KI-FACHKONFERENZ

Artificial Intelligence Act

Dr. Oliver Maspfuhl

Frankfurt am Main,
22.11.2021

I. Wie normierbar ist der KI-Begriff?

Quelle: ...

Was sind typische Aspekte von KI?

EU proposal für KI regulation

‘artificial intelligence system’ (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I* and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with;

*(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;
(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
(c) Statistical approaches, Bayesian estimation, search and optimization methods.

„Intelligence“ means only the goal is set explicitly a priori, not the way to achieve it

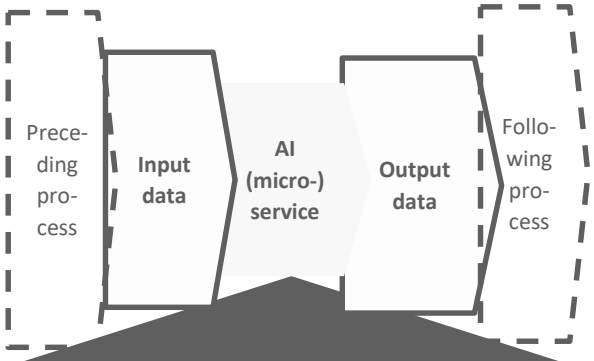
Nach der klassischen Definition ist KI…

Menschlich denken	**Rational denken**
Menschlich handeln	**Rational handeln**

- Rationales Verhalten hat höchste Relevanz in der Finanzwelt
- KI kann Menschen in **Leistung und Effizienz** übertreffen

