



583 Bericht von Shan Huang, Renke Schmacker und Hannes Ullrich

KI-Diagnose überzeugt Hausärzt*innen weniger als ein gleichwertiger Schnelltest

- Experiment vergleicht Einfluss einer KI-Diagnose mit etabliertem Schnelltest
- Ärzt*innen passen erste Einschätzung nach KI-Diagnose weniger stark an als nach Schnelltest
- Schulungen und nutzerorientierte Gestaltung von KI-Werkzeugen entscheidend, damit diagnostisches Potenzial von KI umgesetzt wird

590 Interview mit Hannes Ullrich

592 Kommentar von Johannes Geyer

Konzept der Rentenkommission hat eine entscheidende Lücke

Zur Lage der Privatindustrie in der Sowjetzone Deutschlands (SBZ)

Die Privatindustrie der SBZ
Nach offiziellen sowjetzonalen Angaben

Jahr	Wert der Bruttoproduktion ¹⁾		Arbeiter ²⁾ und Angestellte in 1000	Produktionswert je Beschäftigten in DM
	in Mill. DM	in vH der Gesamtproduktion		
1950	5 484,2	23,5	578	9 488
1951	.	.	594	.
1952	.	.	574	.
1953	5 400,7	14,5	488	11 067
1954	6 253,9	15,2	492	12 711
1955	6 504,3	14,7	498	13 061

¹⁾ Meßwert auf Preisbasis 1950. — ²⁾ Berechnet nach „Statistische Praxis“, Ost-Berlin, Jg. 1956, Nr. 7, S. 89.

© CC BY 4.0

Seit 1945 ist die Privatindustrie in der SBZ durch Sequestrierungen (Umwandlung von Privatbetrieben in SAG-Betriebe), durch die Enteignungen ab Sommer 1946 und durch steuerliche und strafrechtliche Maßnahmen seitens der Zonenbehörden aus einer Reihe von Zweigen der Grundstoff- und Schwerindustrie sowie der metallverarbeitenden Industrie fast völlig verdrängt worden. Dieser Prozeß der gewaltsamen Industriereform nach den Leitsätzen des Marxismus-Leninismus wurde ab 1950 durch geräuschlosere Methoden des planwirtschaftlichen Dirigismus ergänzt, die der Wirtschaftsverwaltung der SBZ erlaubten, von Fall zu Fall die Privatindustrie als Nothelfer einzusetzen und ihre Bedeutung offiziell hervorzuheben, ohne ihr so viel Spielraum zu lassen, daß sie ein eigenes Gewicht wiedererlangen und sich aus den Fesseln der Planwirtschaft hätte lösen können.

Aus dem Wochenbericht Nr. 38 vom 21. September 1956

IMPRESSUM



DIW Berlin — Deutsches Institut für Wirtschaftsforschung e. V.

Anton-Wilhelm-Amo-Straße 58, 10117 Berlin

www.diw.de

Telefon: +49 30 897 89-0 E-Mail: kundenservice@diw.de

93. Jahrgang 16. September 2026

Herausgeber*innen

Prof. Anna Bindler, Ph.D.; Prof. Dr. Tomaso Duso; Sabine Fiedler; Prof. Marcel Fratzscher, Ph.D.; Prof. Dr. Peter Haan; Prof. Dr. Claudia Kemfert; Prof. Dr. Alexander S. Kritikos; Prof. Dr. Alexander Kriwoluzky; Prof. Karsten Neuhoﬀ, Ph.D.; Prof. Dr. Sabine Zinn

Chefredaktion

Prof. Dr. Pio Baake; Claudia Cohnen-Beck; Sebastian Kollmann; Kristina van Deuverden

Lektorat

Hermann Buslei

Redaktion

Dr. Hella Engerer; Petra Jasper; Adam Mark Lederer; Frederik Schulz-Greve; Sandra Tubik

Gestaltung

Roman Wilhelm; Stefanie Reeg; Eva Kretschmer, DIW Berlin

Umschlagmotiv

© imageBROKER / Steffen Diemer

Satz

Satz-Rechen-Zentrum Hartmann + Heenemann GmbH & Co. KG, Berlin

Der DIW Wochenbericht ist kostenfrei unter www.diw.de/wochenbericht abrufbar. Abonnieren Sie auch unseren Wochenberichts-Newsletter unter www.diw.de/wb-anmeldung

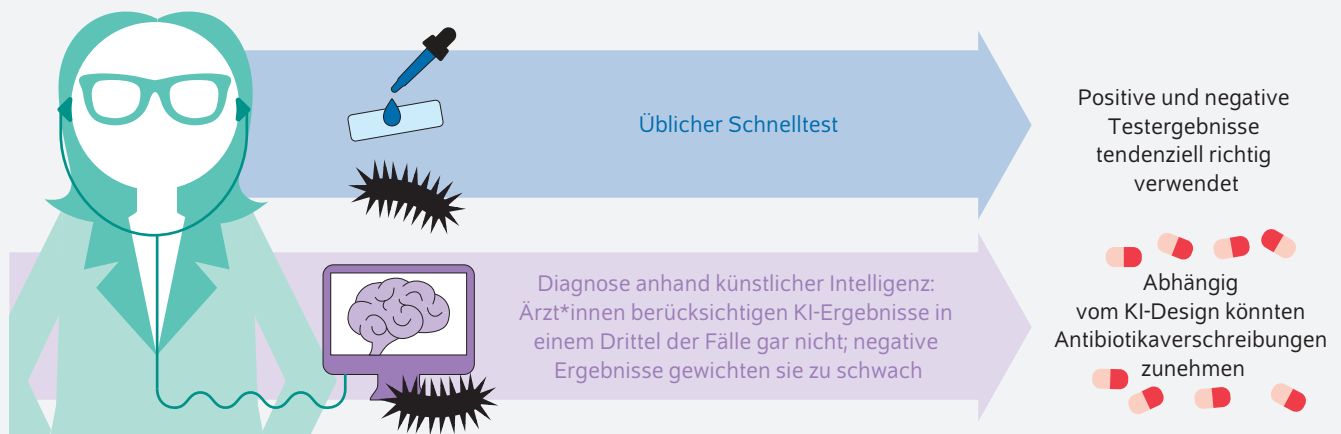
ISSN 1860-8787

KI-Diagnose überzeugt Hausärzt*innen weniger als ein gleichwertiger Schnelltest

Von Shan Huang, Renke Schmacker und Hannes Ullrich

- Experiment mit dänischen Hausärzt*innen vergleicht Einfluss einer KI-Diagnose mit vertrautem Schnelltest bei identischem Informationsgehalt
- Ärzt*innen passen Einschätzung eines Harnwegsinfektionsrisikos nach KI-Diagnose um 41 Prozent weniger stark an als nach Urin-Schnelltest
- Etwa ein Drittel der Ärzt*innen ignoriert KI-Diagnose fast vollständig, aus Skepsis und Unsicherheit über den Nutzen der KI
- Wer KI nutzt, gewichtet negative KI-Ergebnisse zu schwach und überschätzt redundante Testergebnisse; Antibiotikaverschreibungen könnten unnötigerweise zunehmen
- Schulungen und eine nutzerorientierte Gestaltung von KI-Werkzeugen sind entscheidend, damit Hausärzt*innen diagnostisches Potenzial von KI umsetzen

Harnwegsinfektion: Im Vergleich zu etabliertem Schnelltest werden KI-Diagnosen trotz gleichem Informationsgehalt seltener berücksichtigt und häufiger nicht richtig verwendet



Quelle: Eigene Berechnungen.

