Prof. Dr. Stefan Harrendorf
Universität Greifswald
Rechts- und Staatswissenschaftliche Fakultät
Lehrstuhl für Kriminologie, Strafrecht, Strafprozessrecht
und vergleichende Strafrechtswissenschaften

UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456

Prognoseentscheidungen im Justizvollzug durch künstliche Intelligenz

Kann, soll und darf menschliches Verhalten mit Hilfe künstlicher Intelligenz vorhergesagt werden?

50. Arbeits- und Fortbildungstagung der Bundesvereinigung der Anstaltsleiterinnen und Anstaltsleiter im Justizvollzug e.V.
Berlin, 2. bis 6. September 2024

Prognoseentscheidungen im Justizvollzug durch künstliche Intelligenz

1. Einleitung
2. Künstliche Intelligenz
3. Prognoseentscheidungen im Justizvollzug
4. Einsatzgebiete künstlicher Intelligenz im Justizvollzug
5. Treffsicherheit algorithmenbasierter automatisierter Verfahren
6. Rechtliche Grenzen
7. Praktische Probleme
8. Fazit

1. Einleitung

Düsseldorf

Algorithmus soll Straftäter in NRW vor Suizid bewahren

Düsseldorf. · Wie Künstliche Intelligenz aus Sachsen zukünftig Suizidabsichten von NRW-Strafgefangenen erkennen soll.

NIEDERSACHSEN

Künstliche Intelligenz soll Gewalt im Gefängnis verhindern

In Niedersachen erlaubt eine Gesetzesnovelle den Einsatz von KI, um Gefangene vor Gewalt und Suiziden zu schützen. Das Pilotprojekt kostet fast eine Millionen Euro.

Cognify: Forscher malt sich ein virtuelles KI-Gefängnis für Straftäter aus

VON ANDRÉ WESTPHAL JUL 6, 2024 | 34 KOMMENTARE

Aus Zürcher Gefängnissen werden "Smart Prisons"

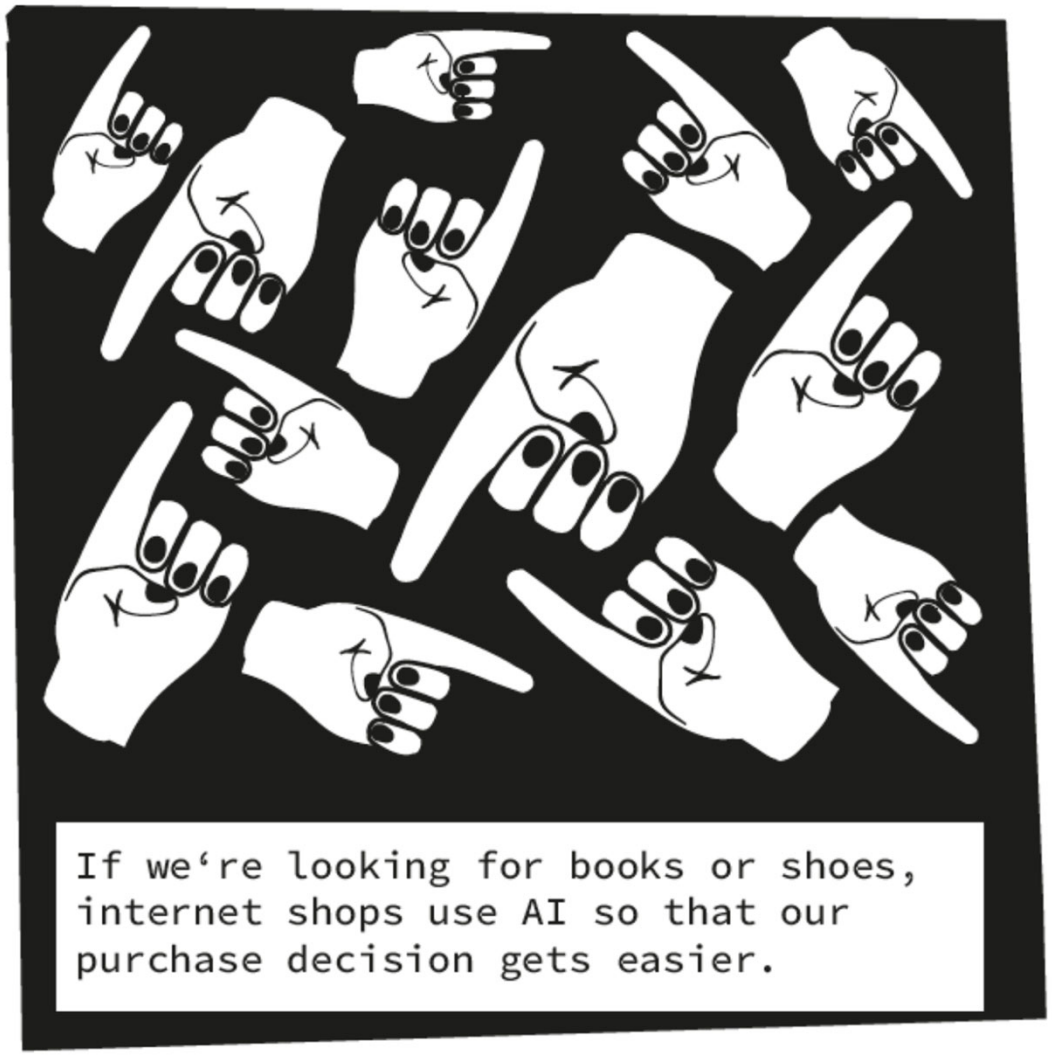
At prisons in Finland, inmates are learning AI and taking online tech courses as a bridge to life on the outside

