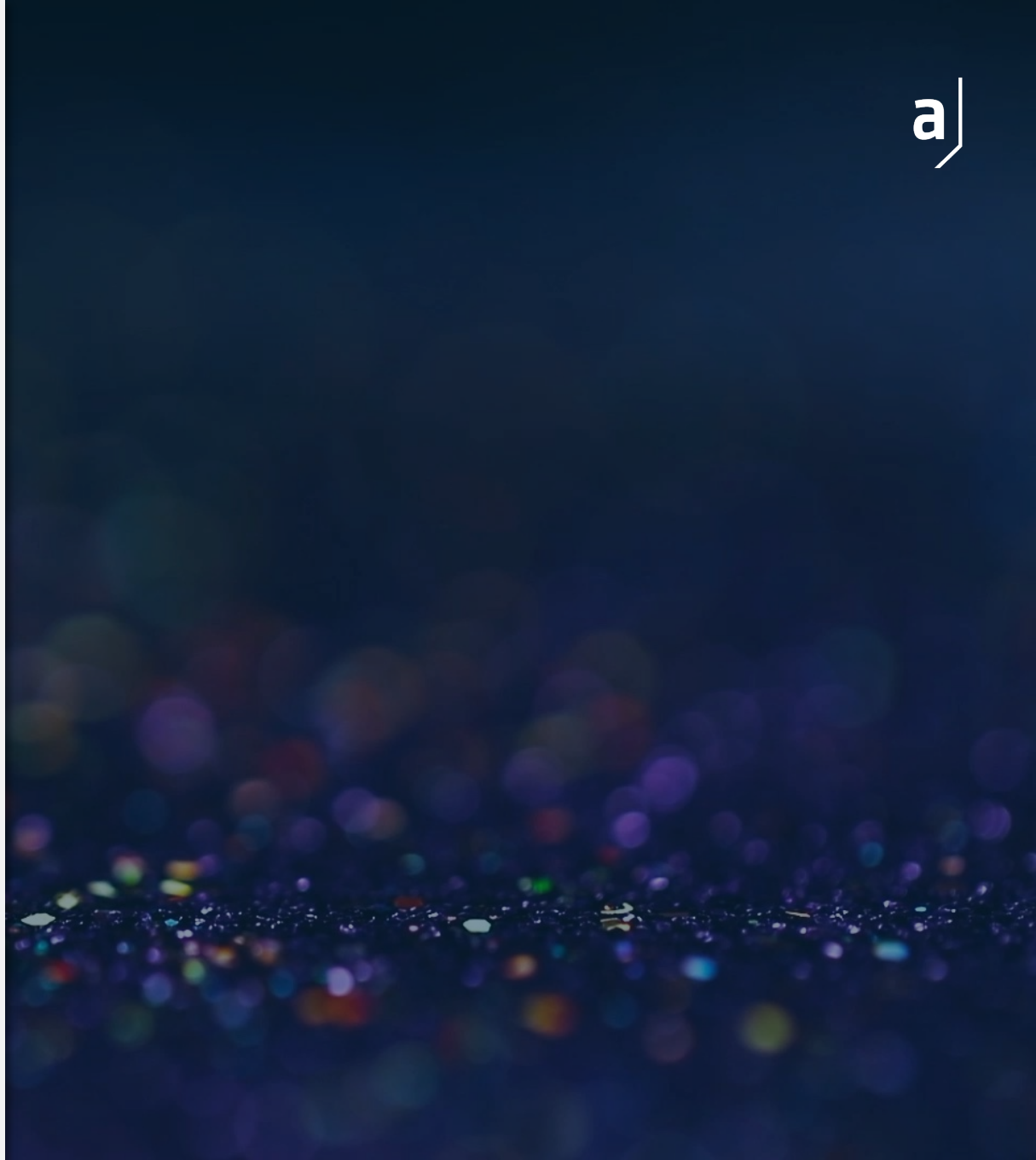
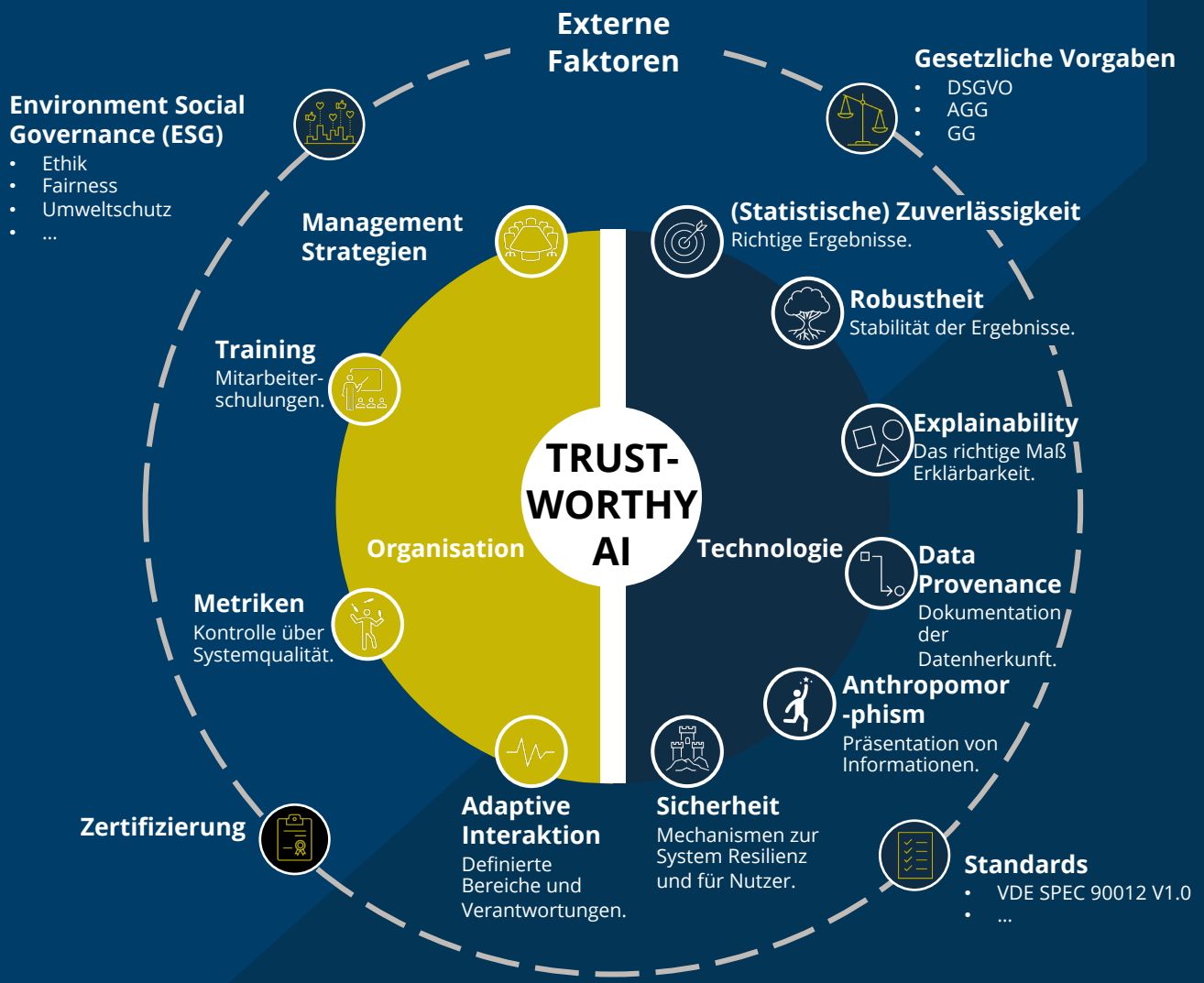
MAZ-Kommunikationstagung: Content Creation mit KI

