



**Imperial College
London**

Trusted AI in Europa

Michael Huth

*Erstes Treffen des deutschen Chapters von
CLAIRE*

London, 28. September 2021



Trusted AI: Teil unserer Welt und Verantwortung

Kate Crawford, Microsoft Research,
Interview vom 6.6.21, The Observer

- KI ist weder künstlich noch intelligent
- Natürliche Ressourcen und menschliche Arbeit treiben maschinelles Lernen an ...
- ... und bringen regressive Stereotypen in die Algorithmen
- "Hey, Alexa, order me some toilet rolls"

Photo: The Observer



Trusted AI: European Values/Precision Inside?



- Intel hatte lange und erfolgreiche Markenkampagne
- EU Academia und Industrie können von ähnlicher Markenarbeit zu Trusted AI profitieren.
- Zwei Aspekte scheinen in diesem Zusammenhang wichtig zu sein:
 - "European Values Inside" (Privatsphäre, Recht auf eine Zukunft, ...)
 - "European Precision Inside" (ähnlich zum "German precision engineering")



Trusted AI in der EU: Geopolitische Betrachtungen

-
- Divergenz in Regulierung von Anwendungen, zB Gesichtserkennung im öffentlichen Raum
 - Divergenz im Datenschutz
 - Konvergenz im Anwenden von Methoden die AI robust/resilient machen
 - Divergenz in ethischer Verankerung von KI

Welche Menschen vertrauen auf welche Weise KI?

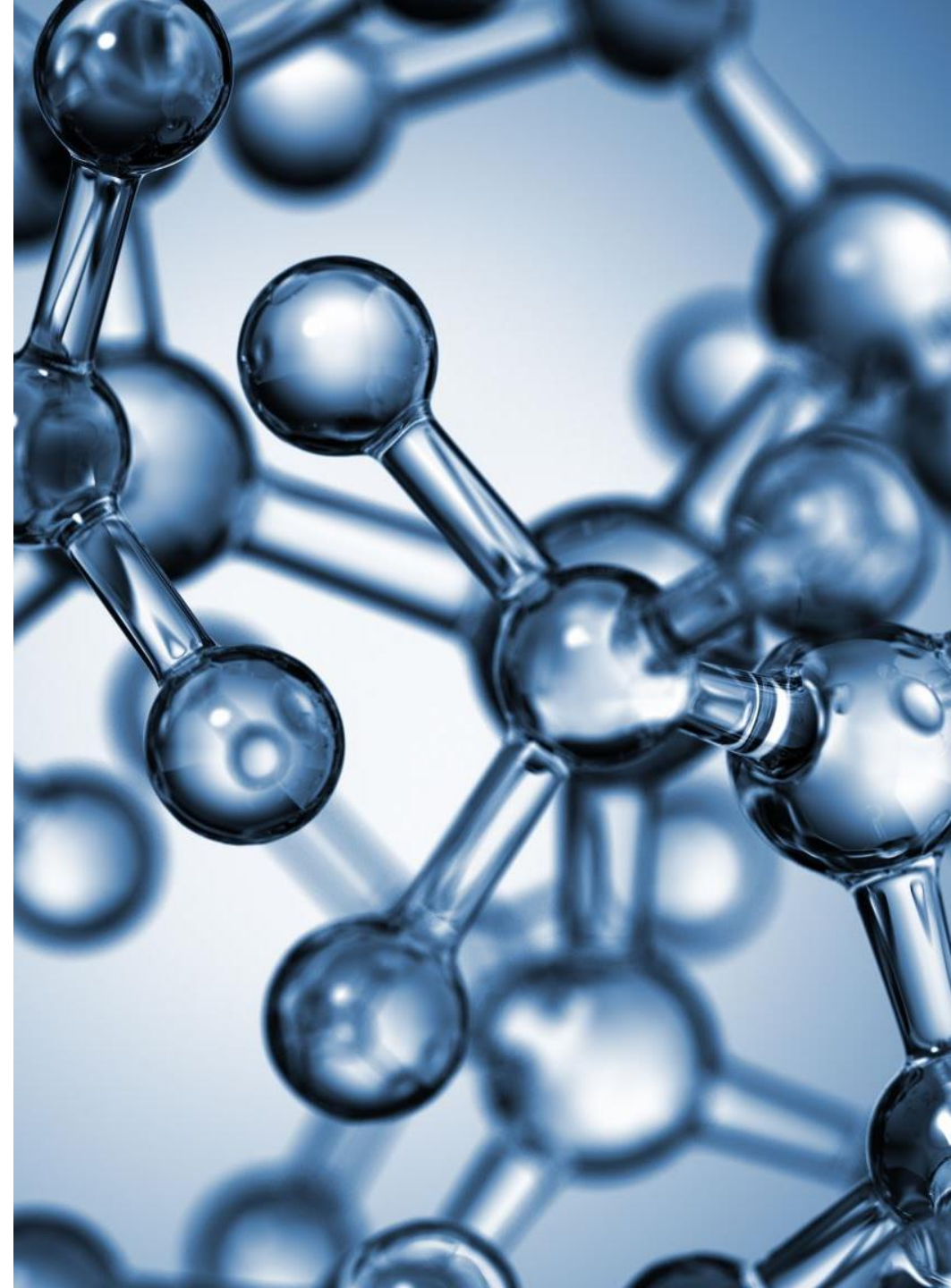
Konsument*innen von Produkten.

