LLM-gestützte Simulationen in der rechtsempirischen Forschung Ein Werkstattbericht

Johannes Kruse & Pascal Langenbach

A. Einleitung

Der computertechnische Fortschritt und Innovationen im Bereich der künstlichen Intelligenz (KI) beschäftigen zunehmend auch die Rechtswissenschaft. Dies ist nur folgerichtig, schließlich wird KI-Anwendungen ein disruptives Potential zugeschrieben, das neben anderen gesellschaftlichen Bereichen auch das Recht bzw. das Rechtssystem nachhaltig beeinflussen kann. Dies betrifft keineswegs allein die Rechtspraxis, also die Rechtsanwendung in Behörden, Gerichten oder der Anwaltschaft. Gerade die jüngste Entwicklungsstufe, die generative KI in Gestalt sogenannter großer Sprachmodelle (engl. Large Language Models oder LLMs), könnte ebenso große Veränderungen für das rechtswissenschaftliche Arbeiten und Forschen mit sich bringen. Gleichwohl finden diese potentiellen Auswirkungen für die Rechtswissenschaft – gerade im Vergleich zur Rechtspraxis (Begriffe wie „legal tech" oder der „robojudge" sind schon seit Jahren debattenprägend) – jedenfalls im deutschen Sprachraum bislang (zu) wenig Beachtung. Zentrale Fragen (etwa: wie kann/könnte künstliche Intelligenz die rechtswissenschaftliche Methodik und allgemein das juristische Denken verändern) werden bislang allenfalls vereinzelt aufgegriffen. Dabei sind die denkbaren Anwendungsszenarien zahlreich: Sie umfassen die Auswertung und Systematisierung von Rechtsprechung und Literatur, die Entwicklung dogmatischer Figuren und sogar das vollständige Verfassen rechtswissenschaftlicher Texte.

Christoph Engel ist eine Ausnahme. Neben den Anwendungsmöglichkeiten in der Rechtspraxis erstreckte sich seine Neugier von Anfang an auch auf die Rechtswissenschaft. Mit einem der beiden Autoren ist er etwa der Frage nachgegangen, inwieweit sich das Verfassen juristischer Kommentare, eine Kerntätigkeit der dogmatischen Rechtswissenschaft,

an LLMs delegieren lässt.[1] Eingedenk der Tatsache, dass LLMs aus der Forschung zum Natural Language Processing hervorgegangen sind, liegen solche Anwendungen „eigentlich" auf der Hand. Weniger offensichtlich sind dagegen die Anwendungsmöglichkeiten von LLMs in einer rechtswissenschaftlichen Subdisziplin, für die Christoph Engel seit mehr als 20 Jahren steht wie wohl kein anderer: der experimentellen Rechtsforschung.

In verschiedenen sozialwissenschaftlichen Disziplinen werden LLMs bereits heute vermehrt eingesetzt, um zuvor mit menschlichen Versuchspersonen durchgeführte Experimente zu replizieren.[2] Die Ergebnisse dieser Versuche fallen im Einzelnen unterschiedlich aus, sind in der Tendenz aber durchaus vielversprechend. Im Rahmen einer großangelegten Replikationsstudie mit 276 ökonomischen Experimenten konnte etwa eine Vorhersagegenauigkeit von 78 % erreicht werden.[3] Im Bereich der experimentellen Rechtsforschung gibt es (noch) keinen entsprechenden Replikationstrend. Groß angelegte Untersuchungen, die eine Vielzahl von Studien in den Blick nehmen, stehen aus. Die Replikation einzelner Studien wurde aber bereits in Angriff genommen, wobei sich die entsprechenden Arbeiten an einer Hand abzählen lassen.[4] Einer der Vorreiter ist (einmal mehr) *Christoph Engel*: Gemeinsam mit *Richard McAdams* hat er eine LLM-gestützte Pipeline genutzt, um das Konzept des ordinary meaning (dem im US-amerikanischen Recht zentrale Bedeutung zukommt) zu erhellen.[5] Eine zu diesem Thema von *Kevin Tobia* durchgeführte Umfrage[6]

1 *Engel/Kruse*, Kommentar ohne Autor: Können Sprachmodelle das Kommentieren übernehmen?, JZ 79 (2024), 997; *dies.*, LLM as a law professor: Having a large language model write a commentary on freedom of assembly, Artificial Intelligence and Law 2026 (im Erscheinen).

2 *Anthis et al.*, LLM Social Simulations Are a Promising Research Method (2023), Preprint arXiv:2504.02234, 3.

3 *Hewitt et al.*, Predicting Results of Social Science Experiments Using Large Language Models (2024), 11; siehe weiterhin *Anthis et al.* (Fn. 2), 3; *Chen/Hu/Lu*, Predicting Field Experiments with Large Language Models (2025), Preprint arXiv:2504.01167, 1 f.

4 *Arbel/Hoffman*, Generative interpretation, NYUL Rev. 99 (2024), 451; *Arbel*, The Silicon Reasonable Person,University of Alabama School of Law Legal Studies Research Paper Series 2025; *Kruse*, The Ordinary Meaning Bot: Simulating Human Surveys with LLMs, MPI Collective Goods Discussion Paper, No. 2025/12, https://ssrn.com/abstract=5378203.

5 *Engel/McAdams*, Asking GPT for the Ordinary Meaning of Statutory Terms, University of Illinois Journal of Law, Technology and Policy 235 (2024).

6 *Tobia*, Testing Ordinary Meaning, Harv Law Rev 134 (2020), 726.

konnte so teilweise repliziert werden.[7] Weitere Studien stammen etwa von *Yonathan Arbel* oder einem der Autoren.

Der vorliegende Beitrag folgt dem von *Christoph Engel* in den letzten Jahren insoweit eingeschlagenen Pfad und unternimmt mit einer LLM-gestützten Simulationspipeline den Versuch, ein rechtsempirisches Experiment zu replizieren. Um der (deutschen) Rechtswissenschaft entsprechende Möglichkeiten aufzuzeigen, ist der folgende Beitrag als Werkstattbericht konzipiert: Eine mögliche Vorgehensweise für rechtsempirische Replikationsstudien wird vorgestellt und bestehende Hürden sowie Fallstricke aufgezeigt. Exemplifiziert wird das Vorgehen „natürlich" mit der Replikation einer Studie von *Christoph Engel* – und *Nina Grgić-Hlača* als Co-Autorin. Diese Studie untersucht die Auswirkungen der durch den Supreme Court von Wisconsin[8] geforderten Warnungen bei der Verwendung des COMPAS-Tools in Strafverfahren und ist 2021 im *Journal of Legal Analysis* erschienen.[9]

Die vorliegende Untersuchung kann als Replikationsstudie im vorstehenden Sinne verstanden werden. Der klassische Replikationsbegriff passt indes nur teilweise. Klassische Replikationsstudien wollen die Robustheit eines berichteten (experimentellen) Effekts untersuchen. *Darum geht es auch vorliegend*. Zugleich wollen sie überprüfen, ob es sich bei den in der zu replizierenden Studie gefundenen Effekten um Zufallsbefunde handeln könnte (und mitunter auch, ob die Erstellung der Originalstudie angemessen und transparent erfolgte). *Darum geht es vorliegend eher nicht*. Die Befunde eines de lege artis durchgeführten sozialwissenschaftlichen Experiments können bzw. sollten aktuell wohl noch nicht allein auf Grundlage einer KI-basierten Replikation in Zweifel gezogen werden. Im Fall divergierender Ergebnisse zwischen Originalstudie und KI-Simulation liegt die Rechtfertigungslast dem Grunde nach bei der KI-Simulation. Dementsprechend kann man eine gelungene Replikation zwar als Beleg für die Robustheit der mit Menschen gewonnenen Befunde deuten. Gelingt die Replikation hingegen nicht, dürfte zunächst die Leistungsfä-

7 Was sich in einem hohen alignment zwischen den LLM-generierten und den echten Antworten niederschlägt, vgl. *Engel/McAdams* (Fn. 5); mit einer verbesserten Vorhersagepipeline ließ sich die Übereinstimmung sogar noch deutlich steigern, vgl. *Kruse* (Fn. 4), 3 f.

8 Loomis v. Wisconsin, 881 N.W.2d 749 (Wis. 2016).

9 *Engel/Grgić-Hlača*, Machine Advice with a Warning about Machine Limitations. Experimentally Testing the Solution Mandated by the Wisconsin Supreme Court, Journal of Legal Analysis 13 (2021), 284.

higkeit KI-basierter Simulationen zu hinterfragen sein. Schließlich kann man experimentelle Studien mit menschlichen Versuchspersonen selbstverständlich niemals KI-basiert vollständig nachbilden. Replikation im vorliegenden Sinn meint dementsprechend keine Wiederholung des Originalexperiments, sondern stets eine *konzeptionelle* Replikation der gefundenen Ergebnisse.