Input-output mapping im Maschinelen Lernen und physikalischen Modellen

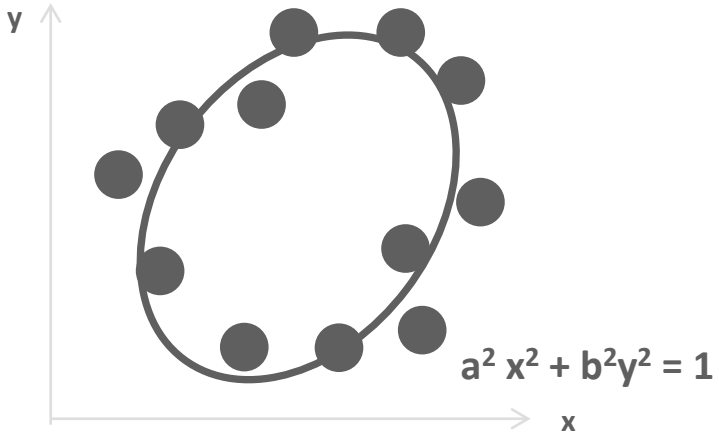
ML-Modell



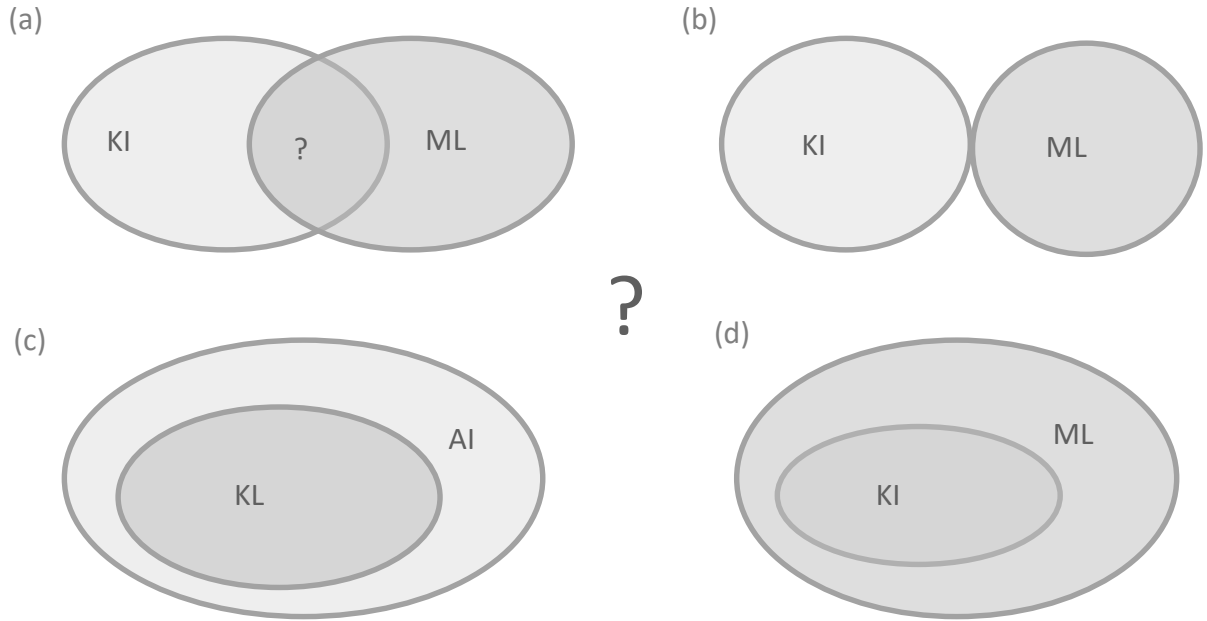
- Mathematische Repräsentation als Abbildung f zwischen rellen Vektorräumen $y = f(x)$ oder allgemeiner $f(x, y) = 0$
- Abbildung wird “gelern” durch fitten mathematischer Modelle mit freien Parametern an Beispielwerte für x und y (**ML**)
- Moderne ML- Lernalgorithmen können extrem nichtlineare Beziehungen abbilden, die trotz vieler Parameter **robust auf neue Daten generalisieren**

Physikalisches Modell

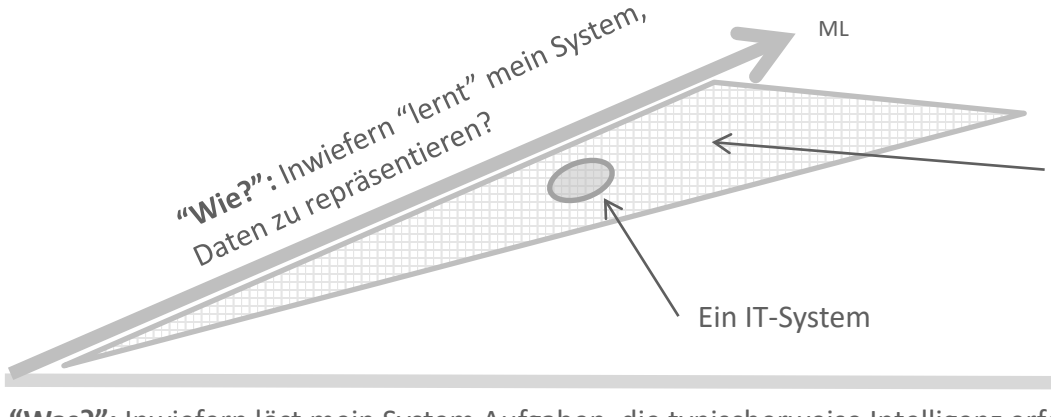
- (Nichtlinerare) Struktur aus fundamentaler **Theorie** abgeleitet
- **Wenige** fundamentale Parameter



Quiz: Welche Relation besteht zwischen KI und ML?



KI und ML sind unabhängige Aspekte

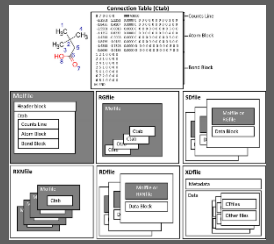


- AI und ML sind **orthogonale Konzepte**, keine Teilmengen voneinander
- Effiziente KI beruht heute fast immer auf ML

"Was?": Inwiefern löst mein System Aufgaben, die typischerweise Intelligenz erfordern, also nicht durch eine explizite Vorschrift beschrieben werden können?

ML ohne KI?