Daniel T. Allen and Mark Abadi Aug 11, 2020, 3:35 PM MESZ Share Save

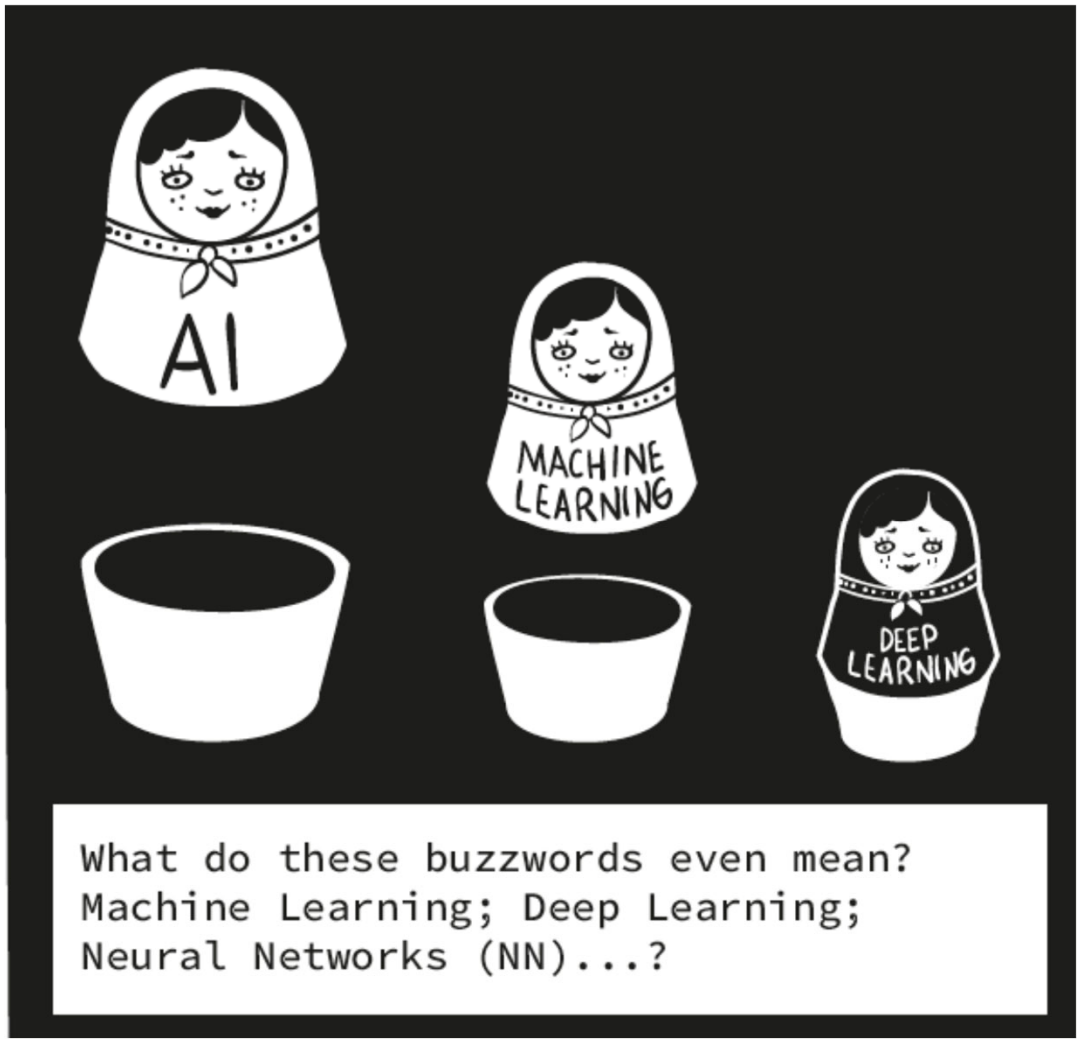
An inmate at the Turku prison in Finland tests out VR equipment. Tom Bateman for Business Insider Weekly