© CC BY 4.0, creativecommons.org/licenses/by/4.0

ZITAT

„Der Nutzen einer KI hängt nicht allein von ihrer statistischen Genauigkeit ab, sondern auch davon, ob und wie Ärzt*innen sie in ihre Diagnostik einbeziehen. Schulungen und eine durchdachte Gestaltung der Werkzeuge sowie Lernen durch Experimentieren sind entscheidend, damit das Potenzial von KI in der Praxis ausgeschöpft werden kann.“ — Hannes Ullrich —

MEDIATHEK



Audio-Interview mit Hannes Ullrich
www.diw.de/mediathek

KI-Diagnose überzeugt Hausärzt*innen weniger als ein gleichwertiger Schnelltest

Von Shan Huang, Renke Schmacker und Hannes Ullrich

ABSTRACT

Künstliche Intelligenz (KI) hält auch in die hausärztliche Versorgung zunehmend Einzug. Allerdings ist wenig darüber bekannt, wie Ärzt*innen KI-generierte Diagnoseinformationen mit etablierten Testverfahren und ihren eigenen klinischen Einschätzungen kombinieren. Ein Experiment mit 372 dänischen Hausärzt*innen analysiert, ob sich die Einschätzung des Risikos einer Harnwegsinfektion nach einer KI-Diagnose beziehungsweise nach einem vertrauten Urin-Schnelltest unterscheidet, wenn beide Verfahren den gleichen Informationsgehalt haben. Ärzt*innen passen ihre Einschätzung nach einer KI-Diagnose um 41 Prozent weniger stark an als nach dem Schnelltest. Etwa ein Drittel der Ärzt*innen ignoriert die KI-Diagnose fast vollständig. Wer die KI nutzt, gewichtet ein negatives KI-Ergebnis systematisch zu schwach und überschätzt den Zusatznutzen davon, den Urin-Schnelltest zu sehen, wenn dieser bereits von der KI verwendet wird. Diese Verzerrungen könnten die Antibiotikaverschreibungen bei Anwendung von KI unnötig erhöhen. Die Ergebnisse der Studie sprechen dafür, dass Schulungen und eine durchdachte Gestaltung von KI-Diagnosewerkzeugen entscheidend sind, damit Hausärzt*innen das Potenzial von KI in der Praxis umsetzen.

Die verbreitete Anwendung von künstlicher Intelligenz (KI) verspricht spürbare Produktivitätsgewinne: im Kundenservice, beim professionellen Schreiben und in der Softwareentwicklung erhöht KI-Unterstützung die Menge und Qualität von Arbeitsergebnissen, insbesondere bei weniger erfahrenen Beschäftigten. Bei einer breiteren Betrachtung des Arbeitsmarkts über Beschäftigtengruppen hinweg fallen die gemessenen Produktivitäts- und Lohneffekte sowie die Verbreitung von KI dagegen sehr unterschiedlich aus.¹

Dies gilt vor allem in Sektoren, in denen Fachkräfte über spezifische Expertise und Erfahrung verfügen und Fehlentscheidungen hohe Folgekosten haben können. Hier sind Entscheidungen die Ergebnisse von Abwägungen menschlicher Expert*innen, die verfügbare etablierte, aber auch zunehmend KI-generierte Informationen mit ihrem eigenen kontext-spezifischen Wissen verknüpfen müssen. Der Nutzen einer KI hängt daher nicht nur von überprüfbaren technischen Merkmalen ab, sondern auch davon, ob und wie spezialisierte Fachkräfte KI-Informationen nutzen und mit ihrem Wissen kombinieren.²

Die hausärztliche Versorgung ist ein Paradebeispiel für ein solches Entscheidungsumfeld: Diagnostische Informationen sind knapp und Entscheidungen müssen unter Zeitdruck getroffen werden. Eine am DIW Berlin entstandene Studie untersucht diesen Bereich anhand einer landesweiten Umfrage und eines Experiments mit dänischen Hausärzt*innen.³ An der Untersuchung, durchgeführt im Jahr 2025, nahmen 21,5 Prozent aller dänischen Hausarztpraxen

¹ Vgl. Shakked Noy und Whitney Zhang (2023): Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science*, 381(6654), 187–192; Anders Humlum und Emilie Vestergaard (2025): The Unequal Adoption of ChatGPT Exacerbates Existing Inequalities among Workers. *Proceedings of the National Academy of Sciences*, 122(1), e2414972121.

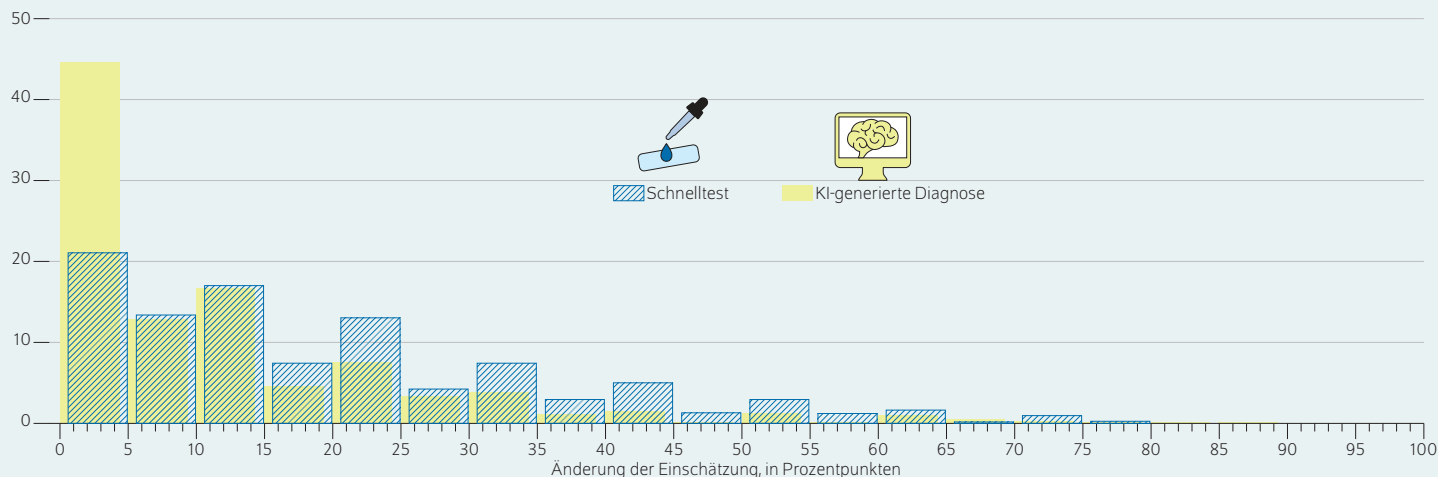
² Erste Evidenz hierzu liefern ein Experiment mit Radiolog*innen, in dem eine Aufteilung von Entscheidungen und nicht die Kombination von Wissen zwischen der KI und Radiolog*innen am produktivsten war, sowie ein Feldexperiment in nigerianischen Praxen, in dem sich klinische Abläufe durch die Nutzung generativer KI zwar verbesserten, aber nicht die eigentlichen Behandlungsergebnisse. Vgl. Nikhil Agarwal et al. (2023): Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. NBER Working Paper Nr. 31422; und Jason Abaluck et al. (2026): Does LLM Assistance Improve Healthcare Delivery? An Evaluation Using On-Site Physicians and Laboratory Tests. NBER Working Paper Nr. 34660.