TRUSTWORTHY AI

Lilian Do Khac



TRUSTWORTHY AI RAHMENWERK



KI-LEBENSZYKLUS

AI-Lifecycle

Specification → Implementation → Operationlization

CRISP-DM

Data Understanding	Data preparation	Modeling	Evaluation	Model deployment
• Data collection • Data description • Inspection of data • Evaluation of data	• Selection of data • Data cleaning • Transformation of data • Formation of data • Feature engineering	• Selection of model • Create base model • Optimize model • Evaluation of model	• Evaluation of results • Evaluation of total process	• Delivery of the model

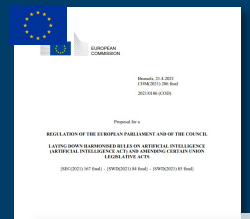
AI-Risks

- Failure or malfunction of ML application
- Oracle
- Evasion
- Compromise of ML application components
- Poisoning
- Model or data disclosure

EU-KI VERORDNUNG

HLEG Ethics guidelines for trustworthy AI

DIN/DKE – Normungsroad map v1.0



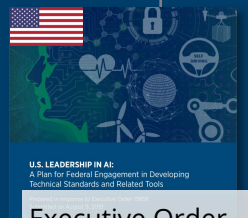
Proposal EU Artificial Intelligence Act

VDE – VDE SPEC 90012 V1.0 Principles to Practice

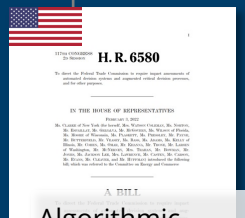
DIN/DKE – Normungsroad map v2.0

Norms/ Standards /Law

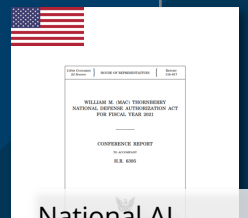
2019 | 2020 | 2021 | 2022 | 2023 | 2024/2025



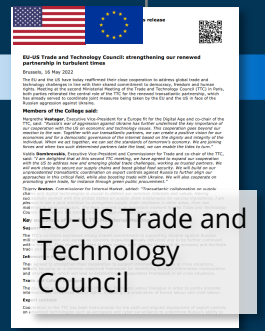
Executive Order 13859 U.S. Leadership in AI



