unsicherheit wird lediglich reichhaltiger.

Literatur

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. PLoS ONE, 11(10), e0164856.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. MIT Press.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. Cognition, 181, 21–34.

Binns, R. (2022). Human judgment in algorithmic loops: Individual justice and automated decision-making. AI & Society, 37(3), 1169–1179.

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making, 33(2), 220–239.

Chen, B., Hermstrüwer, Y., Langenbach, P., Stremitzer, A., & Tobia, K. (2025). Mitigating the Judicial Human-AI Fairness Gap. Center for Law & Economics Working Paper Series, Nr. 7.

Chen, B., Stremitzer, A., & Tobia, K. (2022). Having Your Day in Robot Court. Harvard Journal of Law & Technology, 36(1), 127–169.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2. Aufl.). Lawrence Erlbaum Associates.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (S. 797–806). ACM.

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. Philosophy & Technology, 29(3), 245–268.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580.

Engel, C. (2001). Offene Gemeinwohldefinitionen. Rechtstheorie, 32(1), 23–52.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. Report for the Administrative Conference of the United States.

Henning, M., & Langenbach, P. (2024). Procedural fairness in algorithmic decision-making: The role of explanations and human oversight. Journal of Behavioral Public Administration, 7(1), 1–28.

Hermstrüwer, Y. (2020). Fairnessprinzipien der algorithmischen Verwaltung: Diskriminierungsprävention beim staatlichen Einsatz von Machine Learning. Archiv des öffentlichen Rechts, 145(4), 479–521.

Hermstrüwer, Y., & Langenbach, P. (2023). Human-AI cooperation in legal decision-making: Evidence from algorithmic assistance in the courtroom. Law and Human Behavior, 47(3), 234–267.

Imai, K., & Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. Political Analysis, 21(2), 141–171.

Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2020). Simple rules to guide expert classifications. Journal of the Royal Statistical Society Series A: Statistics in Society, 183(3), 771–800.

Katz, D. M., Bommarito, M. J., & Blackman, J. (2017). A general approach for predicting the behavior of the Supreme Court of the United States. PLoS ONE, 12(4), e0174698.

Kehl, D., Guo, P., & Kessler, S. (2017). Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. The Quarterly Journal of Economics, 133(1), 237–293.

Lehr, D., & Ohm, P. (2017). Playing with the data: What legal scholars should learn about machine learning. University of California Davis Law Review, 51(2), 653–717.

Liu, Y., & Li, Z. (2025). AI-assisted judicial reasoning: Opportunities and challenges for legal decision-making. Journal of Legal Technology, 8(1), 45–62.

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. Telematics and Informatics, 67, 101751.

Mayson, S. G. (2018). Bias in, bias out. Yale Law Journal, 128(8), 2218–2300.

Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

Pearl, J. (2001). Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (S. 411–420). Morgan Kaufmann.

Political Asylum/Immigration Representation Project (2016). Pro Bono Asylum Manual. Boston College Law School.

Re, R. M., & Solow-Niederman, A. (2019). Developing artificially intelligent justice. Stanford Technology Law Review, 22(2), 242–289.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics, 15(6), 657–680.

Selbst, A. D. (2021). An institutional view of algorithmic impact assessments. Harvard Journal of Law & Technology, 35(1), 117–186.

Stevenson, M. T. (2018). Distortion of justice: How the inability to pay bail affects case outcomes. The Journal of Law, Economics, and Organization, 34(4), 511–542.

Stevenson, M. T., & Doleac, J. L. (2019). Algorithmic Risk Assessment in the Hands of Humans. NBER Working Paper Nr. 26448.

Tyler, T. R. (1997). The psychology of legitimacy: A relational perspective on voluntary deference to authorities. Personality and Social Psychology Review, 1(4), 323–344.

Tyler, T. R., & Huo, Y. J. (2002). Trust in the Law: Encouraging Public Cooperation with the Police and Courts. Russell Sage Foundation.

Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International, 22(4), 97–112.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law, 7(2), 76–99.

Wischmeyer, T. (2020). Artificial intelligence and transparency: Opening the black box. In T. Wischmeyer & T. Rademacher (Hrsg.), Regulating Artificial Intelligence (S. 75–101). Springer.

Završnik, A. (2020). Criminal justice, artificial intelligence systems, and human rights. ERA Forum, 20(4), 567–583.