³ Vgl. Shan Huang, Renke Schmacker und Hannes Ullrich (2026): AI Adoption by Human Experts: Evidence from Primary Care Physicians. DIW Discussion Paper Nr. 2168 (online verfügbar, abgerufen am 4. September 2026. Dies gilt auch für alle anderen Online-Quellen dieses Berichts, sofern nicht anders vermerkt).

Abbildung

Anpassungen der Infektionseinschätzung nach Signalquelle

Häufigkeit in Prozent



Lesehilfe: In rund 45 Prozent der Fälle ändern Ärzt*innen ihre Einschätzung über die Wahrscheinlichkeit einer Infektion nach einer KI-generierten Diagnose um weniger als fünf Prozentpunkte. Bei einem Schnelltest geschieht dies in nur rund 20 Prozent der Fälle.

Quelle: Eigene Berechnungen.

© CC BY 4.0

Ärzt*innen passen ihre Einschätzung nach einem Signal der KI deutlich schwächer an als nach dem Schnelltest-Ergebnis.

teil, insgesamt 372 Hausärzt*innen. KI wird dort bereits heute genutzt: 48 Prozent der Befragten gaben an, generative KI-Werkzeuge in der Praxis zu nutzen; eine vergleichbare Befragung unter britischen Hausärzt*innen kam ein Jahr zuvor auf einen Anteil von 20 Prozent.⁴ Rund 49 Prozent der dänischen Befragten erwarten einen hohen Nutzen von KI bei der Diagnostik. KI wird zwar vermehrt für Diagnosezwecke entwickelt und zugelassen, aber es ist unklar, ob und wie KI bei konkreten diagnostischen Entscheidungen genutzt wird.

Im Experiment der DIW-Studie wird die Qualität der diagnostischen Information konstant gehalten und nur ihre Quelle verändert. So kann untersucht werden, wie Hausärzt*innen eine KI-generierte Diagnose im Vergleich zu einem etablierten Schnelltest in Diagnose- und Verschreibungsentscheidungen einbeziehen. Hierfür werden Behandlungssituationen betrachtet, in denen die Diagnose von Harnwegsinfektionen – eine der am häufigsten behandelten Erkrankungen in dänischen Hausarztpraxen⁵ – im Mittelpunkt stehen.

Bakteriell verursachte Harnwegsinfektionen werden hauptsächlich mit Antibiotika behandelt. Um Antibiotikaresistenzen einzudämmen, sollten Antibiotika jedoch nur eingesetzt werden, wenn die Diagnose hinreichend sicher ist. Dies stellt Ärzt*innen vor eine Herausforderung, da typische

Symptome einer Harnwegsinfektion auch durch nicht-bakterielle Ursachen hervorgerufen werden können. Beim ersten Kontakt wird in der Regel eine Symptombeschreibung aufgenommen, über mögliche Faktoren, die eine Infektion begünstigen können, gesprochen, und eine Urinprobe abgenommen, die anhand eines ungenauen Schnelltests vor Ort oder in einem Labor sehr viel genauer untersucht werden kann.⁶ Eine gesicherte Diagnose aus dem Labor liegt häufig erst nach zwei bis vier Tagen vor. Die Entscheidung, ob unmittelbar mit einer Antibiotikabehandlung begonnen oder gewartet wird, fällt jedoch bereits beim ersten Patientenkontakt. Gerade in dieser Situation kann KI einen Mehrwert bieten, indem sie zusätzliche, zum Zeitpunkt der Behandlung verfügbare Informationen (zum Beispiel aus früheren Infektionen, Laborergebnissen, Krankenhausaufenthalten oder Verschreibungen) nutzt, um die Wahrscheinlichkeit einer bakteriellen Infektion vorherzusagen.⁷

Ärzt*innen reagieren auf KI-Diagnose schwächer als auf den vertrauten Schnelltest

Um den Effekt der Signalquelle (Schnelltest oder KI-Diagnose) von deren inhaltlicher Qualität zu trennen, entwickelt

⁴ Vgl. Charlotte R. Blease et al. (2024): Generative Artificial Intelligence in Primary Care: An Online Survey of UK General Practitioners. *BMJ Health & Care Informatics*, 31(1), e101102.

⁵ Vgl. Betsy Foxman (2002): Epidemiology of Urinary Tract Infections: Incidence, Morbidity, and Economic Costs. *The American Journal of Medicine*, 113 (1, Supplement 1), 5–13.

⁶ Ärzt*innen stützen sich meist auf eine Reihe von Urin-Schnelltests, die Hinweise auf Leukozyten und Nitrit liefern, aber nur eingeschränkt genau sind. Vgl. Anastasiya Chernaya, Christian Søborg und Mette Midttun (2022): Validity of the Urinary Dipstick Test in the Diagnosis of Urinary Tract Infections in Adults. *Danish Medical Journal*, 69(1), A07210607.

⁷ Konkret können Methoden des maschinellen Lernens anhand zahlreicher beobachtbarer Faktoren den Befund der mikrobiologischen Laboranalyse vorhersagen. Vgl. Michael A. Ribers und Hannes Ullrich (2019): Künstliche Intelligenz und Daten können bei der Eindämmung von Antibiotikaresistenzen helfen. *DIW Wochenbericht* Nr. 19, 335–341 (online verfügbar).