Konzeptionelle Replikationsstudien mit KI-Agenten können unterschiedliche Zielsetzungen verfolgen. Einerseits kann es darum gehen, LLM-gestützte Simulationen mit KI-Agenten als neue sozialwissenschaftliche Methode zu etablieren. Die Durchführung KI-basierter sozialwissenschaftlicher Experimente ist ein relativ neues Phänomen. Vor diesem Hintergrund können gelingende Replikationen (insbesondere wenn diese sich über eine Vielzahl von Studien aus unterschiedlichen Disziplinen erstrecken) das generelle Vertrauen in die Validität der so gewonnenen Befunde stärken. Ist ein hinreichendes Vertrauensniveau erreicht, könnte man zukünftig womöglich teilweise auf menschliche Versuchspersonen verzichten und Experimente rein KI-basiert durchführen. Praktisch können KI-basierte Experimente auch als Vorstufe für Experimente mit Menschen dienen; um das Forschungsdesign zu testen oder Hypothesen zu entwickeln.[10] Insoweit bedarf es vor allem großangelegter Replikationsstudien, die für einen bestimmten Forschungsbereich oder eine bestimmte Methode den Nachweis erbringen, dass eine Replikation mit KI-Agenten grundsätzlich möglich ist.

Andererseits kann eine Replikationsstudie bei einer konkreten Forschungsfrage ansetzen und versuchen, die bereits vorliegenden Erkenntnisse zu erweitern. Auch hier geht es zunächst um die Frage, ob bzw. inwieweit sich ein mit Menschen durchgeführtes Experiment KI-basiert replizieren lässt. Ist dieser Nachweis erbracht, kann man den Ansatz ausbauen, um weitergehende Erkenntnisse zu gewinnen. Einen Erkenntnisgewinn verspricht insbesondere die Überwindung forschungspraktischer Hürden, denen Studien mit Menschen naturgemäß ausgesetzt sind (zum Beispiel auf Grund ethischer Grenzen oder begrenzter Forschungsmittel). Hierzu bedarf es weniger einer Untersuchung in der Breite als in der Tiefe. Man konzentriert sich auf eine einzelne Studie/Forschungsfrage, die sodann en détail beleuchtet wird. Freilich gewinnt man auch auf diese Weise

10 *Dillion et al.*, Can AI language models replace human participants?, Trends in Cognitive Sciences 27 (2023), 597, 598; *Hewitt et al.* (Fn. 3).

Einsichten über die generelle Tauglichkeit LLM-gestützter Simulationen, die das Vertrauen in die neuartige Methodik stärken oder schwächen können. Insoweit überschneiden sich die beiden Zielsetzungen.

Wie erwähnt gibt es bereits großangelegte Replikationsstudien, die in unterschiedlichen Disziplinen versuchen, den Nachweis zu erbringen, dass sich sozialwissenschaftliche Untersuchungen mit KI-Agenten replizieren lassen. Gerade vor diesem Hintergrund verfolgt der Beitrag primär die zweite Zielsetzung. Im Rahmen der LLM-gestützten Simulation erfolgt eine eingehende Auseinandersetzung mit einer einzelnen Studie. Auf diese Weise wird untersucht, inwieweit sich die praktischen Begrenzungen der ursprünglichen Studie überwinden ließen. Daneben verfolgt der Beitrag ein generelles Erkenntnisinteresse, indem er nach Gründen für das Gelingen oder Nichtgelingen KI-basierter Replikationen im Bereich experimenteller Rechtsforschung fragt.

Der vorliegende Beitrag versteht sich als Werkstattbericht. Er berichtet kein abgeschlossenes Unterfangen, sondern zeichnet „work in progress" nach. Es kommt ihm daher weniger darauf an, inwieweit die konkret unternommene Replikation gelingt. Vielmehr will er aufzeigen, wie man eine LLM-gestützte Replikation im Bereich der experimentellen Rechtsforschung durchführen könnte und welche Herausforderungen dabei zu meistern sind. Der Beitrag gliedert sich wie folgt: Zunächst erfolgt eine kurze Darstellung der Originalstudie (unter B.I), um das Fundament für die Replikationsstudie zu legen. Danach wird die Replikationsstudie dargestellt (unter B.II), wobei deren vier Module jeweils gesondert betrachtet werden. Anschließend werden die Simulationsergebnisse beschrieben sowie (kurz) analysiert (unter B.III). Der Beitrag schließt mit einem Ausblick auf die Einsatzmöglichkeiten, Chancen und Risiken LLM-gestützter Simulationen in der empirischen Rechtsforschung (unter C.).

B. Ein Replikationsversuch

I. Die Originalstudie

1. Kontext

In den Vereinigten Staaten können Richterinnen und Richter in vielen Bundesstaaten zur Unterstützung (einiger) ihrer Entscheidungen im Rahmen von Strafverfahren, etwa hinsichtlich der vorläufigen Freilassung eines Angeklagten auf Kaution (bail) oder der Strafzumessung im Fall

einer Verurteilung (sentencing) auf algorithmische Entscheidungshilfen zurückgreifen.[11] Diese Entscheidungsunterstützungssysteme stellen Risikoklassifizierungen für die jeweiligen Angeklagten zur Verfügung, zum Beispiel hinsichtlich der Wahrscheinlichkeit einer erneuten Verhaftung in einem bestimmten Zeitraum. Eine solche Anwendung ist COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). Da es sich um eine kommerzielle Anwendung handelt, sind die der Risikoklassifizierung zugrundeliegenden Algorithmen nicht öffentlich zugänglich. Die Verwendung von COMPAS zur Strafzumessung im US-Bundesstaat Wisconsin wurde vom Supreme Court von Wisconsin überprüft.[12] Im zugrundeliegenden Fall war COMPAS in einem Verfahren gegen Eric Loomis eingesetzt worden, das mit einer Freiheitsstrafe endete. Die Strafzumessung stützte sich in Teilen auf die von COMPAS ausgegebene Risikoklassifizierung. Der Supreme Court von Wisconsin sah in der Verwendung von COMPAS keine Verletzung der „Due-process"-Rechte von Eric Loomis. Allerdings stellte er die Verwendung von COMPAS im Rahmen der Strafzumessung unter die Bedingung, dass die COMPAS-Risikoeinschätzungen mit Warnungen über die Beschränktheit der Klassifizierungen verbunden werden.[13]

2. Forschungsfrage

Die Forderung des Supreme Courts, wonach die Risikoeinschätzungen durch COMPAS den Richtern nur gemeinsam mit einer entsprechenden Warnung zur Verfügung gestellt werden dürfen, impliziert, dass die entsprechenden Warnungen Einfluss auf das Verhalten der Richter haben. Das Gericht scheint zu erwarten, dass die Richterinnen und Richter ein anderes Strafmaß verhängen, wenn sie zusätzlich zur maschinellen Risikoeinschätzung eine Warnung sehen. Andernfalls wäre die Warnung überflüssig.[14] Diesen unterstellten Verhaltenseffekt der Warnung auf das

11 Die Beschreibungen in diesem Absatz entstammen der Originalstudie, *Engel*/Grgić-Hlača (Fn. 9), 288 f., sowie Recent Case, Criminal Law – State v. Loomis, 881 N.W.2d 749 (Wis. 2016), Harv Law Rev 130 (2017), 1530, siehe dort jeweils auch für weitere Verweise auf die Entscheidung und die Prozessmaterialien.

12 State v. Loomis, 881 N.W.2d 749 (Wis. 2016).

13 Siehe für den Inhalt der Warnung *Engel*/Grgić-Hlača (Fn. 9), 288 f.

14 *Engel*/Grgić-Hlača (Fn. 9), 291: „For the solution mandated by the Wisconsin Supreme Court to be effective human decision-makers must decide differently when

Verhalten von Richterinnen und Richtern untersuchen *Christoph Engel* und *Nina Grgić-Hlača* in ihrer Studie.

Um die Forschungsfrage experimentell operationalisierbar zu machen, wandeln sie den ursprünglichen Entscheidungskontext in verschiedener Hinsicht ab. Das Urteil des Supreme Courts von Wisconsin betrifft die Verwendung von COMPAS bei der Strafzumessung. *Engel* und *Grgić-Hlača* untersuchen in ihrer Studie jedoch Entscheidungen über die Freilassung auf Kaution im Vorfeld der Anklage („bail or jail"). Grund hierfür ist die Verfügbarkeit eines Datensatzes, der für diese Entscheidungskategorie sowohl die COMPAS-Bewertung als auch die tatsächliche Rückfälligkeit innerhalb von zwei Jahren enthält.[15] Zudem beschränken sie sich auf die Bewertungen zur Rückfälligkeit (General Recidivism Score), obwohl COMPAS auf unterschiedlichen Dimensionen Risikobewertungen errechnet (zusätzlich einen „Violent Recidivsm Score" und das „Pretrial Release Risk").[16]

3. Studiendesign

Wie die Autor:innen betonen, ist es ausgeschlossen, in echten Strafverfahren einigen Richterinnen oder Richtern die geforderten Warnungen zu zeigen, anderen hingegen nicht.[17] Deshalb weichen sie methodisch auf eine Vignette-Studie aus, in der sie die Kautions-Entscheidung experimentell nachbilden. Dazu erstellen sie aus dem genannten Datensatz 50 Vignetten, die Informationen über die begangene Tat, das Alter, das Geschlecht sowie über die kriminelle Vorgeschichte des/der Angeklagten enthalten. Auf Grundlage dieser Vignetten und der von COMPAS errechneten Risikobewertung treffen die Versuchspersonen ihre Entscheidungen. In ihrer Veröffentlichung bündeln *Engel* und *Grgić-Hlača* die Ergebnisse fünf unterschiedlicher (Teil-)Studien, die im experimentellen Aufbau variieren. Die Variationen haben im Großen und Ganzen keinen wesentlichen Einfluss auf die Ergebnisse.[18] In der Hauptstudie (Teilstudie 1) lesen die Versuchspersonen zunächst die jeweilige Vignette. Dann

just receiving machine advice compared with receiving machine advice that comes with the prescribed warnings."

