Sind technische Regelwerke für sichere KI nur Roadmaps für Forschung und Konsens über Bauchgefühle?

Die Anwendung von Künstlicher Intelligenz (KI) im safety-kritischen Kontext birgt enormes wirtschaftliches Potenzial aber auch große Herausforderungen für die Forschung. Trotz unbeantworteter Forschungsfragen entstehen gerade eine Reihe von technischen Regelwerken, die Hilfestellung geben sollen, um KI im safety-kritischen Kontext einzusetzen. Dies geschieht sowohl in den einzelnen Anwendungsgebieten wie Fahrzeugtechnik als auch anwendungsübergreifend (z.B. in [1] und [2]). Bei den Lösungsansätzen lassen sich dabei zwei Stoßrichtungen erkennen. **Zielbasierte Ansätze** wie der in ISO/DPAS 8800 für Safety und KI bei Straßenfahrzeugen fordern eine Verfolgbarkeit zwischen der Anwendung von Methoden wie Testmethoden und der Qualitätseigenschaft Sicherheit. Aus wissenschaftlicher Sicht ist dies ein sauberer Weg. Aber eben nur ein Weg: Eine generische Roadmap für empirische Forschung, die für den konkreten Fall umgesetzt werden muss, da unklar ist welche Maßnahmen wie effektiv ist und wie Maßnahmenbündel im konkreten Fall zusammenwirken, um spezifische Ziele zu erreichen. **Regelbasierte Ansätze** liefern Anwendern ein Schema wie sie Maßnahmen auswählen, um Sicherheit zu erreichen, aber ohne wissenschaftliche Erklärung des Schemas zur Auswahl. Bezüglich KI gibt es bisher auch keine Nachweise, dass diese Ansätze auch tatsächlich zur gewünschten Sicherheit führen (können). Der Report ISO/IEC TR 5469 für funktionale Sicherheit und KI-Systeme nennt beispielsweise Methoden und Prozesse, um KI im safety-kritischen (anwendungsunabhängigen) Kontext zu nutzen. Die auf dem ISO/IEC TR 5469 aufbauende Spezifikation ISO/IEC TS 22440 wird in Teil 1 Anforderungen definieren, in Teil 2 Hilfestellung für deren Umsetzung geben und in Teil 3 Anwendungsbeispiele liefern. Die wissenschaftliche Erklärung, warum Sicherheit erreicht wird, wenn die Anforderungen in Teil 1 erfüllt sind, wird aber nicht entwickelt. Voraussichtlich werden etablierte Safety-Integritätslevels verwendet, um festzulegen welche Maßnahmen unter welchen Umständen empfohlen werden. Dies wirft einige Fragen auf: Sollen die empfohlenen KI-Maßnahmen für ein Sicherheitslevel ebenso effektiv sein wie die herkömmlichen Maßnahnahmen für dieses Sicherheitslevel? Falls ja, wie misst man die Effektivität der Maßnahmen? Sollte man eine neue Klasse von Fehlern und Maßnahmen definieren, um mit Unsicherheit bei KI umzugehen? Falls ja, wie definiert man KI-Unsicherheit (cf. DIN SPEC 92005 und ISO/IEC AWI TS 25223) und wie berücksichtigt man KI-Unsicherheit beim Integritätslevel (cf. VDE AR 284261 und $lamda_{AI}$ oder Uncertainty Confidence Indicator UCI)?

Der Vortrag gibt eine Übersicht über aktuelle Entwicklungen in der Normung zu Safety und KI, zeigt Ansätze, Zusammenhänge, Inkonsistenzen und Lücken auf, und weist auf aktuelle Debatten in der Normung hin. Dabei geht er auf zugehörige Forschungsbedarfe ein und lädt die Teilnehmer ein, die eigene Meinung einzubringen.

[1] ISO IEC JTC 1 SC 42
[2] CEN CENELEC JTC 21
[3] ISO/DPAS 8800 Road vehicles — Safety and artificial intelligence
[4] ISO/IEC TR 5469:2024 Artificial intelligence — Functional safety and AI systems
[5] ISO/IEC AWI TS 22440-1 Artificial intelligence — Functional safety and AI systems
[6] ISO/IEC AWI TS 22440-2 Artificial intelligence — Functional safety and AI systems
[7] ISO/IEC AWI TS 22440-3 Artificial intelligence — Functional safety and AI systems
[8] DIN SPEC 92005 Artificial Intelligence - Uncertainty quantification in machine learning
[9] ISO/IEC AWI TS 25223 Artificial intelligence — Guidance and requirements for uncertainty quantification in AI systems
[10] VDE AR 2842-61