Kasten 1

Optimale Verwendung zusätzlicher Informationen

Rational urteilende, sogenannte bayesianische Beobachter*innen sollten ihre Einschätzung der Wahrscheinlichkeit, dass ein bakterieller Infekt vorliegt, proportional zur Aussagekraft eines neuen Signals aktualisieren. Die Aussagekraft hängt vom Verhältnis der Wahrscheinlichkeiten ab, ein bestimmtes Ergebnis bei vorliegender beziehungsweise nicht vorliegender Erkrankung zu beobachten. Da der betrachtete Schnelltest eine höhere Sensitivität als Spezifität aufweist, ist ein negatives Ergebnis aussagekräftiger als ein positives – ein negativer Befund schließt eine Infektion vergleichsweise sicher aus.

Die Studie schätzt für jede Ärztin und jeden Arzt, wie stark die tatsächliche Anpassung ihrer Einschätzung von der bayesianischen Vorhersage abweicht, getrennt für positive und negative Signale.¹ Für den Schnelltest liegen beide Werte etwas übereins, und der Unterschied zwischen positiven und negativen Signalen entspricht der theoretisch erwarteten Asymmetrie. Für das KI-Signal liegt der Wert für negative Ergebnisse dagegen deutlich unter dem für positive Ergebnisse. Ärzt*innen setzen ein negatives KI-Signal in ihrer Einschätzung mit deutlich geringerem Gewicht um als bei einer bayesianischen Nutzung der Information zu erwarten wäre. Dies weist darauf hin, dass sie die unterschiedliche Aussagekraft positiver und negativer KI-Ergebnisse nicht erkennen oder nicht anwenden (Tabelle).

Tabelle

Anpassung der Einschätzungen im Vergleich zur Bayes-Regel

Koeffizienten eines aus der Bayes-Regel abgeleiteten Regressionsmodells

	positives Signal	negatives Signal
Schnelltest	1,15	1,17
KI-Diagnose	1,26	0,77

Anmerkung: Ein Wert von 1,00 entspricht der bayesianischen Anpassung von Einschätzungen nach einem Signal, Werte darunter bedeuten eine Unterreaktion auf das Signal, Werte darüber eine Überreaktion.

Lesehilfe: Der Wert 0,77 zeigt an, dass die geschätzte Wahrscheinlichkeit eines bakteriellen Effekts nach einem negativen Signal der KI deutlich zu hoch ist.

Quelle: Eigene Darstellung.

© CC BY 4.0

¹ Vgl. David M. Grether (1980): Bayes Rule as a Descriptive Model: The Representativeness Heuristic. Quarterly Journal of Economics, 95(3), 537–557.

die Studie ein kontrolliertes Experiment: 372 Hausärzt*innen aus 346 Praxen – mehr als ein Fünftel aller dänischen Hausarztpraxen – beurteilen hypothetische Patientenfälle mit typischen Symptomen einer Harnwegsinfektion. Für jeden Fall schätzen sie zunächst die Wahrscheinlichkeit einer bakteriellen Infektion und treffen eine vorläufige Behandlungsentscheidung. Anschließend erhalten sie, zufällig zugeteilt, entweder eine KI-Diagnose oder ein Schnelltest-Ergebnis. Im Fall der KI-Diagnose wird beschrieben, welche Arten von Daten in die Analyse einfließen (einschließlich des Schnelltest-Ergebnisses). Ebenso wird angegeben, dass das verwendete KI-Modell erfolgreich klinische Studien durchlaufen hat. Sowohl Schnelltest als auch KI liefern ein einfaches Ergebnis in Form von „positiv“ oder „negativ“, mit einer identischen statistischen Genauigkeit (80 Prozent Sensitivität, 60 Prozent Spezifität).⁸ Auf Grundlage der jeweils erhaltenen Informationen können die Ärzt*innen ihre Erstein-schätzung beibehalten oder anpassen. Wer ausschließlich auf den Informationsgehalt reagiert, sollte die Einschätzung unabhängig davon ändern, ob das Ergebnis von der KI oder vom Schnelltest stammt.

Im vorliegenden Experiment weichen die Einschätzungen der Ärzt*innen zwischen den beiden Signalquellen jedoch deutlich voneinander ab. Nach einem Schnelltest-Ergebnis ändern die Ärzt*innen ihre Einschätzung der Infektionswahrscheinlichkeit im Schnitt um 18,5 Prozentpunkte, nach einer KI-Diagnose dagegen nur um 10,9 Prozentpunkte (Abbildung). Das entspricht einer um 41 Prozent schwächeren Reaktion auf die KI. Der Unterschied lässt sich nicht durch eine allgemeine Skepsis gegenüber den mitgeteilten Genauigkeitswerten erklären: Weitere Tests der Studie zeigen, dass Ärzt*innen Angaben zur diagnostischen Güte durchaus ernst nehmen und ihre Einschätzung entsprechend anpassen, wenn ihnen höhere Sensitivitäts- und Spezifitätswerte mitgeteilt werden, sowohl beim Schnelltest-Signal als auch beim KI-Signal.

Etwa ein Drittel der Ärzt*innen ignoriert die KI nahezu vollständig

Hinter dem Unterschied in der durchschnittlichen Nutzung verbergen sich deutliche Unterschiede zwischen einzelnen Ärzt*innen. Rund ein Drittel (34 Prozent) der Ärzt*innen reagiert so gut wie gar nicht auf das KI-Signal: In den meisten Fällen verändern sie ihre Einschätzung um weniger als drei Prozentpunkte. Die übrigen zwei Drittel beziehen die KI-Information dagegen substanziell in ihre Entscheidung ein und reagieren auf ein positives KI-Signal sogar genauso stark wie auf ein positives Schnelltest-Ergebnis. Auf ein negatives KI-Signal reagieren jedoch auch sie weniger stark.

Wer das KI-Signal nicht nutzt, unterscheidet sich von den Nutzenden kaum in Alter, Geschlecht oder Praxisgröße – auch nicht in Kennzahlen der Verschreibungsqualität wie

⁸ Die Sensitivität ist die Wahrscheinlichkeit, dass der Test positiv ausfällt, wenn eine Infektion vorliegt. Die Spezifität ist die Wahrscheinlichkeit, dass der Test negativ ausfällt, wenn keine Infektion vorliegt.

der Rate leitliniengerechter Antibiotikawahl oder vermeidbarer Krankenhauseinweisungen wegen Harnwegsinfektionen. Auch die allgemeine Vertrautheit mit generativer KI unterscheidet sich nicht zwischen denen, die das KI-Signal nutzen, und denen, die das KI-Signal nicht nutzen. Allerdings arbeiten Ärzt*innen, die die KI im Experiment nutzen, häufiger in Praxen, die bereits digitale Technologien wie Video- oder E-Mail-Sprechstunden verwenden. Auch aufwändigere Diagnose-Werkzeuge wie die Mikroskopie, deren Verwendung spezieller Schulungen bedarf, kommen dort häufiger zum Einsatz. Die Nutzung des KI-Diagnosewerkzeugs scheint damit eher der allgemeinen digitalen und diagnostischen Ausrichtung einer Praxis zu folgen als individuellen Eigenschaften der Ärzt*innen oder ihrer bisherigen Verschreibungspraxis.