15 *Engel*/Grgić-Hlača (Fn. 9), 286.

16 *Engel*/Grgić-Hlača (Fn. 9), 299.

17 *Engel*/Grgić-Hlača (Fn. 9), 294.

18 *Engel*/Grgić-Hlača (Fn. 9), 286f.

treffen sie Entscheidungen zu dieser Vignette. Sie geben für jede Vignette an, ob sie sich für „jail" oder „bail" entscheiden, wie hoch sie die Rückfallwahrscheinlichkeit einschätzen und wie sicher sie sich ihrer Entscheidung sind. Im Anschluss sehen sie den COMPAS-Risiko-Score und treffen die Entscheidungen erneut. Der Unterschied zwischen der Kontrollgruppe und der Treatmentgruppe besteht darin, dass die COMPAS-Risikoklassifizierung nur für letztere Gruppe mit einer Warnung versehen ist. Bei der im Experiment verwendeten Warnung handelt es sich um die gekürzte Originalversion, die in Wisconsin nach dem Urteil des dortigen Supreme Courts tatsächlich verwendet wurde.[19] Die Hauptstudie untersucht, wie sich die Warnung auf Veränderungen zwischen der ersten und der zweiten Einschätzung auswirkt.[20] In einer zweiten ergänzenden Teilstudie wird dieser sequentielle Aufbau aufgegeben, nun erhalten die Versuchspersonen den COMPAS-Score bereits unmittelbar mit der jeweiligen Vignette.[21] Die hier beschriebene Replikation versteht sich wie dargestellt als ein erster Schritt und beschränkt sich – aus pragmatischen Gründen – auf den Versuchsaufbau der zweiten Teilstudie.

4. Versuchspersonen

Die Versuchspersonen der Originalstudie rekrutieren sich aus dem Pool des Panel-Providers *Prolific*. An der zweiten Teilstudie nahmen letztlich 134 Personen teil. Jeweils 67 Personen in der Kontrollgruppe (keine Warnung) und 67 in der Treatment-Gruppe (Warnung). Alle 134 Personen hatten zuvor zwei Fragen zur Aufmerksamkeitskontrolle richtig beantwortet.[22]

Die Entscheidung über die Strafzumessung und die Entscheidung über die vorläufige Freilassung auf Kaution werden in Wisconsin von Richterinnen und Richtern getroffen. Nur das Verhalten dieser sehr kleinen demografischen Gruppe soll durch die Warnung beeinflusst werden. Hier müsste man also ansetzen, um die Verhaltenswirksamkeit der Warnungen „optimal" zu untersuchen. Dem stehen allerdings forschungspraktische Hürden entgegen. Die meisten Richter:innen (aus Wisconsin)

19 *Engel*/Grgić-Hlača (Fn. 9), 297 f.
20 *Engel*/Grgić-Hlača (Fn. 9), 296.
21 *Engel*/Grgić-Hlača (Fn. 9), 306 ff.
22 *Engel*/Grgić-Hlača (Fn. 9), 326, von den ursprünglich 138 Teilnehmer:innen gelang dies also nur vier Personen nicht.

wird man kaum zur Teilnahme an einer Online-Vignette-Studie bewegen können. Hinzu kommt, dass diese in ihrem beruflichen Alltag regelmäßig mit den rechtlich verbindlichen Warnungen konfrontiert sind.[23] Das bedeutet, dass eine Versuchsanordnung, die einen Entscheidungskontext ohne den Effekt der Warnung darstellt, schwer zu implementieren sein könnte, weil die Richter:innen sich die entsprechende Warnung stets hinzudenken (oder ihre erlernte Reaktion auf den COMPAS-Score von der real vorhandenen Warnung geprägt ist). Vor diesem Hintergrund greifen *Engel* und *Grgić-Hlača* auf den Probandenpool von *Prolific* zurück, der eher nicht aus Richterinnen und Richtern aus Wisconsin besteht. Um sich den Charakteristika von Richter:innen oder wenigstens ihrem Erfahrungshintergrund anzunähern, versuchten *Engel* und *Grgić-Hlača* Versuchspersonen mit Jury-Erfahrung zu rekrutieren. In der zweiten Teilstudie verfügten 67 Prozent der Versuchspersonen über entsprechende Erfahrungen.[24] Für die einzelnen Studien berichten *Engel* und *Grgić-Hlača* verschiedene demografische Eigenschaften ihrer Versuchspersonen, die sich von der Verteilung in der Gesamtbevölkerung (US-Census) unterscheiden.

5. Ergebnisse

In der Studie von *Engel* und *Grgić-Hlača* hatten die Warnungen insgesamt keinen übermäßigen Einfluss auf das Entscheidungsverhalten der Versuchspersonen. Insbesondere für die normativ entscheidende Frage nach der Haftentlassung auf Kaution blieben sie ohne erkennbaren Einfluss.[25] Für die zweite Teilstudie werden folgende Befunde berichtet: Die Warnungen hatten keinen Effekt auf die Entscheidung „bail" vs. „jail", führten aber zu einer signifikanten Erhöhung der geschätzten Rückfallwahrscheinlichkeit; ferner waren sich die Versuchspersonen ihrer Einschätzung sicherer, wenn dem Score eine Warnung beigefügt war.[26]

23 Für diese Einschränkungen *Engel*/Grgić-Hlača (Fn. 9), 298 f.
24 *Engel*/Grgić-Hlača (Fn. 9), 285, 326.
25 Zusammenfassend *Engel*/Grgić-Hlača (Fn. 9), 286.
26 *Engel*/Grgić-Hlača (Fn. 9), 308 f.

II. Die Replikationsstudie

1. Überblick

Die vorliegende Replikationsstudie nutzt Large Language Models, um das Verhalten menschlicher Versuchspersonen in einem experimentellen Setting zu simulieren. Dem liegt die Erwartung zugrunde, dass LLMs durch ihre Trainingsdaten sowie Designziele ein relativ gutes „Verständnis" davon haben, wie sich Menschen in unterschiedlichen Situationen verhalten, und deshalb als eine Art „homo silicus"[27] dienen können. Gerade durch das Fine-Tuning mittels Reinforcement Learning wird diesen Modellen vermittelt, wie Menschen auf bestimmte, sprachlich vermittelte Stimuli reagieren. LLM-gestützte Simulationen können eine beachtliche Vorhersagegenauigkeit erreichen. In einigen Studien zu sozialwissenschaftlichen Experimenten korrelieren GPT-4-Vorhersagen in hohem Maße (r = 0,85 bis 0,91) mit den tatsächlichen Treatment-Effekten.[28] Aus dem Vorgesagten folgt zugleich, dass die Vorhersagegenauigkeit maßgeblich davon abhängt, ob bzw. inwieweit das in Rede stehende Verhalten in den Trainingsdaten repräsentiert ist.[29] Dementsprechend dürften LLMs etwa das Verhalten von Verbrauchern besser abbilden können als von professionellen Richtern. Um das Verhalten der menschlichen Versuchspersonen simulieren zu können, kommen KI-Agenten mit einem spezifischen demografischen Profil zum Einsatz, das auf den (wenigen) bekannten demografischen Eigenschaften der echten Versuchspersonen beruht. Die der Replikation zugrunde liegende Pipeline setzt sich aus vier Modulen zusammen, die jeweils gesondert betrachtet werden.

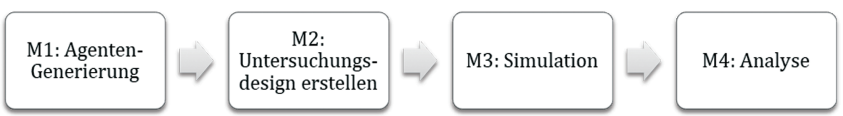

Abbildung 1: Die vier Module der Studie

27 *Horton,* Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, Preprint arXiv:2301.07543 (2023), 1

28 *Hewitt et al.* (Fn. 3), 1; *Anthis et al.* (Fn. 2), 3.

29 Vgl. etwa *Kruse* (Fn. 4), 4 f.

Die vorliegende Studie ist als Prototyp konzipiert und soll in erster Linie als Anschauungsobjekt dienen. Dementsprechend wurden beim Untersuchungsdesign pragmatisch motivierte Entscheidungen getroffen. Um Dauer und Kosten der Durchführung im Rahmen zu halten, wurden insbesondere sowohl bei der Anzahl der KI-Agenten als auch der genutzten Modelle Abstriche gemacht.