Ingenieur*innen bei Ihrer Arbeit.

Einkäufer*innen in Ihren Lieferketten.

Alle, die nicht wissen, dass sie mit KI interagieren.

...



Trusted AI:

*"Bereitstellung von (überprüfbaren)
Garantien für das korrekte Funktionieren
moderner KI-basierter Systeme"*

Setzt voraus dass wir wissen was korrektes und nicht korrektes Funktionieren einer KI im Anwendungsfall sind

Macht es Sinn zu sagen die KI in Douyin (TikTok) ist korrekt oder nicht korrekt?

Es macht Sinn zu sagen, dass ein Neuronales Netz zur Klassifizierung robust unter Eingabeperturbationen ist


Forschung an der Bereitstellung von Methoden die solche überprüfbaren Garantien geben können ist sehr wichtig

Die Verfügbarkeit solcher Methoden ist aber nur ein Baustein für Trusted AI

...

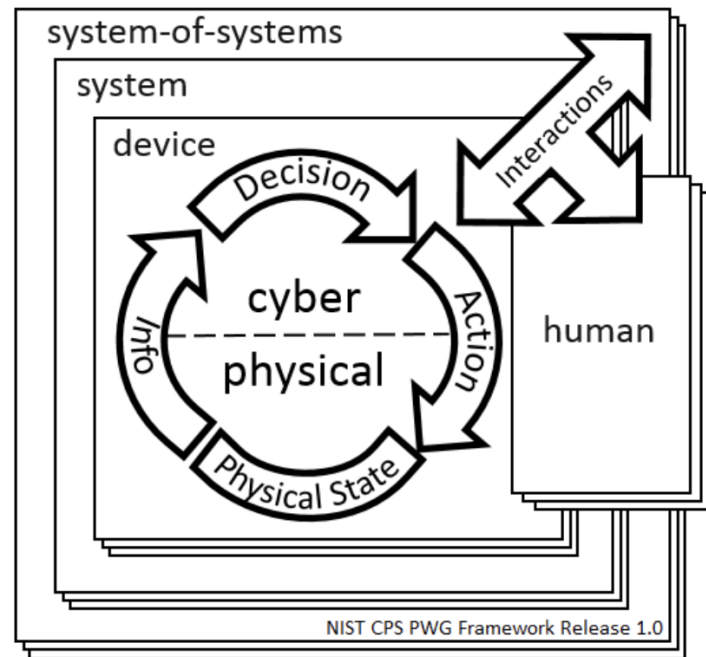


Trusted AI: Eigenschaft eines Gesamtsystems

- Zertifizierung eines neuronalen Netzes bezüglich einer Eigenschaft, zB Robustheit unter kleinen Eingabeperturbationen: *Trusted Neural Net*
 - Was passiert wenn das Modell für Endgeräte komprimiert wird, in Rust re-implementiert wird, Hyperparameter verändert werden, ...?
 - Was ist wenn das zertifizierte Modell irgendwann eine andere Funktion oder Mission in einem System übernimmt?
 - Trusted AI wie Systemsicherheit: *keine kompositionelle Eigenschaft*
 - Was bedeutet das für R&D und Product Owners die Trusted AI in Produkte einbetten?
- 

Bezugssysteme für zuverlässige KI?

Ordnungsrahmen erwünscht für das Planen, Entwerfen, Implementieren und Bewerten von KI
Zuverlässigkeit in der Produktion. Vergleichbare Arbeit in anderen Bereichen, zB von US NIST für *Cyber Physical Systems*:



CPS Conceptual Model

Domains

Manufacturing

Transportation

Energy

Healthcare

... Domain

Aspects

Functional

Business

Human

Trustworthiness

Timing

Data

Boundaries

Composition

Lifecycle

Facets

Conceptualization

Realization

Assurance

Use Case,
Requirements, ...