In offenen Freitextantworten begründen KI-Nutzende ihre Entscheidung meist damit, dass die KI zusätzliche, sonst nicht verfügbare Informationen verwende und mit einer nachvollziehbaren Genauigkeit arbeite. Ärzt*innen, die die KI im Experiment nicht nutzen, äußern dagegen Skepsis gegenüber dem Werkzeug, empfinden es als intransparent und sind unsicher, wie sie das Ergebnis in ihre Entscheidung einordnen sollen.⁹ Aus den Freitextantworten lässt sich die geringere Nutzung der KI damit insbesondere auf Unsicherheiten darüber zurückführen, ob und wie das Werkzeug nützliche Informationen liefert.

Systematische Unterschiede bei Nutzung der KI

Auch unter den Ärzt*innen, die das KI-Signal grundsätzlich nutzen, zeigen sich systematische Verzerrungen. Die Studie identifiziert zwei Muster, nach welchen die Ärzt*innen von einer optimalen, sogenannten bayesianischen Berücksichtigung von Information, abweichen (Kasten 1). Diese Abweichungen spiegeln sich nicht nur in den Einschätzungen von Wahrscheinlichkeiten wider, sondern übertragen sich auch auf die Verschreibungsentscheidungen.

Stärkere Aussagekraft negativer KI-Signale wird ignoriert

Der verwendete, weitläufig bekannte Schnelltest weist eine höhere Sensitivität (80 Prozent) als Spezifität (60 Prozent) auf. Das heißt, er ist besser darin, eine Infektion korrekt auszuschließen, als sie korrekt zu bestätigen. Ein negatives Ergebnis ist deshalb aussagekräftiger als ein positives und sollte die Einschätzung stärker verändern. In der medizinischen Praxis etabliert sich diese Logik häufig über Merkgeln: Ein hochsensitiver Test hilft, eine Erkrankung durch ein negatives Ergebnis auszuschließen.

Im Experiment reagieren Ärzt*innen auf ein negatives Schnelltest-Ergebnis deutlich stärker als auf ein positives.

Sie wenden die Regel also korrekt an, auch wenn die über eins liegenden Werte auf eine etwas zu starke Gewichtung hinweisen. Für das KI-Signal gilt dies nicht, obwohl eine mit dem Schnelltest identische Sensitivität und Spezifität kommuniziert wird. Ärzt*innen reagieren nahezu gleich stark auf ein positives und ein negatives KI-Ergebnis, obwohl das negative Signal objektiv fast doppelt so aussagekräftig ist.

Korrelierte Signale werden als unabhängig angesehen

Die Studie untersucht zusätzlich, ob Ärzt*innen ein KI-Signal abhängig davon verwenden, ob es bereits bekannte Informationen beinhaltet oder vollständig auf tatsächlich neuen Informationen beruht, die für die Ärzt*innen nicht direkt zugänglich sind. Vor allem bei der Verwendung von Signalen, die keine neuen Informationen beinhalten, sind Verzerrungen bei der Anpassung von Einschätzungen ein bekanntes Phänomen (Kasten 2).

Um dies zu untersuchen, erhalten die Ärzt*innen in einem zweiten Teil des Experiments sowohl ein KI-Signal als auch ein Schnelltest-Ergebnis. Der zusätzlich gezeigte Schnelltest bietet keine neue Information und ist für eine rational urteilende Person redundant.¹⁰ Für eine zufällig gewählte Hälfte wird das KI-Signal jedoch so konstruiert, dass es den Schnelltest, anders als im ersten Teil, nicht beinhaltet. Für diese Hälfte sind beide Signale tatsächlich unabhängig voneinander und liefern gemeinsam mehr Information als jedes einzelne für sich. Eine bayesianisch urteilende Person sollte in dieser Situation deutlich stärker auf das zusätzliche Signal reagieren.

Die Ärzt*innen unterscheiden die beiden Situationen jedoch kaum. Ihre Einschätzung ändert sich nach dem zusätzlichen Schnelltest-Ergebnis nahezu identisch, unabhängig davon, ob dieses tatsächlich neue Information enthält oder nicht. Dies kann weder durch bloße Verwirrung über das neu eingeführte Signal noch dadurch erklärt werden, dass Ärzt*innen das KI-Signal ignorieren und sich nur auf den Schnelltest verlassen. Ein ähnlicher Umgang mit verschiedenen Informationsquellen wurde bereits bei Radiolog*innen für bildgebende KI-Signale dokumentiert, was auf eine grundlegende Herausforderung hindeutet.¹¹

Antibiotikaverschreibungen könnten zunehmen

Im Experiment verschreiben Ärzt*innen nach einem positiven KI-Signal (KI sagt Status „bakterielle Infektion“ voraus) nicht signifikant mehr Antibiotika als nach einem positiven Schnelltest. Dagegen verschreiben sie bei einem negativen KI-Signal 31 Prozent häufiger Antibiotika als nach einem negativen Schnelltest-Ergebnis.

⁹ Haftungs- oder Verantwortungsfragen, die in politischen Debatten zu KI im Vordergrund stehen, werden nicht als Grund angegeben. Dies könnte sich durch das dänische Haftungsrecht erklären lassen, in dem einzelne Ärzt*innen bei Behandlungsfehlern kaum persönlich haftbar gemacht werden. Vgl. Patienterstatninger: Oplysningspligt; Styrelsen for Patientsikkerhed (2026): Tilsyn i tal, Mai 2026 (online verfügbar).

¹⁰ Da das KI-Signal das Schnelltest-Ergebnis enthält, ist dieses vollständig redundant, wenn die KI die Information aus dem Schnelltest optimal verwendet. Ist dies nicht der Fall, nimmt die Redundanz ab.

¹¹ Vgl. Agarwal et al. (2023), a. a. O.

Kasten 2

Verwendung von Signalen ohne zusätzlichen Informationsgehalt

Ein in der Verhaltensökonomik bekanntes Phänomen betrifft den Umgang mit mehreren Informationsquellen, bei dem Akteure den gemeinsamen und überlappenden Informationsgehalt verschiedener Quellen nicht berücksichtigen.¹ Wenn Signale, die teilweise auf denselben zugrunde liegenden Informationen beruhen, wie unabhängige Beobachtungen behandelt werden, können die daraus resultierenden Einschätzungen als übermäßig sicher wahrgenommen werden.

Zwei Testergebnisse sind nur dann jeweils eigenständig informativ, wenn sie unabhängig voneinander zustande kommen. Ist ein Ergebnis bereits in einem anderen enthalten, liefert es so gut wie keine zusätzliche Information. Eine rational urteilende Person sollte ihre Einschätzung dann nicht weiter anpassen.