2. Modul 1: Agentengenerierung

Modul 1 erzeugt synthetische Versuchspersonen (KI-Agenten) mit realitätsnahen demografischen Profilen als Ersatz für menschliche Versuchspersonen. Zunächst werden N = 268 Basis-Profile generiert, die die bekannten demografischen Charakteristika der echten Untersuchungsteilnehmer:innen abbilden. Deshalb konnten nur solche demografischen Merkmale verwendet werden, über die die Originalstudie Auskunft gibt. Das sind etwa das Geschlecht (63 % männlich, 37 % weiblich), die ethnische Zugehörigkeit (70 % weiß, 5 % afroamerikanisch, 7 % hispanisch, 17 % asiatisch, 1 % sonstige) oder die Jury-Erfahrung (67 % ja, 33 % nein). Weiterhin die politische Orientierung (50 % liberal, 18 % konservativ und 32 % moderat) sowie der Bildungsgrad (94 % Bachelorabschluss oder höher, 6 % sonstige).

Für jedes demografische Profil wird mittels GPT-4o (Temperature = 0,9) ein kurzer, realistischer Biografietext erzeugt (Name; Alter 25–75; US-Wohnort; Beruf; Familiensituation; Hobbys/Interessen; beruflicher Werdegang; gesellschaftliches Engagement; persönliche Wertvorstellungen; individuelles Merkmal). Die Details stammen aus dem Modell (ohne externe Statistikquellen). Ein Beispielprofil lautet:

> *Grant Mitchell, 38, resides in Portland, Oregon, where he works as a social media strategist for a nonprofit focused on environmental justice. Raised in a small town in Nebraska, Grant was the first in his family to earn a college degree, graduating with a Bachelor's in Environmental Science from the University of Oregon. He is married to Lena, a fellow activist he met at a climate change rally, and together they have a six-year-old daughter, Ruby. Grant is known for his passionate, progressive views, which he often expresses through his writing and activism, but his intensity can sometimes alienate more moderate peers. He has never served on a jury, attributing it to a combination of sheer luck and frequent travel. A fervent cyclist, he enjoys participating in long-distance races and is an advocate for bike-friendly city planning. His career has been diverse: from working as a campaign manager for local green politicians to teaching workshops on sustainable living. Grant's activism extends into his com-*

munity involvement, where he organizes monthly clean-up drives and educational seminars on environmental policies. He is deeply committed to his values of equality and sustainability, though some criticize him for being too unyielding in debates. One unique, yet painful experience in his life was surviving a car accident in his early twenties, which left him with a permanent limp and intensified his commitment to public transportation advocacy. Despite his admirable dedication, Grant occasionally struggles with balancing his activism and family commitments, causing tension at home.

Die N = 268 Basis-Profile wurden randomisiert auf die Bedingungen aufgeteilt: 134 Agenten enthalten den reinen COMPAS-Score (Baseline), 134 einen COMPAS-Score inklusive Warnung (Treatment). Die 268 Basis-Profile werden 1:1 über die Modelle gespiegelt.

Wie bereits erwähnt, dürfte der Rückgriff auf den Teilnehmer-Pool des Panel-Providers *Prolific* und damit die Auswahl der Versuchspersonen in der Originalstudie in erster Linie pragmatisch motiviert gewesen sein. Obgleich man primär an dem Verhalten echter Richter:innen (als den tatsächlichen Adressat:innen der Warnungen) interessiert ist, greift man auf Laien zurück. Unter Kostengesichtspunkten war man sicherlich auch mit Blick auf die Zahl der Versuchspersonen limitiert. Mit KI-Agenten lassen sich diese Beschränkungen überwinden. Zwar sind auch die für die synthetischen Agenten notwendigen API-Abfragen nicht kostenlos (und auch die Antwortzeiten der Modelle sind zu berücksichtigen), im Grundsatz lässt sich die Beobachtungszahl aber skalieren. Aus Kostengründen wurde für die Simulation darauf verzichtet eine größere Anzahl an KI-Agenten zu erstellen. Allerdings übersteigt die Anzahl die der Originalstudie bereits deutlich. Überdies wurden als zweite Agentenpopulation „echte" Richterinnen und Richter erstellt. Das zugrundeliegende demografische Profil beruht auf Daten zu amerikanischen (Bundes-)Richter:innen.[30]

3. Modul 2: Erstellen des Untersuchungsdesigns

Modul 2 bereitet die experimentelle Umgebung vor. Jeder Agent sieht entweder (i) die Vignette mit COMPAS-Risikoinformation oder (ii) dieselbe

30 Die Zahlen stammen aus den Datensätzen des Federal Judicial Center (FJC) zur Demografie der Article-III-Richter; die ABA „Profile of the Legal Profession 2024 – Judges" fasst diese datenbasiert und mit Stichtag 01.08.2024 zusammen, www.americanbar.org/news/profile-legal-profession/judges/.

Vignette mit COMPAS-Risikoinformation plus Warnhinweis. Die Zuweisung der Agenten zur Kontroll- oder Treatment-Gruppe nimmt Modul 1 vor. Pro Agent wird eine Serie von Vignetten zugewiesen (Standard: 10, konfigurierbar); in der Originalstudie erhielten die Versuchspersonen 50 Vignetten. Zur Sicherung der Inhaltsäquivalenz werden die Vignetten zwischen den Bedingungen mittels stabiler Signaturen (CASE_ID, Text-Hash) spiegelbildlich gematcht und verifiziert. Beide Bedingungen enthalten damit identische Fälle in identischer Reihenfolge.

Die Simulation verwendet die ersten 10 Vignetten der Originalstudie, die ein niedriges bis mittleres Risikoniveau (COMPAS Decile 1–6) aufweisen. Für deren Verwendung ließe sich anführen, die entsprechenden Fälle seien rechtspolitisch besonders relevant. Richterinnen und Richtern kommt insoweit tendenziell ein größerer Entscheidungsspielraum zu, sodass hier auch ein größeres Potential für algorithmische Entscheidungshilfen bestehen dürfte. Die Simulation stellt (in Abweichung von der Originalstudie) vorrangig auf die kategoriale Risikoeinstufung (low/medium/high) ab, mit dem Decile-Score als ergänzender Information in Klammern (z. B. „General Recidivism Risk: high (Decile: 8)"). Hierdurch soll einer dokumentierten Schwäche von LLMs Rechnung getragen werden, nämlich den eingeschränkten numerischen Reasoning-Fähigkeiten sowie der Tendenz einer inkonsistenten Verarbeitung granularer Skalen.[31] Während Menschen Decile-Scores intuitiv ordinieren und in Risikoeinschätzungen übersetzen können, zeigen LLMs bei direkter numerischer Instruktion („Decile 8 von 10") häufig Schwierigkeiten, die Richtung (höher = riskanter) und die Größe der Abstände korrekt zu berücksichtigen. Die kategoriale Kodierung soll diese Fehlerquelle reduzieren und die Wahrscheinlichkeit erhöhen, dass das Modell die Risikoinformation adäquat verarbeitet.[32] Gleichzeitig bleibt der numerische Score sichtbar, um die Informationsäquivalenz zur Originalstudie weitgehend zu wahren.

Jede Vignette folgt – wie auch in der Originalstudie – einer standardisierten Struktur: Demographische Basisinformationen (Geschlecht, Alter), aktueller Tatvorwurf mit Deliktklassifikation (misdemeanor/felony), kriminalhistorische Angaben (Anzahl der Vorstrafen oder der Delikte in Kindheit und Jugendalter) sowie die COMPAS-Risikoprädiktion. Ein Beispiel: *„The defendant is a female aged 43. They have been charged with:*

31 Vgl. *Engel/McAdams* (Fn. 5), 247 m.w.N.

32 *Imani et al.*, Mathprompter: Mathematical reasoning using large language models, Preprint arXiv:2303.05398 (2023); *Engel/McAdams* (Fn. 5), 268.

Driving under the influence. This crime is classified as a misdemeanor. They have been convicted of 0 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record. The COMPAS tool made the following prediction about this defendants' general recidivism score: General Recidivism Risk: low DECILE:1"

Wie in der Originalstudie werden für jede Vignette drei abhängige Variablen erhoben:

- Q1 (Recidivism Likelihood) auf einer 5-Punkt-Skala von „Extremely unlikely" bis „Extremely likely". Dies erfasst die subjektive Rückfallwahrscheinlichkeit.
- Q2 (Grant Bail) als binäre Entscheidung (Yes/No)
- Q3 (Confidence) auf einer 5-Punkt-Skala von „Completely guessing" bis „Completely confident". Dies misst die Entscheidungssicherheit.