- Die Berechnung der physikalischen oder chemischen Eigenschaften komplexer Moleküle verlangen das numerische Lösen eine partiellen Differentialgleichung
- Diese Aufgabe kann durch einen Optimierungsalgorithmus eindeutig beschrieben und beliebig exakt gelöst werden – **keine menschliche Intelligenzleistung**



https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00460-5

Diskussionspunkt 1 – Brauchen wir eine Definition von KI, um KI zu regulieren?

A. Ja, sonst wissen wir nicht worüber wir sprechen

B. Nein, wir sollten nur auf das Ergebnis schauen und Leistungen unabhängig von ihrer Realisierung bewerten

C. Ich habe noch keine Meinung dazu

II. Wie normierbar ist Vertrauenswürdigkeit?

Quelle: ...

Ausgangspunkt einer Policy: Die Sieben Anforderungen der HLEG

1. **Vorrang menschlichen Handelns und menschliche Aufsicht***: z. B. Grundrechte, Vorrang menschlichen Handelns und menschliche Aufsicht*

2. **Technische Robustheit und Sicherheit:** *z. B. Widerstandsfähigkeit gegen Angriffe und Sicherheitsverletzungen, Auffangplan und allgemeine Sicherheit, Präzision, Zuverlässigkeit und Reproduzierbarkeit*

3. **Schutz der Privatsphäre und Datenqualitätsmanagement:** *z. B. Achtung der Privatsphäre, Qualität und Integrität der Daten sowie Datenzugriff*

4. **Transparenz:** *z. B. Nachverfolgbarkeit, Erklärbarkeit und Kommunikation*

5. **Vielfalt, Nichtdiskriminierung und Fairness:** *z. B. Vermeidung unfairer Verzerrungen, Zugänglichkeit und universeller Entwurf sowie Beteiligung der Interessenträger*

6. **Gesellschaftliches und ökologisches Wohlergehen:** *z. B. Nachhaltigkeit und Umweltschutz, soziale Auswirkungen, Gesellschaft und Demokratie*

7. **Rechenschaftspflicht:** *z. B. Nachprüfbarkeit, Minimierung und Meldung von negativen Auswirkungen, Kompromisse und Rechtsbehelfe.*

Vertrauenswürdigkeit ist ein menschlicher Begriff und problematisch für KI

Gründe für Vertrauen

Mensch

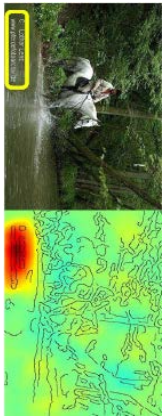
- Gute Erfahrungen aus der Vergangenheit
- Zurückführung auf **gute inhärente (charakterliche) Prinzipien bzw. allgemeine Intelligenz**
- Erwartung, dass Fehler durch Flüchtigkeit passieren können
- **Anpassungsfähigkeit** an besondere Umstände gemäß Prinzipien

(Schwache) KI

- Gute Erfahrung aus historischen Daten
- **Inhärente Prinzipien bzw. allgemeine Intelligenz unklar**
- Erwartung, dass keine Fehler aus Flüchtigkeit passieren, dafür nicht-erklärbare statistische Fehler
- **Keine Anpassungsfähigkeit** an besondere Umstände gemäß Prinzipien

(Schwache) KI ist nicht Denken

- Neurale Netzwerke nutzen Architekturen nach dem Vorbild des Gehirns, aber niemand weiß heute genau wie das Gehirn Informationen verarbeitet!
- Menschen lernen viel schneller und effizienter, offenbar wirken hier höhere **Abstraktionsmechanismen** und **physisches Kontextverständnis**
- Visualisierung von Feature-Importances zeigt auf, dass die Mustererkennung oft auf sehr spezifischen Sub-Pattern oder gar **irrelevanter virtueller Kontextinformation** beruht
- Studien zeigen, dass KI-Systeme durch beliebig kleine Manipulationen in den Eingangsdaten verwirrt werden können
- Wir brauchen Methoden und theoretische Absicherungen gegen solche Angriffsszenarien und **Robustheit gegen kleine Variationen** in den Inputdaten