Design / Produce
/ Test / Operate

Argumentation,
Claims,
Evidence

Activities

Artifacts

Model of a CPS

CPS

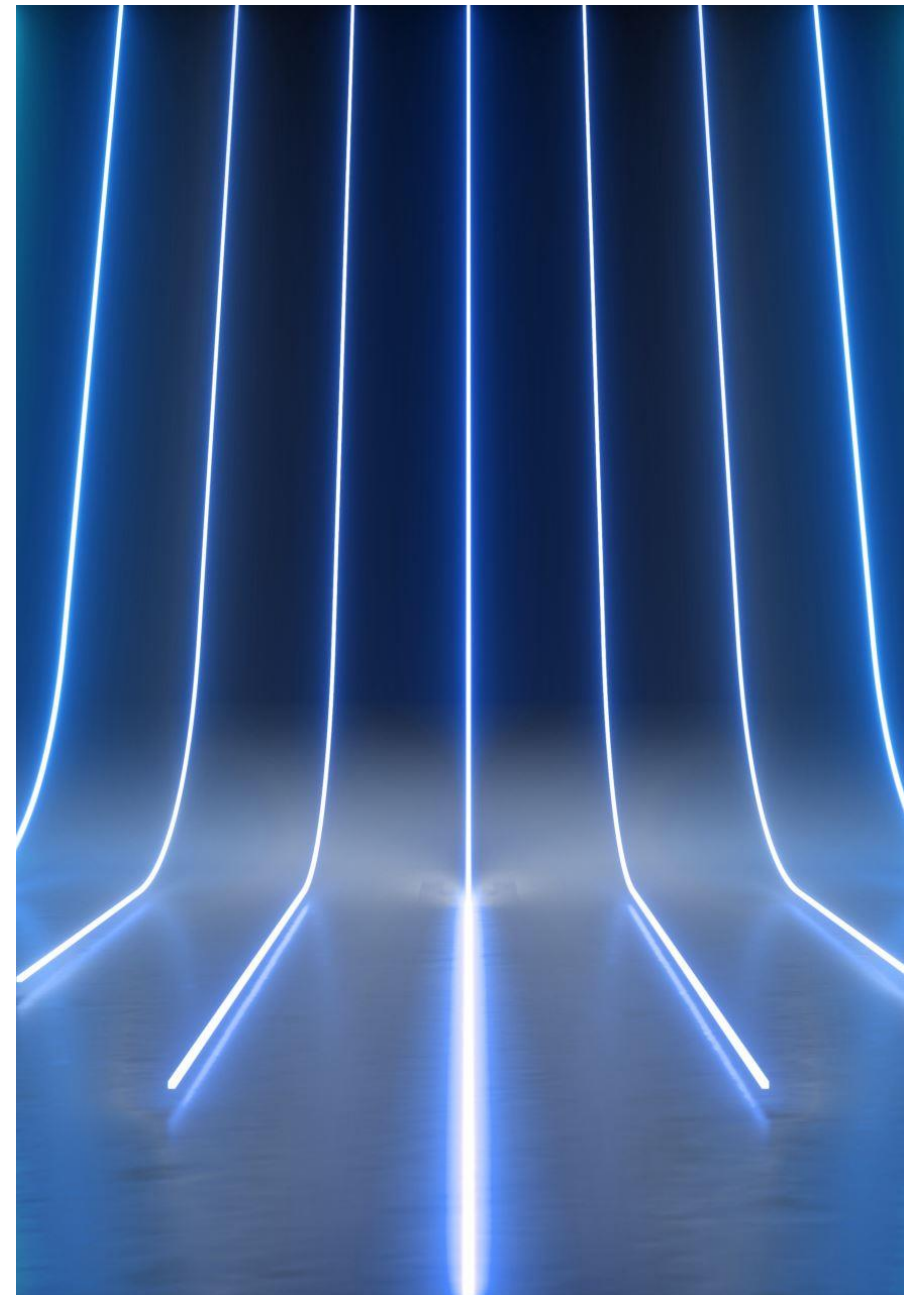
CPS Assurance

NIST CPS PWG Framework Release 1.0

CPS Framework – Domains, Facets, Aspects