Die Pipeline unterstützt optional kurze Begründungstexte (2–3 Sätze), die mehreren Zielen dienen können: (i) Qualitative Validierung, ob das LLM die experimentelle Manipulation tatsächlich verarbeitet hat (zum Beispiel durch explizite Referenzen auf den Warntext in der Treatment-Gruppe); (ii) Reasoning-Transparenz durch Exploration der Mechanismen hinter den Entscheidungen (nutzt das Modell stereotype Heuristiken oder differenzierte Abwägungen?); (iii) Sycophancy-Detektion durch Identifizierung von Fällen, in denen das Modell übermäßig „hilfreich" oder sozial erwünscht antwortet, statt realistische Varianz zu zeigen.[33] Gerade neuere LLMs werden zunehmend auf „Assistenz"-Verhalten optimiert, was zur Unterdrückung kontroverser oder sozial unerwünschter Antworten führen kann – ein Analogon zum Social-Desirability-Bias in Umfragen mit menschlichen Teilnehmerinnen und Teilnehmern.[34] Die Anforderung strukturierter Begründungen folgt zudem der Empfehlung von *Abdurahman et al.* (2025): „[A]uthors must rigorously demonstrate that the LLM possesses the necessary capabilities and cognitive processes inherent to human-data production."[35]. Die Begründungsoption wurde in der finalen Simulation zwar deaktiviert, kam indes in mehreren Testläufen zur Anwendung.

33 *Anthis et al.* (Fn. 2), 3.

34 *Anthis et al.* (Fn. 2), 3.

35 *Abdurahman et al.*, A Primer for Evaluating Large Language Models in Social-Science Research, Advances in Methods and Practices in Psychological Science 8 (2025), 15.

4. Modul 3: Durchführung der Simulation

Modul 3 führt im Anschluss die eigentliche Simulation durch – wiederum mittels einer mehrstufigen Pipeline, die für jede Kombination (Agent × Vignette × Bedingung × Modell) strukturierte Vorhersagen generiert. Der Prozess gliedert sich in vier Hauptkomponenten: (1) Prompt-Generierung aus einer Prompt-Bank, (2) API-Aufruf mit Rate-Limiting, (3) Response-Parsing und Normalisierung, (4) Aggregation über Wiederholungen. Operativ kommen zwei Modellfamilien zum Einsatz: GPT-4o (OpenAI, November 2024) und Claude 3.5 Sonnet (Anthropic, Oktober 2024). Beide werden getrennt konfiguriert und aufgerufen; Effekte, die in beiden Modellfamilien auftreten, gelten als robuster.

Anstatt das Modell in eine Rolle schlüpfen zu lassen („Du bist ein 52-jähriger…"), instruiert der Prompt das LLM ausdrücklich als analytischen Prädiktor: Es soll – gestützt auf das demografische Agentenprofil, die Instruktionen und den Falltext – vorhersagen, wie eine reale Person mit diesen Merkmalen typischerweise antworten würde. Diese Perspektivverschiebung soll Stereotyp-„Performances" entgegenwirken und die Nachvollziehbarkeit der erzeugten Antworten erhöhen. Eine zentrale methodische Herausforderung bei LLM-basierten Simulationen ist die Prompt-Lotterie – die Abhängigkeit der Ergebnisse von idiosynkratischen Formulierungsdetails.[36] So besteht ein wesentlicher Unterschied darin, ob man ein Modell nach seiner eigenen Einschätzung oder der „richtigen Antwort" fragt oder es explizit um seine Einschätzung bittet, wie eine näher beschriebene Person die Frage beantworten wird.[37] Teilweise führen schon kleine Unterschiede bei der Eingabe zu gänzlich unterschiedlichen Ergebnissen. Um diese Empfindlichkeit zu kontrollieren, wird eine Prompt-Bank implementiert. Das ist eine Sammlung von zehn inhaltlich äquivalenten, stilistisch aber variierten Prompts pro experimenteller Bedingung. Für jede Kombination (Agent × Vignette × Bedingung × Modell) wird zufällig eine Variante gezogen, wobei die Antwortstruktur stets identisch bleibt (Q1 Rückfallerwartung, Q2 Bail-Entscheidung, Q3 Entscheidungssicherheit). Dieses Ensembling über Prompts folgt der von *Hewitt et al.* (2024)[38] vorgeschlagenen Simulationsstrategie und erhöht nachweis-

36 Vgl. *von der Heyde/Haensch/Wenz*, Vox Populi, Vox AI? Using Large Language Models to Estimate German Vote Choice, Social Science Computer Review (2025), 2.

37 *Von der Heyde/Haensch/Wenz* (Fn. 35), 2; *Kruse* (Fn. 4), 5.

38 *Hewitt et al.* (Fn. 3).

lich die Robustheit gegenüber einzelnen Formulierungsartefakten. Optional kann der Prompt – wie soeben dargestellt – kurze Begründungen anfordern. Der Kern-Prompt lautet (auszugsweise) wie folgt:

SYSTEM PREAMBLE:

You are an expert in behavioral prediction and survey response analysis.

You are NOT role-playing as the participant. You are ANALYZING and PREDICTING likely responses based on the participant profile and the case.

PARTICIPANT PROFILE: {profile}

INSTRUCTIONS SHOWN TO PARTICIPANT: {instructions}

CASE PRESENTED TO PARTICIPANT: {vignette}

ALGORITHMIC INFORMATION: {compas_prediction}

[Only in compas_warning:]

WARNING TO PARTICIPANT: {warning_text}

YOUR TASK:

Provide your PROFILE-BASED PREDICTION:

Q1_Recidivism: „Extremely unlikely" | „Somewhat unlikely" | „Neither likely nor unlikely" | „Somewhat likely" | „Extremely likely"

Q2_Grant_Bail: „Yes" | „No"

Q3_Confidence: „Completely guessing" | „Somewhat guessing" | „Neutral" | „Somewhat confident" | „Completely confident"

CRITICAL RULES:

– Use EXACTLY the labels above for Q1/Q2/Q3 (no alternatives).

– Provide one definitive answer per question.

[Optional; in Tests aktiviert, im finalen Lauf deaktiviert:]

JUSTIFICATION: Provide 2–3 sentences explaining how the profile and case features lead to these predictions.

Die Entscheidung für GPT-4o und Claude 3.5 Sonnet beruht auf Erfahrungswerten zur Verwendung LLM-gestützter Simulationen. In anderen Kontexten konnten mit den beiden hier verwendeten Modellen die besten Vorhersagen erzielt werden.[39] Die Verwendung zweier Modelle dient auch der Robustheitsprüfung. Während modellspezifische Unterschiede auf Artefakte der Trainings- oder Alignment-Strategie hindeuten können, sprechen konsistente Effekte über beide Modelle hinweg für eine Genera-

39 Vgl. *Kruse* (Fn. 4) für GPT-4o.

lisierbarkeit der Befunde. Die Wahl der Temperature (0,7) ist das Ergebnis eines Kompromisses:[40] Einerseits sollen die Modelle nicht zu deterministisch antworten, da sich anderenfalls die Varianz in den Antworten menschlicher Versuchspersonen nicht abbilden ließe. Andererseits soll die Streuung nicht so groß sein, dass ein zufälliges Rauschen systematische Effekte überlagern könnte.

Operativ werden pro Kombination (Agent × Vignette × Bedingung × Modell) fünf Wiederholungen mit wechselnden Prompt-Varianten ausgeführt. Die Antworten werden zur besseren Vergleichbarkeit in feste Kategorien übersetzt. Bei Q2 (Yes/No) zählt die Mehrheit; bei Q1 und Q3 die häufigste Kategorie als Ergebnis. Zusätzlich berechnet das System Konsistenz- und Unsicherheitsmaße über die fünf Wiederholungen (u. a. Agreement-Rate, Varianz, normalisierte Entropie) und fasst sie in einem zusammengesetzten Zuverlässigkeitsindikator Model_Confidence zusammen, der hohe Übereinstimmung, niedrige Entropie und vollständige Struktur der Antworten belohnt. Am Ende werden lediglich die aggregierten Werte ausgegeben.

5. Modul 4: Analyse

Modul 4 analysiert die Ergebnisse und vergleicht sie mit den Ergebnissen der Originalstudie. Im Unterschied zur Originalstudie (50 Vignetten) erhalten die KI-Agenten jeweils nur 10 Fallbeschreibungen. Ein Vergleich mit den Gesamtergebnissen der Originalstudie ist daher nicht sinnvoll möglich. Vielmehr müssen die Ergebnisse vignettenspezifisch verglichen werden. Der insoweit erforderliche Zugriff auf die Rohdaten der Originalstudie stand den Autoren während der Arbeit an dem vorliegenden Beitrag nur eingeschränkt zur Verfügung (auch um das Überraschungsmoment der Festschrift nicht zu gefährden). Ein umfassender Vergleich konnte nur mit den Ergebnissen der ersten der in der Originalstudie berichteten Teilstudien durchgeführt werden. Wie dargestellt verwendet die erste Teilstudie einen sequentiellen experimentellen Ablauf: Die Versuchspersonen entscheiden zunächst ohne Kenntnis der maschinellen Risikoeinschätzung. Sodann erhalten sie den COMPAS-Score und die Möglichkeit, ihre vorherige Einschätzung anzupassen. In der vorliegen-

40 Und folgt im Übrigen den Empfehlungen vergleichbarer Studien, siehe etwa *von der Heyde/Haensch/Wenz* (Fn. 35), 8 f.

den Studie hingegen entscheiden die KI-Agenten pro Vignette nur einmal (wobei ihnen stets die maschinelle Risikoeinschätzung zur Verfügung steht); dies entspricht dem Vorgehen in der zweiten Teilstudie der Originalstudie. Vor diesem Hintergrund werden die Simulationsergebnisse im Folgenden mit den abschließenden, unter Kenntnis des Risiko-Scores getätigten Angaben aus der ersten Teilstudie verglichen. Dies schränkt die Aussagekraft der folgenden Analysen selbstverständlich dahingehend ein, dass Unterschiede zwischen der Simulations- und der Originalstudie auch darauf beruhen könnten, dass die Daten aus der Originalstudie in dieser Hinsicht abweichend erhoben wurden.