S.Lapuschkin et al,
Unmasking Clever Hans predictors and assessing what machines really learn (Nature communications)
Layer-wise Relevance Propagation for Deep Neural Network Architectures

Wie kommt man zu Vertrauenswürdigkeit? - Deep Learning als Chance verstehen

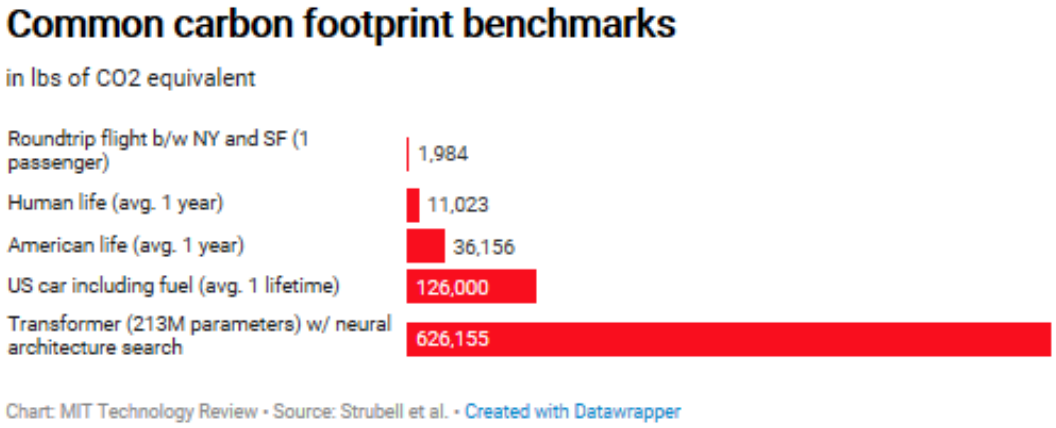
- Vertrauenswürdigkeit von KI muss anders bewertet werden als die von Menschen
- Fokus muss vom „was?“ zum „wie?“ gehen - und damit zur **statistischen Modellvalidierung**
- Dabei sollte man so jedoch so weit wie möglich die menschliche **Verallgemeinerungsfähigkeit** in den Mittelpunkt stellen
- D.h. die Fähigkeit des Modells, **tiefergehende Muster** abzubilden
- Das ermöglicht es auch, als menschlicher Nutzer neue Erkenntnisse zu gewinnen
- Zusätzlich **umfassende Beurteilung** weiterer Aspekte – s. Richtlinien der HLEG

... und warum ist Machine/Deep Learning dabei nützlich?

- Machine/Deep Learning kann den Abstraktionsprozess beim Lernen formal nachvollziehen
  - DNN etc. können nicht-semantische Daten zu bedeutungstragenden Mustern zusammensetzen (**Representation Learning**)
  - Dies kann analysiert und visualisiert werden – aber nur für Probleme, die für Menschen intuitiv lösbar sind
- ML/DL Modelle sind **keine Black Boxes** (die Berechnungen sind vollkommen transparent)
- Aufgrund tiefer Feature-Hierarchien können sie sogar besser geeignet sein, Zusammenhänge zu erklären – *wenn man sie entsprechend designed*

KI-Systeme unterscheiden sich in der Effizienz von menschlicher Intelligenz

…und auch das gehört zur Vertrauenswürdigkeit

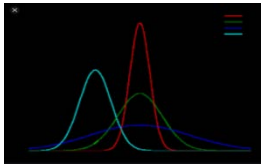


- Energieeffizienz von Lernmethoden ist daher heute ein wichtiges Forschungsgebiet!