2. Künstliche Intelligenz

- Künstliche Intelligenz (KI) versucht intelligentes menschliches Entscheidungs-verhalten zu imitieren bzw. automatisiertes Verhalten zu ermöglichen.
    - Starke KI = dem Menschen ebenbürtige, umfassende KI-Agenten
        → nicht realisiert und aktuell auch nicht realisierbar
    - Schwache KI = KI-Anwendungen für konkrete Problemlagen
        → Musik-, Film-, Serienempfehlungen in Streamingportalen, Kaufempfehlungen auf Shoppingseiten;
        → Autonomes Fahren
        → Übersetzungstools (z.B. DeepL) oder Chatbots (z.B. ChatGPT) auf der Basis großer Sprachmodelle (LLMs).

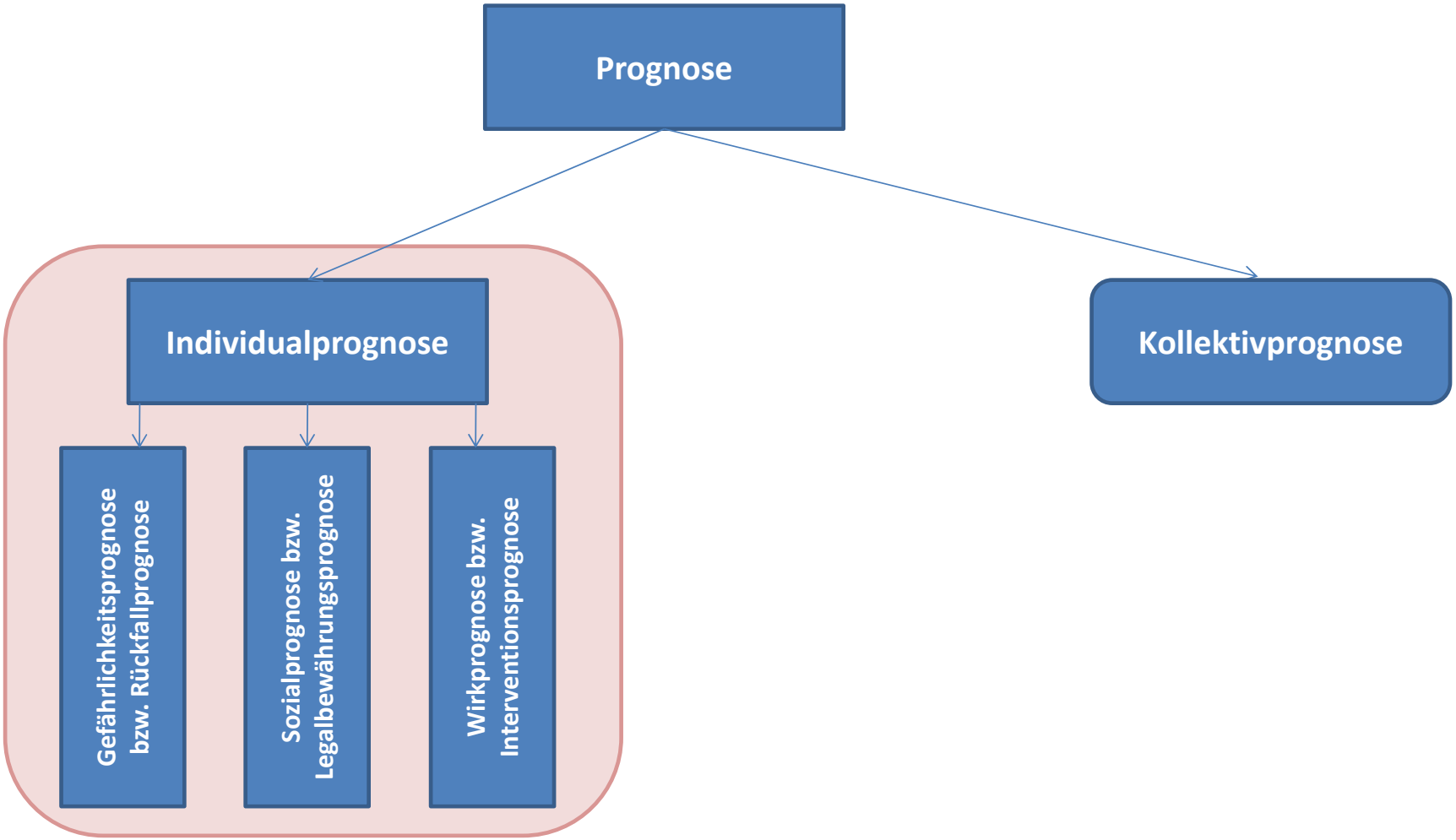


Quelle: Schneider/Ziyal, We Need to Talk, AI: A Comic Essay on Artificial Intelligence, 2019, S. 22



Quelle: Schneider/Ziyal, We Need to Talk, AI: A Comic Essay on Artificial Intelligence, 2019, S. 16

3. Prognoseentscheidungen im Justizvollzug



3. Prognoseentscheidungen im Justizvollzug

Anwendungen im Strafvollzug:

Kollektivprognose

Prognose der Entwicklung der Gefangenenzahlen im Strafvollzug

→ Bau neuer Justizvollzugsanstalten oder Schließung alter