Nutzt zum Beispiel ein KI-Werkzeug das Ergebnis eines Schnelltests optimal, um seine eigene Einschätzung zu berechnen, ist der separat mitgeteilte Schnelltest für eine Person, die das KI-Ergebnis bereits kennt, redundant. Wertet die Person den Schnelltest dennoch als zusätzliche Bestätigung und vernachlässigt damit die Korrelation beider Signale, passt sie ihre Einschätzungen auf Basis des Testergebnisses zu stark an.

In der Studie wird dieses Prinzip direkt getestet, indem identische Fälle einmal nur mit dem KI-Signal und einmal zusätzlich mit dem bereits im KI-Signal enthaltenen Schnelltest-Ergebnis gezeigt werden. Passen Ärzt*innen ihre Verschreibungsentscheidung nach dem redundanten Schnelltest-Ergebnis an, obwohl dieses keine neue Information enthält, ist das ein direkter Beleg für die Vernachlässigung der Korrelation beider Signale.

¹ Vgl. Benjamin Enke und Florian Zimmermann (2019): Correlation Neglect in Belief Formation. *Review of Economic Studies*, 86(1), 313–332.

In dieselbe Richtung beeinflussen die Verzerrungen bei der Verwendung beider Signale die Verschreibungsentscheidungen. Sahen Ärzt*innen zunächst nur ein positives KI-Signal und danach zusätzlich einen positiven, aber redundanten Schnelltest, verdoppelte sich der Anteil der Antibiotikaverschreibungen nahezu.

Fazit: Sorgfältige Gestaltung und Schulung entscheiden über den Nutzen von KI in der Arztpraxis

Der Nutzen KI-gestützter Diagnostik hängt nicht allein von deren statistischer Genauigkeit ab, sondern auch davon, ob

und wie Ärzt*innen die von der KI bereitgestellte Information verwenden. Ein Drittel der Ärzt*innen in der vorliegenden Studie ignoriert die KI-Diagnose weitgehend, und selbst unter den Nutzenden können unvertraute, asymmetrische und teils redundante Signale zu mehr statt weniger Antibiotikaverschreibungen führen. Angesichts wachsender Antibiotikaresistenzen ist dies ein unerwünschter Nebeneffekt, der sowohl bei der Entwicklung diagnostischer KI als auch bei ihrer Einführung in die Praxis berücksichtigt werden sollte.

Auch in Deutschland gewinnt der KI-Einsatz in der hausärztlichen Versorgung an Bedeutung. Seit dem 1. Oktober 2025 sind alle Praxen verpflichtet, die elektronische Patientenakte zu nutzen, wodurch perspektivisch mehr Patientendaten für diagnostische KI-Anwendungen zur Verfügung stehen könnten.¹² Die Bundesärztekammer betont in einer aktuellen Stellungnahme, dass die Verantwortung für Diagnose und Therapie stets ärztliche Aufgabe bleiben müsse, und warnt ausdrücklich vor unreflektiertem Vertrauen in automatisierte Empfehlungen.¹³ Die vorliegenden Ergebnisse legen jedoch nahe, dass in der Praxis auch das entgegengesetzte Risiko besteht: zu geringes statt zu großes Vertrauen in KI-Signale sowie eine unzureichende Einordnung ihrer statistischen Eigenschaften. Gezielte Schulungen und eine praxisnahe, nutzerorientierte Gestaltung von KI-Werkzeugen können dazu beitragen, beide Risiken der Über- und Unterreaktion zu reduzieren.

Transparenz und Erklärbarkeit genügen allerdings nicht, damit KI richtig eingesetzt wird. In der DIW-Studie bestehen systematische Fehleinschätzungen, obwohl Sensitivität und Spezifität der KI explizit angegeben und erklärt werden. Experimentelle Evidenz zeigt zudem, dass transparente Modelle das Erkennen von KI-Fehlern sogar erschweren können und die Gefahr besteht, dass bestehende Fehleinschätzungen eher bestätigt als korrigiert werden.¹⁴ Die Befragungen dänischer Hausärzt*innen und Patient*innen zeigen, dass die Akzeptanz der KI-Ergebnisse von professioneller Autonomie, einer reibungslosen Einbindung in bestehende Arbeitsabläufe und einer unveränderten Arzt-Patienten-Beziehung abhängt.¹⁵ Die Gestaltungsaufgabe besteht damit darin, ärztliche Kontrolle und Verantwortung zu erhalten und zugleich zu verhindern, dass systematische Fehler in der Informationsverarbeitung den Nutzen der KI entwerten.

¹² Vgl. Bundesministerium für Gesundheit: Die ePA für alle (online verfügbar). Seit dem 1. Oktober 2025 sind bundesweit alle Praxen, Krankenhäuser und Apotheken verpflichtet, die elektronische Patientenakte zu nutzen.

¹³ Vgl. Bundesärztekammer (2025): Stellungnahme „Künstliche Intelligenz in der Medizin“, Vorstand am 14. Januar 2025. *Deutsches Ärzteblatt*, 122(4).

¹⁴ Vgl. Forough Poursabzi-Sangdeh et al. (2021): Manipulating and Measuring Model Interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52; Kevin Bauer, Moritz von Zahn und Oliver Hinz (2023): Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *Information Systems Research*, 34(4), 1582–1602. Andere Evidenz zeigt, dass Erklärungen helfen können, wenn sie sichtbar machen, wie eine KI-Diagnose von der eigenen Logik abweicht. Vgl. Moritz von Zahn et al. (2025): Knowing (Not) to Know: Explainable Artificial Intelligence and Human Metacognition. *Information Systems Research*, im Erscheinen.

¹⁵ Vgl. Julie G. Mikkelsen et al. (2023): Patient Perspectives on Data Sharing regarding Implementing and Using Artificial Intelligence in General Practice – A Qualitative Study. *BMC Health Services Research*, 23(1), 335.

In der DIW-Studie konnte zwar nur eine einmalige KI-Nutzung untersucht werden, die Ergebnisse unterstreichen aber die Bedeutung systematischer Evaluierungen medizinischer KI-Werkzeuge nach ihrer Einführung in die klinische Praxis.¹⁶ Ein aktueller Kommentar in einer führenden medizinischen Fachzeitschrift fordert entsprechend, nicht allein die technische Genauigkeit zu prüfen, sondern zu untersuchen, wie sich ärztliche Einschätzungen, Arbeitsabläufe und letztlich Behandlungsergebnisse durch den Einsatz von KI verändern.¹⁷ Sowohl die Medizinprodukteverordnung als auch

die EU-KI-Verordnung sehen eine Überwachung nach dem Markteintritt vor, die unter anderem die Leistung, Sicherheit und Gebrauchstauglichkeit eines Systems berücksichtigen soll.¹⁸ Da die tatsächliche Leistung solcher Systeme davon abhängt, wie menschliche Expert*innen ihre Ergebnisse interpretieren und mit anderen Informationen kombinieren, sollte das Monitoring auch systematische Fehler und Verzerrungen in der menschlichen Verarbeitung von KI-generierten Informationen erfassen.