Umsetzung einer Policy: Implementierbare Definitionen

1.Nicht-technische Interpretation der Bedrohung

Beispiel:
Verzerrungen in Trainingsdaten



2. Technische Definition einer Messgröße für die Bedrohung

Beispiel:
Statistisches Maß für gleiche bedingte Verteilungen

3. Richtlinie für die Implementierung Vertrauenswürdige Systeme

Beispiel:
Monitoring eines Fairness-Bias-Maßes auf den Modellergebnissen für verschiedende Gruppe

- Dies ist nur möglich durch den teilweisen Perspektivwechsel vom „Was“ zum „Wie“

Diskussionspunkt 2 – Sollte man und kann man eine enge oder weite Normierung der Vertrauenswürdigkeit anstreben?

A. Eng und technisch präzise

B. Umfassend und prozessbasiert

C. Ich habe noch keine Meinung dazu

III. Wie normierbar ist Credit Scoring?

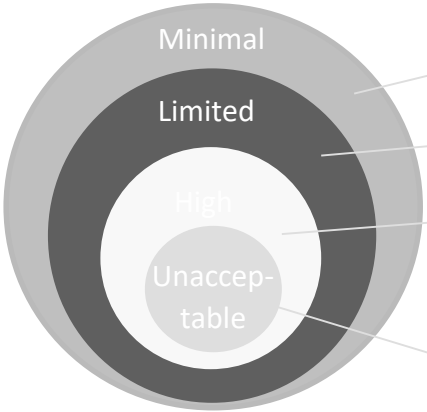
Quelle: ...

Kommende Regulierung antizipieren

Seit 2019 veröffentlichte Vorschläge zur KI-Regulierung

2019 Ethical Guidelines for Trustworthy AI of the High-Level Expert group of the European Commission	➡ **Ethik** (“Was?”)
2020 Whitepaper in AI of the European Commission	➡ **Hochrisikosektoren**
2021 Proposal for a Regulation of the European Parliament and of the Council	➡ **Risikoklassen** (“Wie?)

Risikobasierter Ansatz



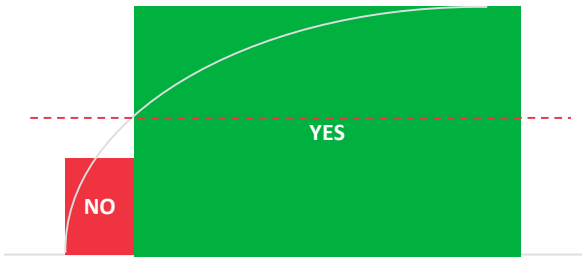
- Keine zusätzlichen Verpflichtungen
- Transparenzvorschriften gemäß Artikel 52
- Anwendungen müssen Anfordungen nach Artikel 6/7 erfüllen, „Conformity“ muss mittels „Conformity Assessment“ nachgewiesen werden
- Definition in Artikel 5, Einsatz nicht erlaubt

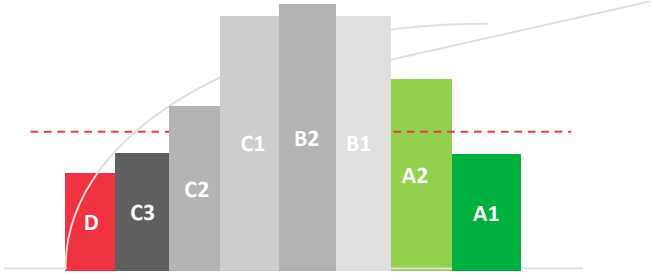
(Hochrisiko-) Sektorbasierter Ansatz

- Healthcare
- Transport
- Energy
- Public sector

2021 Regulierung Act 3: Scoringmodelle sind high-risk?

- Scoringmodelle sind Risikomodelle und sollen die Bank vor Verlusten schützen (nur indirekt auch den Kunden)
- Scoringmodelle ermitteln **Ausfallwahrscheinlichkeiten** für eine Ratingklassifikation, aber **keine binäre** Kreditentscheidung
- Innerhalb einer Ratingklasse zahlen alle Kunden eine Risikprämie (**Versicherungsprinzip**)
- Keine Aussage über individuelles Ausfallverhalten
- Konservativitätspuffer sind nicht nicht “fair” aus Sicht des Kunden, aber **notwendig**
- Beispiel: Kredite warden bei einer Rückzahlwahrscheinlichkeit von 90% in der Regel abgelehnt!