Quelle: <https://pages.nist.gov/cpspwg/>

Trusted AI Forschung: Beispiele vom Imperial College London und anderswo



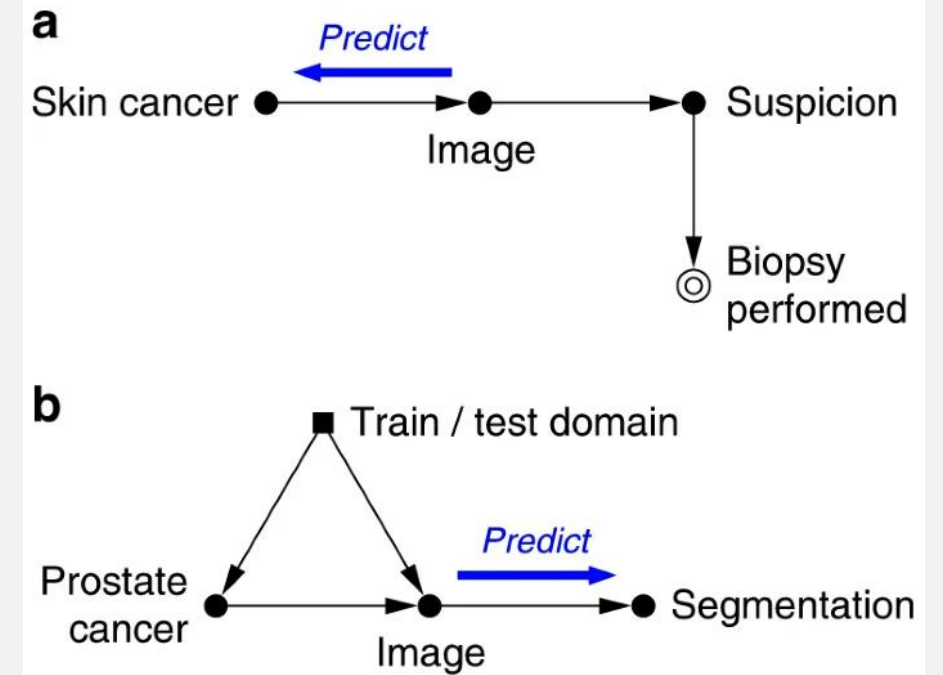
Kausalität zu verstehen ist wichtig in der medizinischen Bildanalyse