→ Personalplanung

Individualprognose

- Wahrscheinlichkeitsaussagen über künftiges Verhalten, insbesondere Flucht und künftige Straffälligkeit
- Drei Typen:

1. Sozialprognose (positive Verhaltensprognose)

→ Erwartung künftigen Legal- bzw. Wohlverhaltens

2. Gefährlichkeitsprognose (negative Verhaltensprognose)

→ Erwartung/Gefahr weiterer (ggf. erheblicher) Straftaten bzw. der Flucht

3. Wirkprognose / Interventionsprognose

→ Miteinbeziehung der voraussichtlichen Wirkung der Interventionen im Strafvollzug in die Prognose

3. Prognoseentscheidungen im Justizvollzug

Beispiele für die drei Prognosetypen aus dem Kontext des Justizvollzugs:

1. Sozialprognose

→ Eignung für den offenen Vollzug (z.B. § 16 Abs. 2 StVollzG Bln, § 15 Abs. 2 StVollzG M-V, § 12 Abs. 2 NJVollzG)

→ Lockerungsprognose (z.B. § 42 Abs. 2 StVollzG Bln, § 38 Abs. 2 StVollzG M-V, § 13 Abs. 2 NJVollzG)

→ Im Vollstreckungsrecht: Bewährungsprognose bei Strafrestaussetzung (§ 57 Abs. 1 S. 1 Nr. 2 StGB, auch i.V.m. § 57 Abs. 2 StGB bzw. § 57a Abs. 1 S. 1 Nr. 3 StGB; § 88 Abs. 1 JGG)