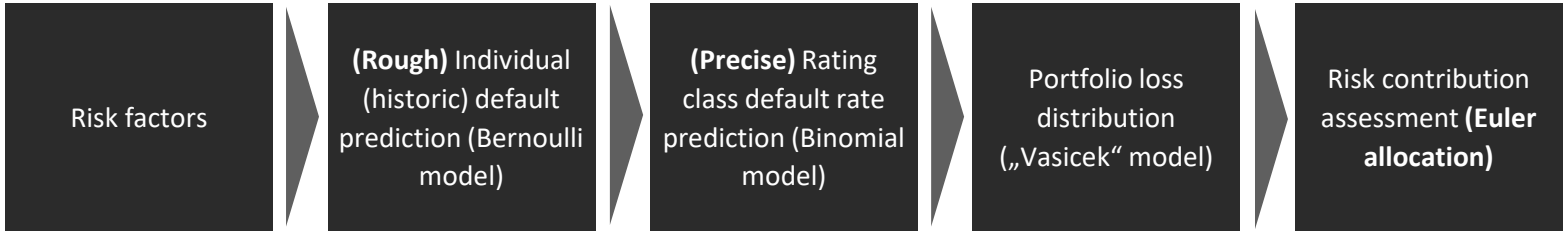




Keine individuelle Prognose für den Kreditausfall in der Klasse

2021 Regulierung Act 3: Scoringmodelle sind high-risk?

Was ist der Risikopreis?



Fairnessaspekte

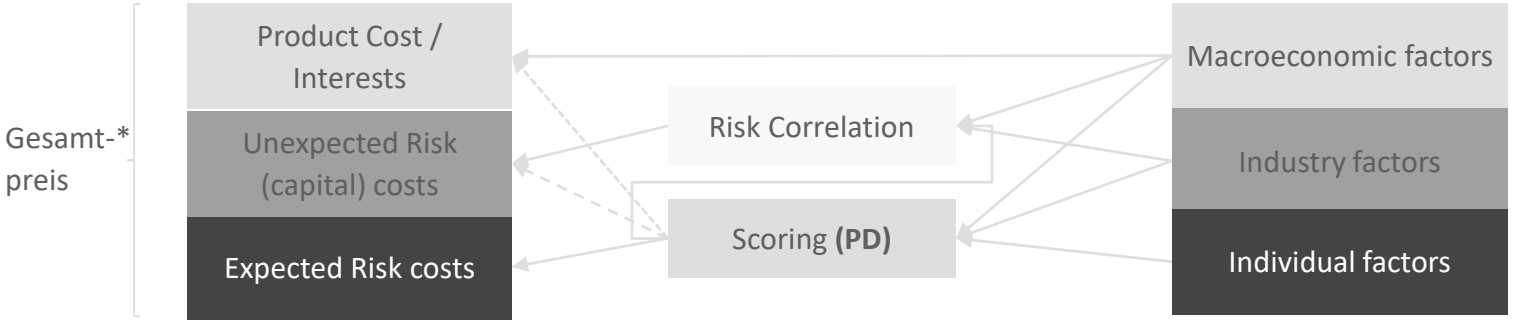
- Zuornung zu einer Klasse kann aus unterschiedlichsten Gründen erfolgen
- Zuordnug zur einem Ratingsystem aus sehr formalin Gründen
- Ground truth ist unbekannt
- Selection Bias
- *“When we leave out attributes, can the price estimation become unfair?” / “We do a risk prediction…”*

Methodische Aspekte

- Stochastische Systeme können nicht vorhergesagt werden
- KI ersetzt Systeme die nicht mit Regeln funktionieren -> Kann man diese mit Regeln nornieren?
- Bei Fehlern: Sind diese auf systemnatische Faktoren zurückzuführen? Failty factors? Hidden variables?

