# DIE ETHISCHE GEFAHRENANALYSE ALS FUNKTIONALER ANSATZ ZUR BEWERTUNG VON MACHINE LEARNING RISIKEN IN LUFTVERTEIDIGUNGSSYSTEMEN

H. Budig*, V. Gollnick†

* Technische Universität Hamburg, Institut für Lufttransportsysteme, Harburger Schloßstraße 22A, Hamburg, Deutschland

† Technische Universität Hamburg, Professor und Direktor des Instituts für Lufttransportsysteme, Harburger Schloßstraße 22A, Hamburg, Deutschland

**Zusammenfassung**

Die *Funktionale Gefahrenanalyse* (FHA) ist in der zivilen und der militärischen Entwicklung zur Bewertung der Tolerierbarkeit von Fehlerhäufigkeiten technischer Systeme weit verbreitet. Sie berücksichtigt dabei den Schweregrad von möglichen negativen Auswirkungen konkreter Fehlfunktion. Sie fokussiert sich jedoch ausschließlich auf technische Fehlfunktionen und deren Auswirkungen und bezieht ethische Auswirkungen nicht mit ein. Wir stellen die *Ethische Gefahrenanalyse* als Erweiterung der FHA vor. Hierbei zeigen wir, dass ethische Gefahren aufgrund von technischen Fehlfunktionen, aber auch bei korrekter technischer Funktion, entstehen können. Zudem wird im Luftverteidigungskontext die Zugehörigkeit einer betroffenen Person zu einer bestimmten Konfliktteilnehmergruppe relevant. Durch die zunehmende Aufmerksamkeit, die KI-basierte Systeme in der militärischen Anwendung bekommen, fokussieren wir uns für die EHA insbesondere auf Machine-Learning Systeme. Die aktuelle Forschungsliteratur zeigt, dass etablierte Verfahren zur funktionalen Risikobewertung für Machine Learning Systeme nicht ausreichend sind. Zur Adaption wird eine neue Methodik entwickelt, die den Weg für eine objektive Beurteildung der Zuverlässigkeit dieser Systemen ebnen soll. Die DIN ISO 5469 wird hierfür als eine methodische Grundlage verwendet. Zudem wird ein neues Klassifikationsschema zur Schweregradbewertung ethischer Gefahren notwendig. Wir wenden die EHA auf ein konkretes Beispiel der luftgestützten Objektdetektion an. Es wird dargelegt, welche technischen Fehlfunktionen und ethischen Gefahren auf einer konreten Funktionsebene auftreten können, und welche Auswirkungen zu erwarten sind. Diese Auswirkungen können bestimmten Zuverlässigkeitsklassen von konkreten ML-Algorithmen gegenübergestellt werden. Die EHA soll langfristig Entwickler und Entscheider unterstützen, indem es einen Bewertungsrahmen bietet, welche ML-Algorithmen für den Einsatz im entsprechenden Kontext und System zulässig sind.

**Keywords**

Machine Learning; Ethik; Funktionale Sicherheit; Luftverteidigungssysteme

## NOMENKLATUR

### Abkürzungen

| | |
|---|---|
| AVSI | Aerospace Vehicle Systems Institute |
| BNN | Bayesian Neural Network |
| CNN | Convolutional Neural Network |
| DIN | Deutsches Institut für Normung |
| DSP | Digitaler Signalprozessor |
| DT | Decision Tree |
| EHA | Ethical Hazard Assessment |
| FHA | Functional Hazard Assessment |
| GP | Gaussian Process |
| IEC | International Electrotechnical Commission |
| ISO | International Organization for Standardization |
| LLM | Large Language Model |
| LR | Linear Regression |
| ML | Machine Learning |
| RBS | Rule-based System |
| RF | Random Forest |
| RL | Reinforcement Learning |
| RNN | Recurrent Neural Network |
| RTCA | Radio Technical Commission for Aeronautics |
| SAE | Society of Automotive Engineers |
| SVM | Support Vector Machine |
| TR | Technical Report |