2. Gefährlichkeitsprognose

→ Unterbringung in der Sozialtherapie (z.B. § 18 Abs. 2 StVollzG Bln, § 17 Abs. 2 StVollzG M-V, § 104 Abs. 1 NJVollzG)

3. Wirkprognose / Interventionsprognose

→ Vollzugsplanung (z.B. § 9 StVollzG Bln, § 8 Abs. 1 StVollzG M-V, § 9 NJVollzG)

→ Unterbringung in der Sozialtherapie (z.B. § 18 StVollzG Bln, § 17 StVollzG M-V, § 104 NJVollzG)

4. Einsatzgebiete künstlicher Intelligenz im Justizvollzug

- Broschüre „Artificial Intelligence Applications in Corrections“ des Criminal Justice Testing and Evaluation Consortiums beim National Institute of Justice des U.S. Department of Justice nennt bereits 2020 vor allem die folgenden Anwendungsbereiche, in denen KI im Justizvollzug weltweit aktuell entwickelt, getestet oder implementiert wird:

1. Überwachung der Kommunikation der Gefangenen
2. Überwachung von Positions- und Gesundheitsdaten
3. Entdeckung eingeschmuggelter Gegenstände
4. Vereinfachung der Verwaltungsabläufe
5. Bewertung des Rückfallrisikos
6. Bewährungsaufsicht durch Chatbots

4. Einsatzgebiete künstlicher Intelligenz im Justizvollzug

- Für Deutschland und Österreich finden sich Hinweise auf die Erprobung von Systemen zur automatisierten Identifikation von Gewalthandlungen gegen andere Personen und Suizidversuchen.

- So wurde von 2019 bis 2021 das Forschungsprojekt „Ereignisgesteuerte Videoüberwachung mit automatisierter Situationseinschätzung als Instrument der Suizidverhinderung in Justizvollzugsanstalten“ durchgeführt und anhand nachgestellter Prüfszenarien in der JVA Düsseldorf getestet.

→ Fazit: prinzipielle Eignung, aber noch nicht praxisreif.

- Im Jahr 2022 erscheinen zudem mehrere journalistische Beiträge, die die Entwicklung eines Systems zur Erkennung von Gewalt- und Suizidhandlungen im niedersächsischen Vollzug ankündigen; getestet werde dieses in der JVA Oldenburg.

→ Schaffung gesetzlicher Grundlagen in §§ 79a, 81, 81a NJVollzG n.F.

→ Danach keinerlei Updates mehr, insbesondere auch kein Projektbericht!

- Weiter fortgeschritten in Österreich (vgl. Rothmann/Mayer MschrKrim 2024, 267 ff. zum Projekt „Künstliche Intelligenz im Strafvollzug“ (KIIS)

→ Aber auch dort keine klaren Evaluationsergebnisse, schwerpunktmäßig rechtlich-ethischer Beitrag!

→ Jeweils geht es nicht um Prognose!

5. Treffsicherheit algorithmenbasierter automatisierter Verfahren

- „Automatisierte“ Verfahren können unmittelbar durch nicht psychowissenschaftlich ausgebildete Entscheider*innen computergestützt angewendet werden und errechnen einen Rückfall- (oder Flucht-)Risikowert für die zu beurteilende Person algorithmenbasiert.

- KI-gestützte Verfahren sind dabei solche, die durch Algorithmen menschliches Entscheidungsverhalten imitieren bzw. intelligentes automatisiertes Verhalten ermöglichen.

  → Nicht immer trennscharfe Unterscheidung von anderen automatisierten, statistisch-aktuarischen Verfahren möglich!

- Geht es allein um die Treffsicherheit der Prognose, ist die Überlegenheit statistisch-aktuarischer Instrumente gegenüber der klinisch-idiographischen Methode in zahlreichen Studien belegt.

  → Aber andere Probleme, z.B. rechtlicher Natur, u.a., wegen fehlender Individualisierung (dazu auch später noch)!

- Empirische Validierung kombinierter Verfahren noch nicht besonders weit vorangeschritten.

  → Zudem grds. Problem des „Clinical Override“: Verschlechterung der Prognose durch klinisch-individuelle Korrektur aktuarischer Prognoseinstrumente!