Causality matters in medical imaging

[Daniel C. Castro, Ian Walker & Ben Glocker](#)

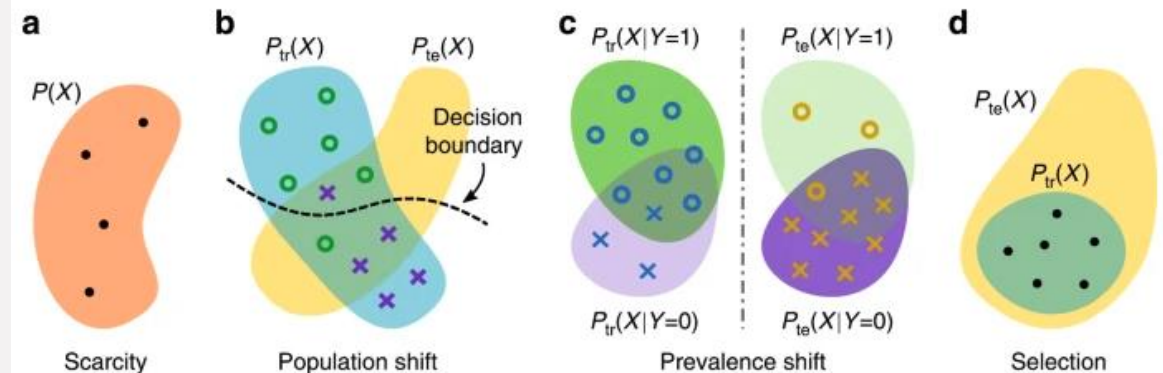
[Nature Communications](#) volume 11

Fig. 2: Causal diagrams for medical imaging examples.



Quelle: zitierte Arbeit

Fig. 1: Key challenges in machine learning for medical imaging.



Formale Verifikation Neuraler Netze: "Safe AI"

Efficient Verification of ReLU-based Neural Networks via Dependency Analysis

Erreichbarkeit von Ausgaben: **Elena Botoeva, Panagiotis Kouvaros, Jan Kronqvist, Alessio Lomuscio, Ruth Misener**

Department of Computing, Imperial College London, UK

{e.botoeva, p.kouvaros, j.kronqvist, a.lomuscio, r.misener}@imperial.ac.uk

formale Analyse kann zeigen
dass unsichere Ausgaben eines
neuralen (feed-forward
ReLU) Netzwerkes nicht
erreichbar sind.

Erfordert Formalisierung
von "unsicher".

Verification problem. Given a network $f: \mathbb{R}^{s_0} \rightarrow \mathbb{R}^{s_k}$
and a specification $(\mathcal{X}_0, \mathcal{X}_k) \subseteq \mathbb{R}^{s_0} \times \mathbb{R}^{s_k}$, the verification
problem determines whether $\forall \mathbf{x}_0 \in \mathcal{X}_0: \mathbf{x}_k \in \mathcal{X}_k$.

Perturbierte Eingaben



Feed Forward Network

Unsichere
Ausgaben

Abhängigkeitsanalyse für
bessere Skalierbarkeit der
Verifikation.

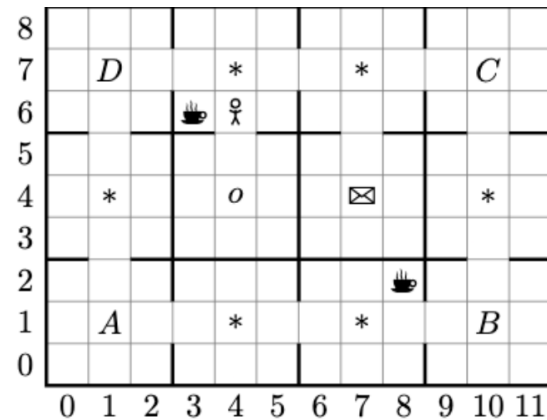
Verifikation mit "Cuts" und
"Splits" aus dem Mixed Integer
Linear Programming.