Wie erwähnt wurde die Simulation mit zwei LLMs durchgeführt. Für die Auswertung werden beide Modelle zunächst gemeinsam behandelt und die Ergebnisse in „gepoolter" Form berichtet. Erst in einem zweiten Schritt werden die Modellunterschiede näher beleuchtet. Die unterschiedlichen LLMs werden genutzt, um das Entscheidungsverhalten einzelner KI-Agenten vorherzusagen. Die KI-Agenten haben aber für beide Modelle jeweils ein identisches demografisches Profil. Hier stellt sich die Frage, ob die Beobachtungen zwischen den LLMs aufgrund der gemeinsamen demografischen Basis der KI-Agenten als abhängige Beobachtungen zu behandeln sind oder ob sie dennoch unabhängige Beobachtungen darstellen. Für die Abhängigkeit spricht, dass für jeden KI-Agenten ein individualisiertes Profil mit Namen, Alter, Beruf und weiteren Merkmalen erstellt wurde. Man könnte also annehmen, in beiden Modellen entscheide „derselbe" KI-Agent. Für die Unabhängigkeit spricht, dass die demografischen Merkmale pro KI-Agent immer noch sehr begrenzt sind und es sich eben nicht um dieselbe Person mit weiteren unbekannten Eigenschaften handelt, sondern dass idiosynkratisches Entscheidungsverhalten erst aus der Kombination der demografischen Basis mit dem jeweiligen LLM entsteht: Ein individueller KI-Agent entstünde also erst durch die konkrete Entscheidungsvorhersage für ein bestimmtes demografisches Profil. In der folgenden Analyse werden die Beobachtungen zwischen den LLMs trotz der gemeinsamen demografischen Basis als unabhängig behandelt.

Die statistische Analyse beruht in einem ersten Zugriff auf zweiseitigen t-Test-Vergleichen der Durchschnittswerte pro Versuchsperson/Agent. Diese Tests berücksichtigen allerdings nicht, dass pro Person/Agent mehrere Entscheidungen (10 Vignetten) getroffen wurden. Um dem Rechnung zu tragen, wurden zusätzlich Mehrebenenregressionen (Mixed-Effects-Modelle) geschätzt. Aus Darstellungsgründen werden im Folgenden in erster Linie die Ergebnisse der t-Tests über Durchschnitts-

werte berichtet. Vereinzelt werden jedoch ergänzend Ergebnisse der Regressionsanalysen eingebracht, die aufgrund der präziseren Modellierung der Datenstruktur grundsätzlich aussagekräftiger sind (höhere Teststärke).

III. Analyseergebnisse

1. Treatmentunterschiede

Tabelle 1 zeigt die durchschnittlichen Bail-Entscheidungen, Konfidenzeinschätzungen und die subjektive Risikowahrscheinlichkeit für die Originalstudie sowie für die Simulationsstudie (jeweils für die Kontroll- und die Treatmentgruppe).

Tabelle 1: Treatmentunterschiede[41]

	Ohne Warnung (Kontrollgruppe)		**Mit Warnung (Treatmentgruppe)**	
	Menschen (Original, N=67)	LLM (Simulation, N=268)	Menschen (Original, N=67)	LLM (Simulation, N=268)
Bail	.70	.72	.70	.72
Konfidenz	.92	1.34	.89	1.22
Rückfälligkeit	–.15	–.42	–.09	–.27

Die Anteile der Pro-Bail-Entscheidungen mit und ohne Warnungen sind in der Originalstudie deskriptiv sehr ähnlich; die Bewertung der Konfidenz fällt ebenfalls ähnlich aus. Schließlich wird die Rückfallwahrscheinlichkeit in der Treatmentgruppe nur unwesentlich höher eingeschätzt als in der Kontrollgruppe. Signifikante Unterschiede zwischen Kontroll- und Treatmentgruppe lassen sich für keine der abhängigen Variablen nachweisen (t-Tests, N=134, ps > .52). Auch unter Verwendung der LLM-Simulation sind die absoluten Werte zwischen den Versuchsanordnungen ähnlich. Für die (abschließende) Bail-Entscheidung lassen sich keine

41 Für Bail wird der Anteil der Pro-Bail-Entscheidungen angegeben. Die Konfidenz und Rückfallwahrscheinlichkeit wurden auf Skalen von -2 bis 2 (in 5 Ausprägungen) erhoben. Angegeben werden Durchschnittswerte.

signifikanten Unterschiede mit und ohne Warnung nachweisen (t-Test, N=536; p = .75). Die deskriptiv zum Teil eher geringen Unterschiede in der Konfidenz und Rückfallwahrscheinlichkeit sind jedoch statistisch signifikant (t-Tests, N=536, ps < .01). Dabei ist auch zu berücksichtigen, dass die Beobachtungszahlen in der Simulationsstudie deutlich höher sind als in der Originalstudie.

2. Unterschiede zwischen den Erhebungsmethoden

Der Anteil der Pro-Bail-Entscheidungen der menschlichen Teilnehmer:innen und der KI-Agenten ist in beiden Experimentalgruppen mit gerundet 70 % bzw. 72 % sehr ähnlich – entsprechend findet sich kein statistisch signifikanter Unterschied zwischen beiden Erhebungsmethoden (t-Tests, N=335, ps >.25). Allerdings ist die Bewertung der Konfidenz für die KI-Agenten sowohl mit als auch ohne Warnungen deutlich höher als die der menschlichen Versuchspersonen (t-Tests, N=335, ps < .01). Die Rückfallwahrscheinlichkeit wird von den KI-Agenten in beiden Gruppen leicht, aber signifikant niedriger eingeschätzt (t-Tests, N=335, ps < .01).

3. Modellspezifische Unterschiede

Die bisherige Analyse beruht auf „gepoolten" Werten, in die sowohl die Simulation durch Claude als auch die von GPT generierten Antworten eingeflossen sind. Während ein solches Vorgehen eine gewisse Robustheit gegen Idiosynkrasien einzelner KI-Modelle verspricht, kann es die Heterogenität verdecken, die mit der Nutzung unterschiedlicher Modelle einhergehen kann. Vor diesem Hintergrund wurden die gepoolten Ergebnisse aufgelöst. Die modellspezifischen Werte sind in Tabelle 2 dargestellt (zu Vergleichszwecken enthält diese wiederum auch die Werte der Originalstudie).

Tabelle 2 zeigt zunächst, dass das Antwortverhalten der Agenten zwischen den beiden Modellen teilweise unterschiedlich ist. Die Bail-Entscheidungen aber fallen zwischen beiden Modellen schon deskriptiv sehr ähnlich aus. Für den Anteil der Pro-Bail-Entscheidungen lassen sich weder in der Kontrollgruppe noch in der Treatmentgruppe signifikante Unterschiede zwischen den LLMs nachweisen (t-Tests, N=268, ps > .26). Dagegen fallen die Bewertungen der Konfidenz und der Rückfallwahrscheinlichkeit zwischen beiden Modellen leicht unterschiedlich aus. Das Modell Claude gibt sowohl eine durchschnittlich höhere Rückfallwahr-

Tabelle 2: Modellspezifische Unterschiede[42]

	Ohne Warnung (Kontrollgruppe)			**Mit Warnung (Treatmentgruppe)**		
	Original (N=67)	Claude (N=134)	GPT (N=134)	Original (N=67)	Claude (N=134)	GPT (N=134)
Bail	.70	.73	.71	.70	.73	.71
Konfidenz	.92	1.41	1.27	.89	1.32	1.12
Rückfall-wahrscheinlichkeit	–.15	–.37	–.47	–.09	–.20	–.35

scheinlichkeit als auch eine höhere Konfidenz für Entscheidungen der KI-Agenten in beiden Experimentalgruppen an (t-Tests, N=268, ps < .05).

Allerdings handelt es sich hierbei um Level-Unterschiede der beiden Modelle. Sowohl die Unterschiede zwischen den Experimentalgruppen als auch die Unterschiede zur Originalstudie sind pro Modell qualitativ vergleichbar mit den Ergebnissen der über die beiden Modelle aggregierten Bewertungen. Die Bail-Entscheidungen sind unter Verwendung beider Modelle mit und ohne Warnungen wiederum sehr ähnlich; statistische Unterschiede zeigen sich nicht (t-Tests, N=268, ps > .78). Die Konfidenz ist in beiden Modellen ohne Warnung höher als mit Warnung und die Rückfallwahrscheinlichkeit ohne Warnung niedriger als mit Warnung (t-Tests, N=268, ps < .01). Verglichen mit der Originalstudie finden sich für den Anteil der Pro-Bail-Entscheidungen keine signifikanten Unterschiede (t-Tests, N=201, GPT ps > .52, Claude ps > .18), wogegen die Konfidenz durch beide LLMs signifikant höher und die Rückfallwahrscheinlichkeit grundsätzlich signifikant niedriger eingeschätzt wird (t-Tests, N=201, ps < .01). Die Einschätzung der Rückfallwahrscheinlichkeit durch Claude und in der Originalstudie ist aber in der Treatmentgruppe nur marginal signifikant unterschiedlich, ein t-Test ergibt einen p-Wert von 0.099, eine Regressionsanalyse, die je Treatment ein Modell mit beiden LLMs schätzt (Anzahl der Gruppen = 335), einen p-Wert von 0.065.