5. Treffsicherheit algorithmenbasierter automatisierter Verfahren

- Wohl bekanntestes automatisiertes Prognosetool ist weiterhin COMPAS (Correctional Offender Management Profiling for Alternative Sanctions).

  → Nicht KI-basiert, der (allerdings nicht offengelegte) Algorithmus basiert auf dem RNR-Ansatz von Andrews/Bonta.

- Zentrale Prognosetools sind die General Recidivism Risk Scale (GRRS) und die Violent Recidivism Risk Scale (VRRS), die jeweils auf 26 Risikoitems basieren.

- Dressel/Farid (2018) gelangten zu dem Ergebnis, dass das Programm keine besseren Prognosen anstelle als Personen ohne oder mit nur wenig strafrechtlich-kriminologischen Vorkenntnissen.

- Neuere Untersuchungen (Lin et al. 2020) belegen hingegen durchaus eine generelle Überlegenheit von COMPAS gegenüber intuitiven Prognosen.

- Eine Übersicht über AUC-Werte (korrekte Einstufung des Rückfallrisikos) aus verschiedenen Evaluationsstudien zu COMPAS und für verschiedene Rückfalldefinitionen findet sich bei Jackson/Mendoza (2020, S. 6 f.); danach war die Treffsicherheit moderat und lag zwischen 0,64 und 0,74 (Zufall = 0,5).

5. Treffsicherheit algorithmenbasierter automatisierter Verfahren

- Doch auch tatsächlich KI-gestützte, auf machine learning basierende Risikoprognosetools sind bereits im Einsatz bzw. in Entwicklung.

- Ghasemi et al. (2021) verglichen auf drei verschiedenen Formen des machine learning basierende Prognosemodelle mit dem statistisch-aktuarischen Tool des LS/CMI (Level of Service/Case Management Inventory).

- Vergleichbare AUC-Werte im Bereich von bis zu 0,75 wurden erreicht, KI-basierte Modelle erhöhten die Treffsicherheit aber gerade bei den Personen mit mittlerem Rückfallrisiko, bei denen der LS/CMI nicht besser als eine Zufallseinstufung operiert, auf etwa 0,6.

- Es gibt weitere Untersuchungen mit ähnlichen Ergebnissen.

- Allerdings unklar, wie gut die Treffsicherheit in vollzuglichen Kontexten jenseits der Bewährungsprognose (also z.B. bei der Eignungsbeurteilung für Lockerungen oder offenen Vollzug) ist.

→ Keine explizit darauf zugeschnittene Untersuchungen!

- Besondere Probleme resultieren hier aus der geringen Missbrauchsrate und damit Basisrate der Prognose, die das Risiko von Fehlprognosen der falsch-positiven Art (irrtümliche Annahme eines tatsächlich nicht gegebenen hohen Missbrauchsrisikos).

→ Personenbezogene Missbrauchsquote für Deutschland beim Langzeitausgang 2021 1,3 %, beim Ausgang 1,2 %, Freigang 0,6 %, vgl. Dünkel et al. (2024).

- Grds. dürften die Ergebnisse aber übertragbar sein, Entwicklung und Training eines passenden Tools vorausgesetzt!

6. Rechtliche Grenzen

Grund- und Menschenrechte

- Mögliche Verstöße:

a) Fehlende Verhältnismäßigkeit von Grundrechtseingriffen bei mangelnder Individualisierung der Prognoseentscheidung

- BVerfG verlangt eine „umfassende Prüfung der Täterpersönlichkeit und der begangenen Taten“, die auf eine breite Tatsachengrundlage gestützt werden müsse (BVerfGE 109, 190, 241).

- Eine bloß abstrakte, auf statistische Wahrscheinlichkeiten gestützte Prognoseentscheidung sei unzulässig; es bedürfe vielmehr „unter Ausschöpfung der Prognosemöglichkeiten einer positiven Entscheidung über die Gefährlichkeit des Betroffenen, um die Freiheitsentziehung zu rechtfertigen“ (BVerfGE 109, 190, 242).