Interagierende symbolische und nicht symbolische KI

Daniel Furelos-Blanco, Mark Law, Anders Jonsson, Krysia Broda, Alessandra Russo:

Induction and Exploitation of Subgoal Automata for Reinforcement Learning. J. Artif. Intell.

Res. 70: 1031-1116 (2021)



Episodisches bestärkendes Lernen: Induktives Logic

Programming lernt einen Automaten der die Eventsequenzen eines Agenten des bestärkenden Lernens erfasst.

Wenn eine neue Eventsequenz nicht vom Automaten generiert werden kann, wird das induktive Lernen den Automaten angepasst um auch diese Sequenz zu beinhalten.

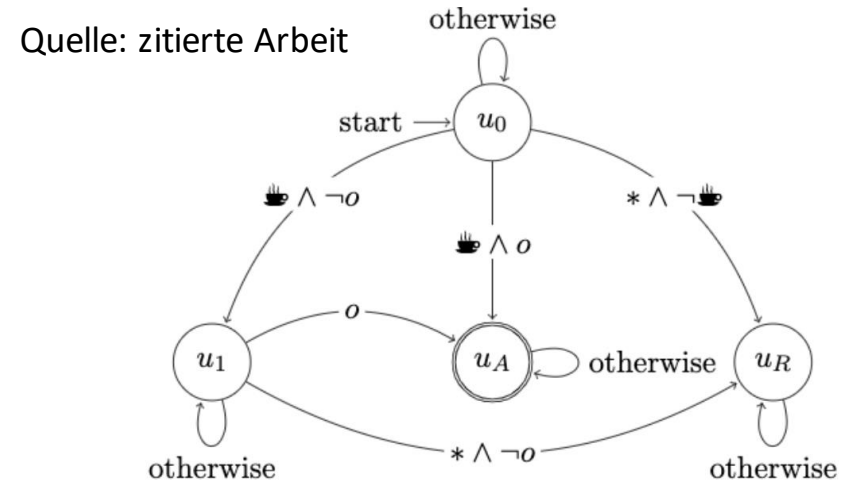


Figure 2: Subgoal automaton for the OFFICEWORLD's COFFEE task.

Dies generiert minimale aber vollständige Automaten, die somit effizienter sind und das bestärkende Lernen besser erklären.

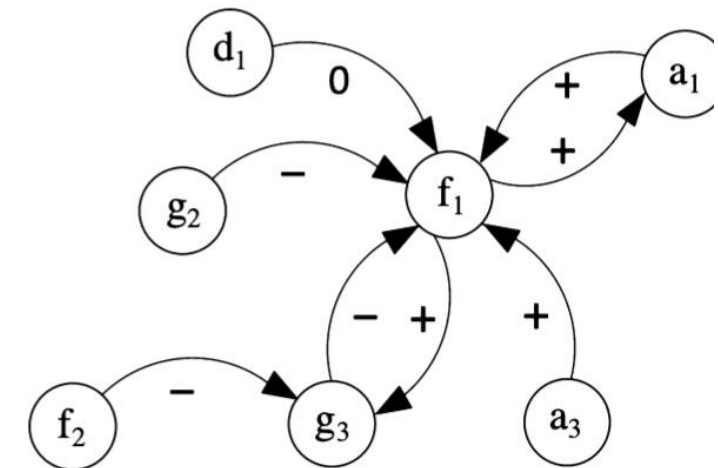
Argumentation Frameworks

- Graphen in denen Ecken Argumente darstellen
- Ecken unterstützen sich oder greifen einander an.
- Wir wollen Mengen von Ecken deren interne Unterstützung konsistent ist und deren externe Angriffe von innen gekontert werden.
- Von solchen Strukturen können auch quantitative Größen abgeleitet werden, zB dialektische Stärke eines Argumentes.
- Nun auch angewendet um über die Struktur eines neuronalen Netzwerkes zu argumentieren.