42 Für Bail wird der Anteil der Pro-Bail-Entscheidungen angegeben. Die Konfidenz und Rückfallwahrscheinlichkeit wurden auf Skalen von -2 bis 2 (in 5 Ausprägungen) erhoben. Berichtet werden Durchschnittswerte.

4. Vignettenunterschiede

Die die obigen Ergebnisse bestätigenden – hier aber nicht berichteten – Regressionsanalysen berücksichtigen bereits statistisch, dass Antworten für mehrere Vignetten abgegeben wurden. Im Folgenden sollen die vignettenspezifischen Werte auch grafisch dargestellt werden. Die nachstehenden Abbildungen (2a, 2b sowie 2c) stellen die Ergebnisse der LLMs denen der Originalstudie (Menschen) gegenüber, jeweils für die Kontroll- sowie die Treatmentgruppe. Jede Abbildung deckt dabei eine der drei abhängigen Variablen ab. Es zeigt sich, dass die LLM-Bewertungen über die Vignetten hinweg eine deutlich geringere Varianz aufweisen. Während die LLM-Bewertungen die menschlichen Abbildungen etwa bei den Bail-Entscheidungen im Durchschnitt gut abbilden, weichen die Bewertungen der einzelnen Vignetten deutlicher voneinander ab. Ein ähnliches Bild ergibt sich für die Rückfallwahrscheinlichkeit. Die KI-Agenten geben hier im Durchschnitt leicht niedrigere Rückfallraten an als die menschlichen Versuchspersonen. Allerdings sind die Unterschiede zwischen den Einschätzungen der unterschiedlichen Vignetten bei Menschen deutlich größer als bei den KI-Agenten. Sowohl bei der Bail-Entscheidung als auch bei der Einschätzung der Rückfallwahrscheinlichkeit scheinen die menschlichen Versuchspersonen also extremer auf die Vignetteninformationen (wozu auch die Risiko-Scores gehören) zu reagieren, als die LLMs dies erwarten.

5. Erweiterung der Originalstudie: Richter-Agenten

Auf aggregierter Ebene bilden die Simulationsergebnisse die Ergebnisse der Originalstudie zumindest in der Grundtendenz gut ab (siehe Tabelle 1). Dies gilt insbesondere für die Bail-Entscheidung. Demgegenüber ergeben sich bei der Bewertung der Konfidenz sowie der Rückfallwahrscheinlichkeit qualitative Unterschiede. Schlüsselt man die Ergebnisse in Bezug auf die einzelnen Vignetten auf, so beruht selbst die durchschnittliche Übereinstimmung bei den Bail-Entscheidungen auf einer deutlich unterschiedlichen Verteilung der Einzelbewertungen.

Geht man vor diesem Hintergrund davon aus, dass LLM-gestützte KI-Agenten menschliche Laienentscheidungen zumindest grundsätzlich zutreffend abbilden können, so eröffnen sich weitergehende Forschungsperspektiven. Dem eigentlichen Erkenntnisinteresse der Originalstudie (den Effekt der Warnungen auf Richterinnen und Richter zu ergründen) folgend, könnte man etwa versuchen, das Verhalten echter Richter zu

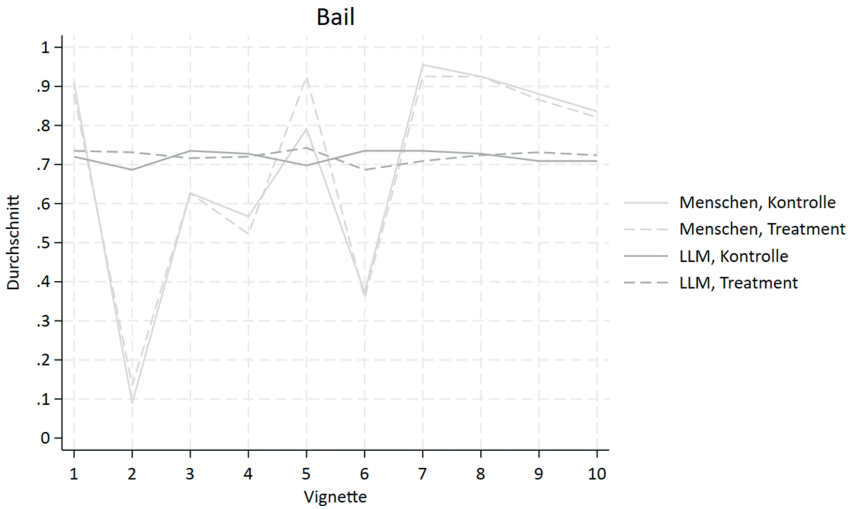

Abbildung 2a: Vignettenunterschiede für Bail-Entscheidung

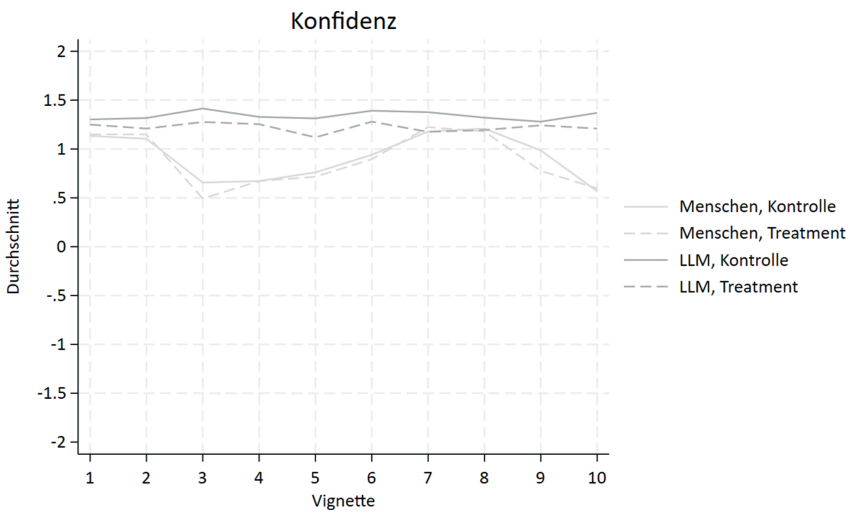

Abbildung 2b: Vignettenunterschiede für Konfidenz

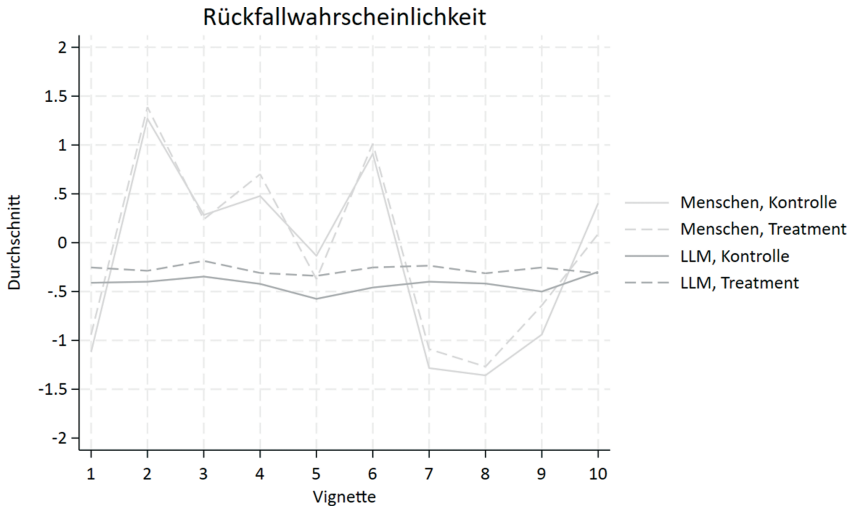

Abbildung 2c: Vignettenunterschiede für Rückfallwahrscheinlichkeit

simulieren. In diesem Sinne wurde im Zug eines Testlaufs ein weiteres Agenten-Sample (mit individuellen demografischen Profilen) erstellt, das nun aber US-amerikanische (Bundes-)Richterinnen und Richter abbilden sollte. Da die Erstellung der Agenten sowie die Bewertung dabei an einigen Stellen von der finalen Simulation abweicht, ist ein Vergleich mit den KI-Agenten als Laien nur begrenzt möglich. Die Untersuchung der Wirkung der Warnungen auf Richter-Agenten ist eher explorativ und wie die gesamten Ergebnisse vorläufiger Natur. Um die darin zum Ausdruck kommenden weitergehenden Forschungsmöglichkeiten zu unterstreichen, sollen die Ergebnisse gleichwohl kursorisch dargestellt werden.