2021 Regulierung Act 3: Scoringmodelle sind high-risk?

- Credit Pricing ist in sich komplex und berücksichtigt viele (auch ML-unabhängige) Faktoren
- Fairnessbetrachtungen nur für das Scoring ist relativ bedeutungslos
- … und alle Beziehungen sind **nicht-linear**



Die klassischen Prüfer von Ratingsystemen sind gut vorbereitet, die richtige Funktionsweise dieser Modelle zu bewerten.

* schematisch

Diskussionspunkt 3 – Kann man Credit Pricing unter die high-risk-Anwendungen zählen? Wer sollte sie prüfen?

A. Ja, Prüfung druch unabhängige Dritte

B. Nein, es sollte weiter die klassische Modellprüfung durch Bafin/Bundesbank stattfinden

C. Ich habe noch keine Meinung dazu

IV. Wie normierbar ist KI-Risiko-Management?

Quelle: ...

High-risk-KI-Systeme - Anforderungen

Establish and implement **risk management system** in light of the **intended purpose** of the AI system

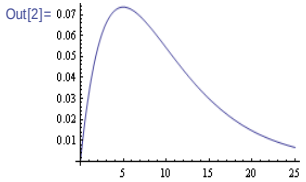
- Use high-quality **training, validation and testing data** (relevant, representative, …)
- Draw up **technical documentation** & setup **logging capabilities** (traceability & auditability)
- Ensure appropriate degree of **transparency** and provide users with **information** on capabilities and limitations of the system & how to use it
- Ensure **human oversight** (measures built into the system and/or implemented by the users)
- Ensure **robustness**, **accuracy** and **cybersecurity**

High-risk-KI Systems – Providing conformity

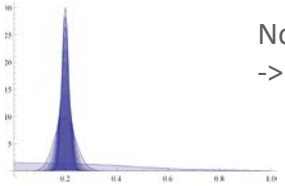
High-risk AI systems shall undergo a new conformity assessment procedure whenever they are substantially modified.
For high-risk AI systems that **continue to learn** after being placed on the market, changes to the high-risk AI system and its performance that have been **pre-determined** by the provider at the moment of the **initial conformity assessment** and are **part of** the information contained in the **technical documentation**, are **not substantial modifications**.

KI for Risk Management: Mustererkennung

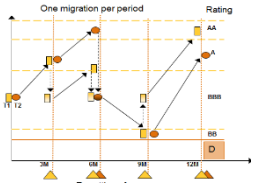
- **Synchrone Abhängigkeiten** (gleichzeitiger/aufeinanderfolgender) Ereignisse zerstören das Versicherungsprinzip -> zentrale Rolle im Risikomanagement
- **Zeitliche Abhängigkeiten** führen über mehrere Periode oder abrupt zu dynamisch veränderten Bedingungen oder Regimewechseln
- Abhängigkeiten oder **Regimewechsel** sind mit klassischen Daten schwer zu messen und vorherzusagen (Dynamik, Psychologie)
- Nötige Informationen zur vorausschauenden Bewertung von Gefahren und Risiken in der Regel
  - noch in keinem Banksystem erfasst, teilweise in externen Quellen
  - nicht zentral verfügbar, sondern verteilt und aus kleinen, schwachen Signalen zusammengesetzt
  - unstrukturiert (Text)
- Anhängigkeiten sind Muster -> versteckt in komplexen Signalen!



Abhängigkeiten -> Schwere Ränder



Normalverteilung -> Diversifikation

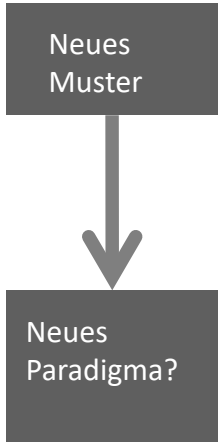


Mehrperiodizität -> Nichtlinearität

KI for Risk Management: Mustererkennung

ML/DL macht Mustererkennung objektiver

- DL vermeidet dir Abhängigkeit von Paradigmen auf der Featureseite – es findet und verifiziert Muster objektiv
- KI „zertifiziert die Expertenmeinung“ über bestehende Paradigmen, der Experte verifiziert neu entdeckte Muster
- Bayes‘sche Methoden können helfen, datengetriebene und expertenbasierte Ansätze zu verbinden