16 Insbesondere sollte untersucht werden, wie sich die KI-Nutzung verändert, wenn Vertrauen und Wissen durch wiederholte Anwendung wachsen. Erste Evidenz aus der Radiologie zeigt, dass die Nutzung von KI mit wachsender Erfahrung zunimmt. Auch die Ablehnung von KI nimmt ab, ohne jedoch vollständig zu verschwinden. Vgl. Paul Goldsmith-Pinkham, Chenhao Tan und Alexander K. Zentefis (2026): Human-AI Collaboration in Radiology: The Case of Pulmonary Embolism. arXiv Preprint arXiv:2601.13379.

17 Vgl. Show Us the Evidence for the Value of Medical AI [Editorial] (2026). Nature Medicine, 32, 1163.

18 Vgl. Verordnung (EU) 2017/745 über Medizinprodukte, insbesondere Art. 83–86 zur Überwachung nach dem Inverkehrbringen und Anhang III zu deren Anforderungen (online verfügbar), sowie Verordnung (EU) 2024/1689, insbesondere Art. 72 zur Beobachtung nach dem Inverkehrbringen und Art. 73 zur Meldung schwerwiegender Vorfälle (online verfügbar).

Shan Huang ist Postdoktorandin an der Stockholm School of Economics | shan.huang@hhs.se

Renke Schmacker ist Wissenschaftler in der Abteilung Staat im DIW Berlin und in der Forschungsgruppe Verhalten auf Märkten am Wissenschaftszentrum Berlin für Sozialforschung (WZB) | rschmacker@diw.de

Hannes Ullrich ist stellvertretender Leiter der Abteilung Unternehmen und Märkte im DIW Berlin und Associate Professor an der University of Copenhagen | hullrich@diw.de

JEL: I11, D81, D83, J24, O33

Keywords: expert decision-making, artificial intelligence, healthcare, diagnostic information

© 2026 Der Artikel ist gemäß der Creative-Commons-Lizenz CC BY 4.0 nachnutzbar: <https://creativecommons.org/licenses/by/4.0/>

INTERVIEW



Hannes Ullrich ist stellvertretender Leiter der Abteilung Unternehmen und Märkte im DIW Berlin

„Menschliche Entscheidungen sind bei der Gestaltung von KI maßgeblich“

1. **Herr Ullrich, Sie haben anhand von Daten aus Dänemark untersucht, ob Ärzt*innen bei der Diagnose einer Harnwegsinfektion mehr einer KI oder eher einem herkömmlichen Test vertrauen. Was unterscheidet die KI-Informationen vom herkömmlichen Diagnoseverfahren?** Für die herkömmlichen Diagnoseverfahren für Harnwegsinfekte nehmen Ärzt*innen in der Regel eine Urinprobe vor Ort und wenden dann einen Schnelltest an, denn man möchte schnell wissen, ob man ein Antibiotikum verschreiben muss oder nicht. Diese Schnelltestverfahren sind aber relativ unsicher, und auf exakte Laborergebnisse muss man bis zu drei Tage warten. Mit der KI hingegen bekommt man einen Teil dieser diagnoserelevanten Informationen sofort, denn die KI kann allein anhand der Daten, die die KI für eine spezielle Person kennt, die Wahrscheinlichkeit schätzen, mit der ein Laborergebnis positiv oder negativ ausfallen wird.
2. **Wer liegt denn in der Regel richtiger? Die KI oder der Schnelltest?** In dem Experiment haben wir die KI so definiert, dass sie genau gleich oft richtig oder falsch liegt wie der Schnelltest.
3. **Welcher Diagnose vertrauen die Ärzt*innen mehr?** In unserem Experiment vertrauen die Ärzt*innen im Durchschnitt dem Schnelltest mehr. Genauer gesagt, ignorieren ein Drittel der Ärzt*innen die KI komplett und nutzen die Information daraus gar nicht. Die anderen zwei Drittel verwenden die KI, und zwar im Durchschnitt ähnlich stark wie den Schnelltest.
4. **Sind Ärzt*innen allgemein skeptisch gegenüber KI?** Ja, wir finden einen Grad an Skepsis. In unserem Experiment haben Ärzt*innen geäußert, dass Sie die KI nicht transparent genug finden. Sie verstehen nicht gut genug, wie aus den Patientendaten die Wahrscheinlichkeitsschätzung entstehen kann und glauben im Grunde nicht, dass die KI das kann. Die, die die KI nutzen, sagen in der Tat das Gegenteil. Sie sagen, sie finden die KI sehr präzise und genau. Sie finden das Ergebnis hilfreich, da es eine zusätzliche Informationsquelle liefert.
5. **Kann das Ignorieren der KI-Diagnose am Beispiel einer Harnwegsinfektion zu mehr Antibiotikaverschreibungen führen?** Das Problem ist in diesem Fall nicht das Ignorieren der KI-Diagnose, sondern dass die Ärzt*innen im Experiment sie nicht optimal nutzen. Und zwar übersehen Sie die Unterschiede zwischen einem positiven und negativen Testergebnis. Wie auch der Schnelltest ist die KI besser darin, eine Infektion auszuschließen als sie zu bestätigen. Ein negatives Ergebnis ist deshalb aussagekräftiger als ein positives. Das berücksichtigen die Ärzt*innen in unserem Experiment jedoch nicht und behandeln positive und negative Signale als genau gleich präzise. Weil sie das negative Signal nicht stark genug berücksichtigen, verschreiben Sie im Durchschnitt mehr Antibiotika, wenn Sie die KI-Diagnose verwenden.
6. **Was bedeuten die Ergebnisse Ihrer Studie im Hinblick auf einen zunehmenden Einsatz von KI in der hausärztlichen Behandlung?** Wir denken, dass unsere Studie darauf hinweist, dass wir bei Entwicklung, Einsatz und Auswertung von KI auch die menschlichen Entscheider in den Mittelpunkt stellen sollten. Denn eine KI, und so wollen wir das ja, ist eine Hilfestellung für uns Menschen und übernimmt in der Regel noch nicht komplett automatisiert alle Entscheidungen. Daher ist es ganz essenziell, zu verstehen, wie Expert*innen, wie wir Menschen KI-Informationen benutzen, um Entscheidungen zu fällen. KI wird bereits im Gesundheitsbereich eingesetzt und die Nutzung nimmt schnell zu. Ich glaube, wir brauchen hier sehr viel mehr Erkenntnisse, um zu verstehen, wie KI Entscheidungen in der Praxis verändert.

Das Gespräch führte Erich Wittenberg.