Algorithmic Accountability Act 2019



National AI Initiative Act 2020

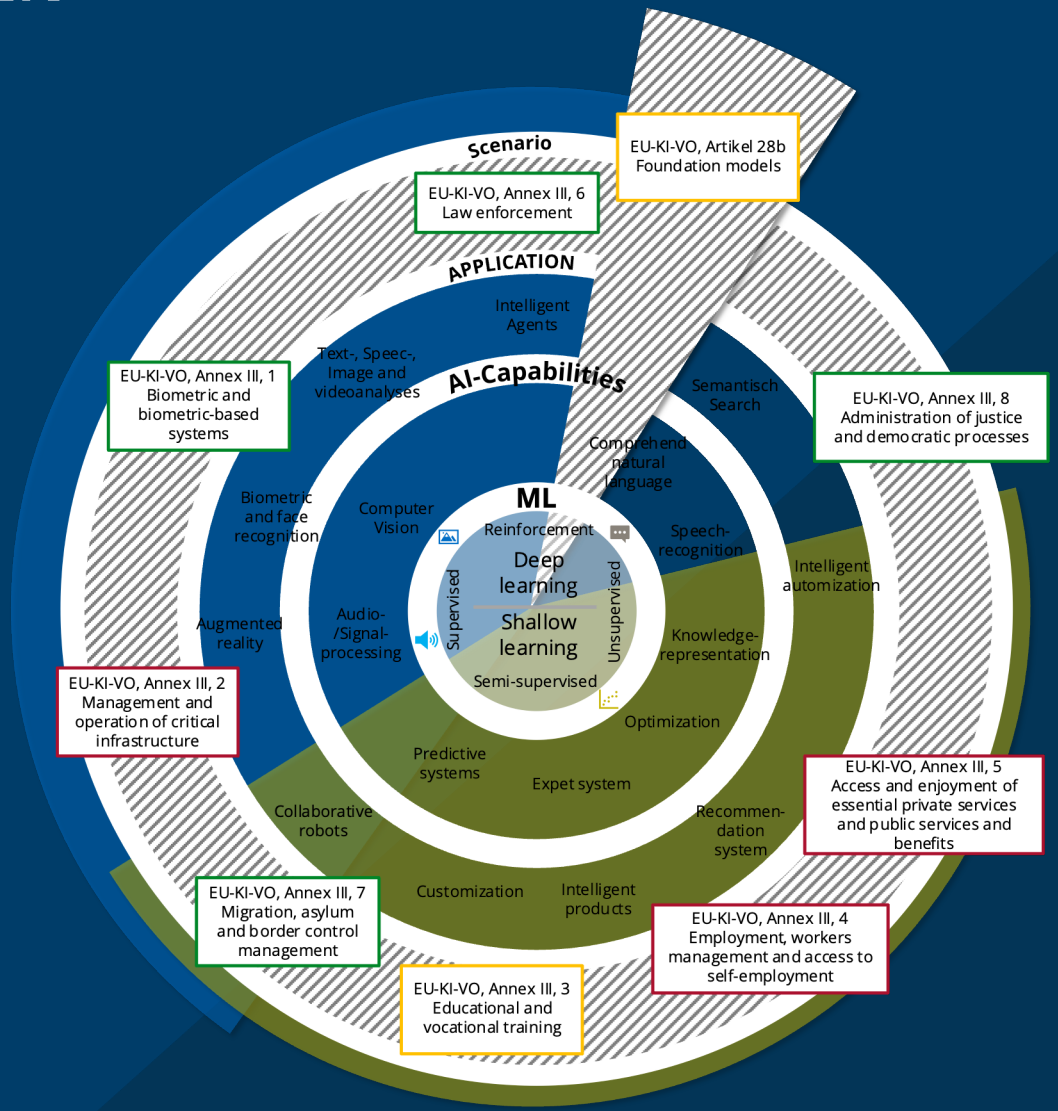


EU-US Trade and Technology Council

NIST - AI Risk Management Framework

Whitehouse – Executive Order

RISIKEN MINIMIEREN



SONSTIGES

Fachartikel

Do Khac, L., Gatzemann, R., Bunkus, T. (2022): Spannung zwischen Recht und „Richtig“, BI Spektrum 3/2022, <<Link>>

Do Khac, L., Kasseck, R., Smolanko, O. (2022): AI Compliance mit MLOps, BI Spektrum 5/2022, <<Link>>

Detering, H., Hammer, C., Do Khac, L. (2023): Compliance für Geschwindigkeit 4/2023, <<Link>>

Blogs

Trustworthy AI – das ganzheitliche Gerüst und was davon zukünftig geprüft wird, <<Link>>

Trustworthy AI – eine wilde Spannung zwischen Mensch und Maschine Interaktion, <<Link>>

Trustworthy AI – menschenzentrierte Anforderungen, <<Link>>

KI-Roulette – Eine grobe Orientierung von erlaubten und nicht erlaubten KI-Anwendungen, <<Link>>