- Auch nach Auffassung des BGH verbietet sich „eine abstrakte, auf statistische Wahrscheinlichkeiten gestützte Prognoseentscheidung […]. Auch wenn bestimmte Persönlichkeitsstörungen von vornherein ein hohes Rückfallrisiko beinhalten, entbindet dies […] nicht von einer individuellen Gefährlichkeitsprognose“ (BGHSt 50, 121, 130 f.).

- Statistisch-aktuarische Prognoseinstrumente (konkret: Static 99 im Rahmen der für die Sicherungsverwahrung erforderlichen Gefährlichkeitsprognose) könnten für die Prognose zwar Anhaltspunkte über die Ausprägung eines strukturellen Grundrisikos liefern, sie seien indes nicht in der Lage, eine fundierte Einzelbetrachtung zu ersetzen (vgl. BGH bei Pfister 2010, 165).

6. Rechtliche Grenzen

→ Entscheidungen betreffen allerdings jeweils die besonders eingriffsintensiven stationären Maßregeln, insbesondere die Sicherungsverwahrung; ggf. geringere Anforderungen bei geringeren Grundrechtseingriffen!

→ Dennoch ist die fehlende Individualisierung problematisch und die verfassungsrechtliche Zulässigkeit rein automatisierter Prognosen fraglich.

→ „Clinical Override“ könnte das Problem zwar lösen, verschlechtert aber die Prognosequalität (s.o.)!

- Allerdings sind auch menschliche Prognoseentscheidungen nicht immer transparent!

b) Menschenwürdeverstoß (Art. 1 Abs. 1 GG) bzw. „inhuman or degrading treatment“ (Art. 3 EMRK) bei Intransparenz des Algorithmus

- Eine solche Intransparenz könnte ggf. eine Objektivierung / Instrumentalisierung beinhalten, wenn die algorithmenbasierten Entscheidungen menschlich nicht mehr nachvollzogen und erklärt werden können.

→ KI als Black Box, insbesondere bei machine learning auf der Basis neuronaler Netze (deep learning)

- Allerdings existieren bereits Ansätze zur Entwicklung sog. „explainable AI“ („erklärbarer KI“).

→ Auch schon im Bereich der Rückfallprognose!

6. Rechtliche Grenzen

c) Verstöße gegen die richterliche Unabhängigkeit (Art. 97 GG) und die Verfahrensfairness (Art. 1 Abs. 1, 2 Abs. 1, 20 Abs. 3 GG und Art. 6 EMRK)

- Auch diesbezüglich stellen sich verschiedene Probleme, die für rein vollzugliche Prognosen nicht relevant sind, sondern erst bei Ersetzung oder unmittelbarer Unterstützung einer Entscheidung des (Vollstreckungs-)Gerichts.

→ Daher hier nicht weiter vertieft!

d) Diskriminierung (Art. 3 Abs. 3 GG und Art. 14 EMRK)

- Die Verwendung KI-gestützter Prognosetools kann Vorurteile unter bestimmten Umständen, nämlich dann, wenn die Trainingsdaten selbst in dieser Form (und sei es nur unterschwellig) „biased“ sind, noch verfestigen.

→ Externes ethisches Korrektiv notwendig!

- Zudem Gefahr zu starker Gewichtung statischer Faktoren, z.B. Vorstrafen.

- Insbesondere zu COMPAS gibt es eine intensive Debatte dazu, inwiefern dieses Tool biased ist und Vorurteile gegenüber Menschen afroamerikanischer Herkunft fortschreibt und in ungünstigere Prognosen für diese Gruppe ummünzt (bejahend Angwin et al. 2016, ablehnend z.B. Rudin et al. 2020, Überblick bei Dressel/Farid 2018).

- Andererseits kann KI auch gezielt genutzt werden, um im menschlichen Entschiedungsverhalten ebenfalls häufig enthaltenden Bias zu reduzieren.

→ Problembewusstsein notwendig!

6. Rechtliche Grenzen

e) Allgemeines Persönlichkeitsrecht und Recht auf informationelle Selbstbestimmung (Art. 1 Abs. 1 i.V.m. Art. 2 Abs. 1 GG) bzw. Recht auf Privatsphäre (Art. 8 EMRK)

- Probleme entstehen hier insbesondere aufgrund der Art (insbesondere bei Daten sensitiver Natur) und Menge („Big Data“) der gesammelten Daten.