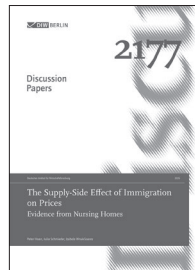
© © Der Artikel ist gemäß der Creative-Commons-Lizenz CC BY 4.0 nachnutzbar: <https://creativecommons.org/licenses/by/4.0/>



Das vollständige Interview zum Anhören finden Sie auf www.diw.de/interview

Discussion Papers Nr. 2177

2026 | Peter Haan, Julia Schmieder, Izabela Wnuk-Soares



The Supply-Side Effect of Immigration on Prices: Evidence from Nursing Homes

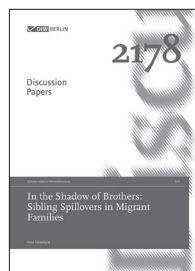
We study how immigration affects prices through changes in labor supply. Using policy-driven increases in immigration to Germany and administrative data covering the universe of nursing homes, we find that one additional foreign-born woman per 100 residents reduces nursing-home care prices by 2.6 percent. The price effects are concentrated in more competitive markets. Examining potential mechanisms, we find that immigration lowers labor costs as nursing homes shift toward lower-paid foreign-born workers. We find no evidence of declines in inspection-based measures of quality or of scale effects. Finally, immigration increases the use of informal care, potentially strengthening competitive pressure from the broader care market.

www.diw.de/publikationen/diskussionspapiere



Discussion Papers Nr. 2178

2026 | Anna Hasselqvist



In the Shadow of Brothers: Sibling Spillovers in Migrant Families

Academic inequality between migrants and natives is often attributed to family dynamics that are difficult to observe. Siblings provide a way to study them: I estimate the effect of having a high-achieving sibling on the educational outcomes of children in migrant and native families. Using exogenous variation in sibling achievement from Sweden's school-entry cutoff in a Regression Discontinuity Design, I show that second-generation migrant girls face a unique disadvantage. A high-achieving brother lowers their grades, while a high-achieving sister has a small positive influence. Spillover effects are negligible for boys and native children regardless of sibling gender. Heterogeneity analysis suggests

that traditional cultural norms contribute to this disadvantage, and parental labor market integration substantially attenuates it.

www.diw.de/publikationen/diskussionspapiere





JOHANNES GEYER

Konzept der Rentenkommission hat eine entscheidende Lücke

Johannes Geyer ist stellvertretender Leiter der Abteilung Staat im DIW Berlin. Der Kommentar gibt die Meinung des Autors wieder.

Über die Empfehlungen der Alterssicherungskommission wird seit Wochen heftig gestritten. Im Fokus stehen die Abschaffung der Rente nach 45 Beitragsjahren und die Minijobs. Wichtiger ist allerdings der Kernauftrag aus dem Koalitionsvertrag: Die Kommission sollte eine „neue Kenngröße für ein Gesamtversorgungsniveau über alle drei Rentensäulen“ entwerfen. Aber genau dort klafft eine Lücke im Konzept.

Als politische Zielgröße einer lebensstandardsichernden Alterssicherung empfiehlt die Kommission für Durchschnittsverdienende eine Nettoersatzquote von mindestens 70 Prozent nach Steuern, für Menschen mit geringerem Einkommen soll sie höher ausfallen. Beim Stellenwert dieser Größe bleibt die Kommission allerdings vage. Die Quote soll kommuniziert werden und Handlungsbedarfe anzeigen, binden soll sie die Politik aber nicht. Offen bleibt auch, ob die 70 Prozent über die gesamte Rentenlaufzeit gelten sollen oder nur beim Rentenzugang – das ist aber ein wichtiger Unterschied.

Praktisch folgt daraus vor allem Statistik: Die Nettoersatzquote soll im Berichtswesen der Bundesregierung etabliert und für typisierte Beispielfälle berechnet werden, die Datenbasis soll besser, ein Statistikkrans für das Monitoring aufgebaut werden. Dazu ist eine lebensbegleitende Finanzbildungsstrategie und der Ausbau der digitalen Rentenübersicht geplant. Alles gut gemeint – aber wäre nicht zuerst zu klären, wie möglichst viele Menschen die 70 Prozent überhaupt erreichen können? Dass es in Teilen der Bevölkerung Vorsorgelücken gibt, ist bekannt. Man müsste priorisieren, was zum Schließen dieser Lücken nötig ist, und das Berichtswesen begleitend nutzen, um nachsteuern zu können. Der Bericht liest sich dagegen, als gäbe es vor allem ein Erkenntnisproblem.

Der elementare Webfehler der Riester-Reform von 2001 wird weder benannt noch behoben: Die zusätzliche Vorsorge wurde nicht obligatorisch eingeführt. Die Kommission bleibt es schuldig zu erklären, wie die Menschen auf die anvisierten Ersatzquoten kommen sollen. Ein Obligatorium schlägt sie weder für die private noch für die betriebliche Vorsorge vor.

Stattdessen soll ein Sozialpartnerdialog die Verbreitung der betrieblichen Altersvorsorge „möglichst flächendeckend“ erhöhen. Dieses Ziel steht seit Jahren im Raum. Schon 2017 nannte Andrea Nahles das Sozialpartnermodell den „Kern des Betriebsrentenstärkungsgesetzes“ und versprach, die seit 2009 anhaltende „komplette Stagnation“ zu beenden. Fast zehn Jahre später stagniert die Verbreitung der betrieblichen Altersvorsorge weiter oder sinkt. Daran hängt tatsächlich, ob diese Reform einen Unterschied macht. Gelingt der Dialog und wird die Zusatzvorsorge flächendeckend, wäre das ein großer Wurf. Gelingt er nicht, bleibt die 70-Prozent-Marke ein Ziel ohne verlässlichen Weg dorthin.

Hinzu kommt: Die Reform ist teuer, vor allem für Beschäftigte und Rentner*innen, und die wachsende kapitalgedeckte Säule erhöht die Komplexität der Vorsorgeentscheidungen erheblich. Wer heute höhere Lasten trägt und morgen vor unübersichtlichen Anlageentscheidungen steht, lässt sich nicht allein mit Kommunikation und gutem Berichtswesen von freiwilliger Zusatzvorsorge überzeugen.

Die 70 Prozent sind ein sinnvoller Kompass. Aber die Kommission delegiert die entscheidende Frage – Verpflichtung ja oder nein? – an einen erst noch zu führenden Dialog der Sozialpartner, die sich bisher vor allem gegenseitig gebremst haben. Damit verbunden ist eine zweite offene Frage: Wie viel Raum soll die geförderte private Vorsorge (Frühstartrente, Altersvorsorge-depot) überhaupt noch bekommen, wenn eine gute, rentierliche und kostengünstige betriebliche Vorsorge in die Fläche käme? Liebe sich die Förderung dann nicht auf die Gruppen konzentrieren, die sie wirklich brauchen? Die eigentliche Weichenstellung dieser Rentenreform steht jedenfalls noch aus.

Dieser Kommentar ist in einer längeren Version am 24. August 2026 im Blog „Politische Ökonomie“ erschienen.

© © Der Artikel ist gemäß der Creative-Commons-Lizenz CC BY 4.0 nachnutzbar: <https://creativecommons.org/licenses/by/4.0/>