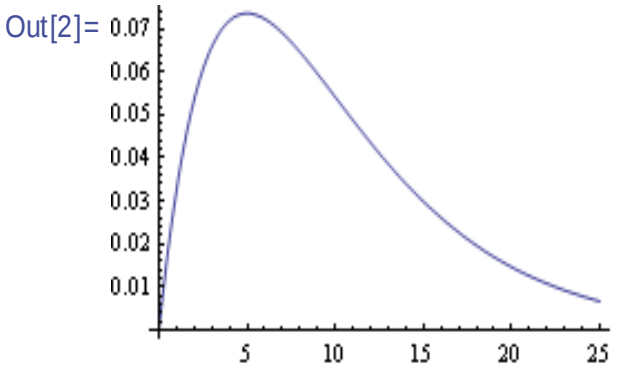


Simulation und Anomalieerkennung sollten Kernanwendungen der KI (im Risikomanagement) sein

- In überwachten Methoden gibt es immer noch eine Paradigmenabhängigkeit über die Zielvariablen
- Nichtüberwachte Methoden finden Anomalien durch Modellierung der Normalität
- KI-Systems können für die Simulation neuer Muster genutzt werden (GANs)
- Die Erkennung neuer Paradigmen ist mit der **Erkennung von Modellgrenzen** verwandt

Risk Management for KI: Modellrisiko 2.0

- Am Ende ist jedoch das Management von KI-Systemen, die auf ML beruhen, eine Aufgabe für das **Modellrisikomanagement**
- Nicht das individuelle statistische Risiko einer Fehlfunktion eines KI-Systems kann gemanaged werden, sondern nur das **Portfoliorisiko**
- Abdeckung des Einzelrisikos abzudecken letztlich nur durch **Versicherungsansatz** möglich: Ersetzung des Schadens im Einzelfall gegen Prämienzahlung



Fat tails… here we go again…

Diskussionspunkt 4 – Sollte Risikomanagement von KI-Anwendungen im Rahmen des klassischen Modellrisikomanagements erfolgen? Kann man diese normieren?

A. Ja, und man sollte das klassische Model-Risk-Framework weiterentwickeln

B. Nein, man braucht völlig neue Ansätze wegen des Echtzeittraining

C. Ich habe noch keine Meinung dazu

Diskussion

Quelle: ...

Diskussionspunkt 1 – Brauchen wir eine Definition von KI, um KI zu regulieren?

A. Ja, sonst wissen wir nicht worüber wir sprechen

B. Nein, wir sollten nur auf das Ergebnis schauen und Leistungen unabhängig von ihrer Realisierung bewerten

C. Ich habe noch keine Meinung dazu

Diskussionspunkt 2 – Sollte man und kann man eine enge oder weite Normierung der Vertrauenswürdigkeit anstreben?

A. Eng und technisch präzise

B. Umfassend und prozessbasiert

C. Ich habe noch keine Meinung dazu

Diskussionspunkt 3 – Kann man Credit Pricing unter die high-risk-Anwendungen zählen? Wer sollte sie prüfen?

A. Ja, Prüfung druch unabhängige Dritte

B. Nein, es sollte weiter die klassische Modellprüfung durch Bafin/Bundesbank stattfinden

C. Ich habe noch keine Meinung dazu

Diskussionspunkt 4 – Sollte Risikomanagement von KI-Anwendungen im Rahmen des klassischen Modellrisikomanagements erfolgen? Kann man diese normieren?

A. Ja, und man sollte das klassische Model-Risk-Framework weiterentwickeln

B. Nein, man braucht völlig neue Ansätze wegen des Echtzeittraining

C. Ich habe noch keine Meinung dazu