Example variations in explanation content for f_1 , with argumentative artefacts in the linguistic explanations highlighted in bold. Note that many other variations are possible.

Requirements	Content	Linguistic Explanation
All supporters of f_1	a_1, a_3	<i>Catch Me If You Can was recommended because you like Leonardo DiCaprio and Tom Hanks.</i>
Strongest attacker and strongest supporter of f_1	a_1, d_1	<i>Catch Me If You Can was recommended because you like Leonardo DiCaprio, despite the fact that you dislike Biographies.</i>
A weak attacker of f_1 and its own attacker	g_3, f_2	<i>Catch Me If You Can was not recommended because it inferred that you don't like Dramas, since you disliked Moulin Rouge.</i>

Quelle: zitierte Arbeit



[Antonio Rago](#), [Oana Cocarascu](#), [Christos Bechlivanidis](#), [David A. Lagnado](#), Francesca Toni:
Argumentative explanations for interactive recommendations. [Artif. Intell. 296](#): 103506 (2021)

Standardisierungen sind wichtig

Fallstudie:
Komprimierung
neuraler Netzwerke

Solche Techniken
erschliessen neue
Anwendungsbereiche,
z.B. KI auf Smartphones.

Standardisierungen sind
da wichtig.

Context-Based Adaptive Binary Arithmetic Coding in the H.264/AVC Video Compression Standard

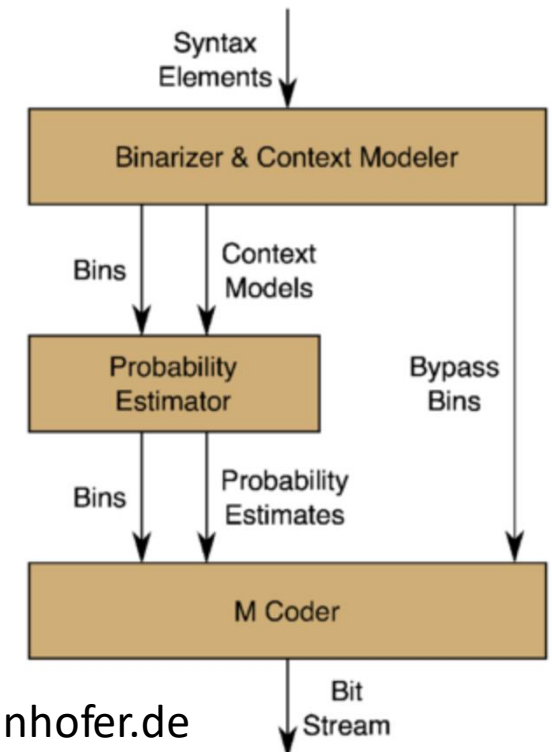
Detlev Marpe, *Member, IEEE*, Heiko Schwarz, and Thomas Wiegand

Das Fraunhofer Institut hat angekündigt dass die obige Komprimierungstechnik nun Teil eines solchen Standards ist, der sich leicht in andere Standards integrieren lässt.

<https://newsletter.fraunhofer.de/-viewonline2/17386/587/5/6RFhct0v/KdFwGRLkqA/1>

Das Verfahren wird nun erweitert für
inkrementelle Komprimierung, z.B. für
Föderiertes Lernen

Quelle: hhi.fraunhofer.de





Trusted AI: Probleme in der Umsetzung

- Evaluierung der Zuverlässigkeit innerhalb existierender Standards
- Produktreife akademischer Lösungen, z.B. private Inferenz
- Verfügbarkeit zuverlässiger Datenmengen
- Erhalten der Zuverlässigkeit im Produktzyklus
- Methoden der Prüfungssicherheit müssen in technischen Produktumgebungen anwendbar sein
- Integration oder Adaption akademischer Tool für professionelles Tooling

Vielen Dank!

Fragen?



**Imperial College
London**