- Dies gilt erst recht, wenn die Art und Weise, nach der die gesammelten Daten verwertet und in Prognoseergebnisse umgesetzt werden, intransparent bleibt.

→ Auch insofern Transparenz essentiell!

- Angemessene Abwägung zwischen Persönlichkeitsrechten und den mit den mittels der Prognosetools verfolgten Zielen notwendig!

Weitere rechtliche Grundlagen

- Zu denken ist an EU-Recht, z.B. die DSGVO und den neuen EU AI Act.

- Zudem dürften klare Rechtsgrundlagen in den Vollzugsgesetzen notwendig sein, zumindest, wenn die eingesetzten Prognosetools menschliche Entscheidungen ersetzen bzw. zumindest determinieren.

- Der bloße Einsatz als Entscheidungshilfe dürfte weniger problematisch sein.

- Allerdings bedarf auch die Sammlung der nötigen Trainingsdaten etc. einer Rechtsgrundlage.

7. Praktische Probleme

a) Automation Bias

- Studien zu Entscheidungsunterstützungssystemen im medizinischen Bereich sowie bei Pilot*innen zeigen, dass es Personen schwerfällt, sich gegen das Ergebnis algorithmischer Berechnungen zu entscheiden, und zwar selbst dann, wenn das Ergebnis des Algorithmus eigentlich nur einen von mehreren Entscheidungsfaktoren darstellen sollte.

- Automation Bias führt dazu, dass Menschen es unterlassen, zusätzlich zu einem algorithmischen Ergebnis selbstständig Informationen einzuholen und zu bewerten, und sogar deutlich gegen das Ergebnis des Algorithmus sprechende Anhaltspunkte bewusst ignorieren.

→ Algorithmen, die lediglich als Entscheidungsunterstützung gedacht waren, determinieren dann faktisch die Entscheidung des Menschen determinieren.

→ Der häufig als nötig erachtete „Human in the loop“ macht sich also selbst überflüssig.

- Weitere Komplikation: Gegenläufige Forderungen zur Bekämpfung von Automation Bias einerseits und Clinical Override andererseits:

→ Macht der „Human in the loop“ die Entscheidung besser, macht er sie schlechter oder kommt es auf die Rahmenbedingungen an?

→ Dazu ist weitere Forschung nötig!

7. Praktische Probleme

b) Automation Aversion

- Schließlich ist die Akzeptanz für automatisierte Entscheidungen in der Bevölkerung zu berücksichtigen.

- Hier zeigt sich eine generelle Aversion gegenüber solchen Entscheidungen gerade dann, wenn es um Entscheidungen mit großer Tragweite geht.

- Nach Fischer/Petersen (2018) ist die „Beurteilung des Risikos, ob ein Straftäter rückfällig wird“ diejenige, bei der mit Abstand am wenigsten dem Computer getraut wurde: nur 2 % stimmten für dessen alleinige Entscheidung, 37 % stimmten der Option eines bloßen Computer-Vorschlags zu und 54 % sagten, es solle allein ein Mensch entscheiden.

- Allerdings könnte das Vertrauen in KI durch mehr Information über ihre Wirkungsweise sowie durch eine zunehmende Integrierung in den Alltag und den damit einhergehenden Gewöhnungseffekt möglicherweise steigen.

8. Fazit

Kann, soll und darf menschliches Verhalten bei Prognoseentscheidungen im Justizvollzug mit Hilfe künstlicher Intelligenz vorhergesagt werden?

Prof. Dr. Stefan Harrendorf
Universität Greifswald
Rechts- und Staatswissenschaftliche Fakultät
Lehrstuhl für Kriminologie, Strafrecht, Strafprozessrecht
und vergleichende Strafrechtswissenschaften

UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456

Vielen Dank!

Kontakt: stefan.harrendorf@uni-greifswald.de

Näher zum Thema: *Butz/Christoph/Sommerer/Harrendorf/Kaspar/Höffler* (2021). Automatisierte Risikoprognosen im Kontext von Bewährungsentscheidungen. In: Bewährungshilfe, S. 241 – 259.