Wie Tabelle 3 zu entnehmen ist, zeigt sich für die KI-Agenten in der Richter-Rolle auf allen drei Bewertungsdimensionen ein signifikanter Effekt der Warnung.

In Gegenwart einer Warnung treffen die Richter-Agenten seltener Pro-Bail-Entscheidungen und sind sich ihrer Einschätzung zudem weniger sicher. Zugleich schätzen sie die Rückfallwahrscheinlichkeit im Durchschnitt leicht geringer ein, als dies ohne Warnung der Fall ist. Dieser etwas überraschende Befund könnte mit den Reaktionen auf einzelne Vignetten zu erklären sein. Zwar ist die Pro-Bail-Rate mit Warnung insgesamt niedriger, eine deutliche Reduktion ergibt sich aber allein für eine spezifische

Tabelle 3: Simulationsergebnisse „echte" Richter

	Ohne Warnung (Kontrollgruppe) (N=134)	**Mit Warnung (Treatment-gruppe)(N=134)**	**t-Test (p-Wert, N=268)**
Bail	.80	.72	.00
Konfidenz	.88	.80	.04
Rückfallwahrscheinlichkeit	.38	.26	.01

Vignette (Vignette 6). Die Rückfallwahrscheinlichkeit wird mit Warnung grundsätzlich niedriger eingeschätzt. Allein für besagte Vignette 6 zeigt sich ein gegenteiliger Effekt: Hier hat die Warnung eine (spürbare) Erhöhung der angenommenen Rückfallwahrscheinlichkeit zur Folge.

Abschließend soll noch auf eine wesentliche Einschränkung des Einsatzes von KI-Agenten zur Simulation des Verhaltens „echter" Richterinnen und Richter hingewiesen werden. So lässt sich zumindest die Frage aufwerfen, ob das Verhalten von Richterinnen und Richtern in den LLM-Trainingsdaten in einem Ausmaß repräsentiert ist, als dass die Modelle verlässliche Verhaltensannahmen ausbilden könnten.

C. Ausblick

Am Anfang dieser Untersuchung stand die Frage, ob und, wenn ja, mit welcher Vorgehensweise sich rechtsempirische Experimente LLM-gestützt replizieren lassen und was dabei zu beachten ist. Der vorliegende Versuch zeigt, dass konzeptionelle Replikationen mit KI-Agenten durchaus im Bereich des Möglichen liegen. Der wichtigste Befund der Originalstudie (die Warnungen führen nicht zu einem nachweisbaren Effekt auf das Entscheidungsverhalten) konnte mit der vorliegenden Simulation repliziert werden: Wie schon die Originalstudie zeigt auch die Simulationsstudie keine Unterschiede für Bail-Entscheidungen mit und ohne Warnung. Auch in Bezug auf die Konfidenz sowie die Rückfallwahrscheinlichkeit sind die Bewertungstendenzen zwischen Kontroll- und Treatmentgruppe in der Simulation und in der Originalstudie deskriptiv ähnlich. Treatmentunterschiede sind allerdings in der Simulation deutlicher und statistisch signifikant. Insoweit stehen die vorliegenden Ergebnisse durchaus in Einklang mit den Beobachtungen größer angelegter Untersuchun-

gen, wonach sich eine Vielzahl von Studien mit KI-Agenten grundsätzlich replizieren lässt.

Solche, eher basalen, Übereinstimmungen dürfen freilich nicht zu dem Schluss verleiten, LLM-gestützte KI-Agenten könnten bereits heute als Substitut für menschliche Versuchspersonen fungieren, entsprechende Studien also bereits ersetzen.[43] Dies zeigt auch der vorliegende Replikationsversuch. Betrachtet man dessen Ergebnisse etwa vignettenspezifisch, so treten teilweise erhebliche Abweichungen vom Verhalten der menschlichen Versuchspersonen hervor. Allgemein weist das Antwortverhalten der KI-Agenten im Vergleich zu den menschlichen Versuchspersonen eine geringere Variabilität auf. Dies dürfte zwar teilweise methodisch induziert sein (die Werte pro KI-Agent sind bereits über mehrere Befragungen aggregiert). In erster Linie zeigt sich hier aber wohl eine generelle Eigenschaft oder Schwäche LLM-gestützter Simulationen. So wird in der Literatur über unterschiedliche Studien hinweg berichtet, dass das Antwortverhalten von LLMs im Vergleich zu menschlichen Versuchspersonen eine deutlich geringere Varianz aufweise.[44] Schon aus diesem Grund sind KI-Agenten kein Substitut für menschliche Versuchspersonen und ist bei der Extrapolation der so gewonnen Erkenntnisse Vorsicht geboten. Rechtlich relevante Folgerungen sollte man grundsätzlich erst dann wagen, wenn die Ergebnisse einer LLM-gestützten Simulation durch verhaltenswissenschaftliche Untersuchungen mit menschlichen Teilnehmern untermauert wurden.

Auch und gerade jenseits der Replikationsfrage fördert die vorliegende Studie berichtenswerte Einsichten zu Tage. Für die empirische Rechtsforschung ist insbesondere das Verhältnis zur bisherigen Methodik von Interesse. Ähnlich sind sich beide Wege vor allem auf konzeptioneller Ebene. Anders als man intuitiv annehmen mag, handelt es sich bei LLM-gestützten Simulationen keineswegs um eine Art (rechts-)empirische For-

43 *Harding et* al., AI language models cannot replace human research participants, AI & Society 39 (2024), 2603; *Wang et* al., Large language models that replace human participants can harmfully misportray and flatten identity groups, Nature Machine Intelligence 7 (2025), 400.

44 Vgl. nur *Abdurahman et al.*, Perils and opportunities in using large language models in psychological research, PNAS Nexus 3 (2024), 245; *Saynova et al.*, Identifying Non-Replicable Social Science Studies with Language Models, Preprint arXiv:2503.10671 (2025); *Bisbee et al.*, The Perils of Large Language Models, Political Analysis 32 (2024), 401; *Park et al.*, Diminished diversity-of-thought in a standard large language model, Behav Res Methods 56 (2024), 5754.

schung light, die man ohne einen theoretisch-konzeptionellen Unterbau aufsetzen könnte. Wie das hier beschriebene Vorgehen zeigt, erschöpft sich eine Simulation nicht in der Formulierung einer Handvoll Prompts. Gedanken über den für eine Forschungsfrage geeigneten experimentellen Aufbau und seine Implementierung (in einem LLM-Setting) muss man sich weiterhin machen. Viele Arbeitsschritte, die bei der Vorbereitung experimenteller Studien anfallen, müssen konzeptionell vergleichbar auch bei LLM-gestützten Experimenten vollzogen werden. So musste vorliegend etwa eine komplexe Pipeline entwickelt und umgesetzt werden, die das Untersuchungsdesign der Originalstudie experimentell sauber abbildet. Dem ging eine Auseinandersetzung mit der Originalstudie voraus.

Relevante Unterschiede (zugunsten LLM-gestützter Simulationen) ergeben sich vor allem unter forschungspraktischen Gesichtspunkten. Obgleich auch LLM-gestützte Simulationen keineswegs kostenlos sind (für die API-Nutzung können zumindest bei größer angelegten Untersuchungen schnell Kosten im hohen dreistelligen Bereich anfallen), fallen die Kosten im Vergleich zu Studien mit Menschen dennoch deutlich geringer aus. Hinzu kommt, dass einige praktische Begrenzungen (etwa mit Blick auf die Verfügbarkeit von Versuchspersonen), die sich nur bedingt kontrollieren lassen, im LLM-Experiment nicht auftreten. So sind Beobachtungszahlen relativ einfach skalierbar. Auch demografische Besonderheiten verfügbarer Versuchspersonen können in der Simulation einfach ausgeglichen werden. Soweit demografische Daten über die für die Forschungsfrage relevanten Gruppen zur Verfügung stehen, lassen sich KI-Agenten mit entsprechenden Eigenschaften ohne größeren Aufwand generieren. Die Erstellung von Richter:innen als KI-Agenten in der vorliegenden Simulation stellt gerade für die empirische Rechtsforschung eine vielversprechende Option dar. Dies macht empirische Rechtsforschung nicht nur zugänglicher (niedrigere Einstiegshürden), sondern auch flexibler. KI-Agenten und LLM-gestützte Simulationen könnten das Feld daher nachhaltig beleben und möglicherweise einen rechtsexperimentellen Innovationsschub auslösen. Forschende können die unterschiedlichsten experimentellen Designs (vorab) ausprobieren, praktisch unbegrenzt (noch so kleinteilige) Anpassungen vornehmen und selbst fernliegende (treffender: bislang als fernliegend erscheinende) Ansätze und/oder Erklärungen testen.

Vor diesem Hintergrund kann der vorliegende Werkstattbericht auch als Plädoyer für eine innovativere und kreativere (empirische) Rechtsforschung verstanden werden – und das dürfte ganz im Sinn des von *Christoph Engel* gelebten Forschungsverständnisses sein.