©2026

doi: 10.25967/680457

## 1. EINLEITUNG

Die *Funktionale Gefahrenanalyse*, *engl. Functional Hazard Assessment*, ist ein etablierter und weit verbreiteter Standard in der zivilen Luftfahrtindustrie. Sie stellt einen fortlaufenden Prozess dar, der während des gesamten Entwicklungsprozesses sowie während des Betriebes des entsprechenden Systems angewendet werden kann, um Risiken bestimmter Fehlfunktionen zu bewerten und zu minimieren. Sie ist in den beiden Normen SAE ARP4754 [1] und SAE ARP4761 [2] definiert. Die FHA arbeitet auf zwei Ebene: Auf der Luftfahrzeugebene, die höher liegende Gesamtfunktionen des Flugzeugsystems bewertet, und auf der Systemebene, die sich mit funktionalen Gefahren auf Teilsystemebene beschäftigt.
In den vergangenen Jahren gibt es vermehrt Publikationen, die sich mit der Anwendung der FHA und ähnlicher Sicherheitsprozesse auf Systeme befassen, die Machine Learning Algorithmen, bzw. *KI*[1], nutzen. Aufgrund der fundamentalen Black-Box Eigenschaften vieler dieser Algorithmen ist eine Rückführung auf logische, direkt messbare und auswertbare Prinzipien nicht mehr möglich, da diese durch den Trainingsprozess, insbesondere bei tiefen, sogenannten *Deep Learning* Algorithmen abstrahiert werden, und dem Menschen nach dem Training nicht mehr fassbar sind. Dies hat zum einen mit der mathematischen Struktur, zum anderen aber auch mit der schieren Größe des Suchraums zu tun, der eine hohe Dimensionalität aufweist [4] [5].
Des weiteren gibt es im Kontext von Luftverteidigungssystemen eine weitere Problematik: Die zunehmende Performanz von ML-Algorithmen, insbesondere im Bereich der Autonomie, lässt Anwendungen in den Blick geraten, die ethische Fragestellungen aufwerfen, und vor dem Einsatz dieser Algorithmen bewertet werden müssen, insbesondere hinsichtlich autonomer Waffensysteme (vgl. [6–8]). Dies sollte bestenfalls bereits während des Entwicklungsprozesses solcher Systems stattfinden, falls deren Anwendung sich nicht von vornherein aus ethischer Perspektive verbietet. [9, 10].
Aus diesem Grund stellt sich die Frage, ob die FHA um eine ethische Komponente erweitert werden kann, sodass auch ethische Auswirkungen während des Entwicklungsprozesses berücksichtigt werden können. Ethische Fragestellungen werden nicht allein von ML-Algorithmen aufgeworfen, sondern auch von konventionellen Systemen, allerdings stellen jene eine besondere Herausforderung dar. Die FHA bietet hier durch ihre Agnostizität hinsichtlich der Struktur der bewerteten Systeme das Potential, auch für ML-basierte Systeme anwendbar zu sein. Die Problematik ist allerdings tiefergehend. ML-Algorithmen lassen sich sehr schwer bezüglich ihrer Fehleranfälligkeit quantifizieren, wiederum ein Resultat aus ihrer nicht reduzierbaren Eigenschaft als Black-Box bzw. der Tatsache, dass diese nicht *zerlegbar* sind (*engl. non-decomposability*). Siehe hierzu auch [11].
Diese Arbeit stellt eine Erweiterung der Funktionalen Gefahrenanalyse vor: die *Ethische Gefahrenanalyse*, welche beide Aspekte vereinen soll, und somit das Gesamtsystem, ethisch wie technisch, von konventionellen wie auch ML-basierten Systemen auf seine funktionale Sicherheit untersuchen kann. Zu diesem Zweck nutzen wir Definitionen zur Klassifizierung des Schweregrades von technisches und ethischen Auswirkungen von Fehlfunktionen und Bewertungsmatrizen, die auf denen der klassischen Funktionalen Gefahrenanalyse aufbauen und von diesen inspiriert sind.

Die Arbeit ist wie folgt gegliedert. Kapitel 2 gibt eine Überblick über vorhandene Literatur zur Funktionalen Gefahrenanalyse sowie über bestehende Veröffentlichungen zur Problematik der Anwendung von Machine-Learning Algorithmen in sicherheitskritischen Bereichen, insbesondere in Luftfahrtanwendungen. Kapitel 3 gibt eine Einführung in das etablierte Vorgehen der Funktionalen Gefahrenanalyse. Kapitel 4 beschreibt bestehende Lücken des FHA bezüglich ML-basierten Systemen und stellt ein Schema zur Bewertung von Algorithmen hinsichtlich der Zuverlässigkeit ihrer Prädiktionen vor. Kapitel 5 bildet den Hauptteil. Es begründet die Notwendigkeit einer zusätzlichen *ethischen* Betrachtung während der FHA und stellt die *Ethische Gefahrenanalyse* inklusive einer Liste an Definitionen vor, die zur Schweregradbewertung von ethischen Auswirkungen eines technischen Systems genutzt werden kann. Kapitel 6 wendet die Ethische Gefahrenanalyse auf ein einfaches Beispiel aus dem Bereich der Luftaufklärung an. Kapitel 7 fasst die Ergebnisse zusammen und schließt mit einem kurzen Ausblick.

## 2. VERWANDTE ARBEITEN

Die beiden grundlegenden Standards zur FHA sind ARP4754 [1] und ARP4761 [2] der SAE. Ersterer beschreibt die Notwendigkeit der Durchführung einer FHA und definiert die grundlegende Vorgehensweise. Die Durchführung des Prozesses ist in SAE ARP4761 Anhang A [2] schließlich vollständig beschrieben. Der Zweck der FHA ist es, Fehlerfälle zu klassifizieren, welche zum einen durch Funktionsverlust, aber auch durch Fehlfunktionen verursacht werden können.
Hierfür wird zwischen zwei Analyseebenen unterschieden: (1) Luftfahrzeugebene und (2) Systemebene, wobei von der Luftfahrzeugebene als der Betrachtung der gröberen Ebene auf die Notwendigkeit der Betrachtung der Systemebene als der Speziellen geschlossen werden kann.
Heutzutage stellen ML-Algorithmen eine Herausforderung für klassische Sicherheitsanalysemethoden dar, da sie neue Funktionsparadigmen implementieren. Die SAE hat mit dem sog. *Statement of Concern* bezüglich dieser Problematik den Luftfahrt-

[1] Wir verzichten hier bewusst auf die Nutzung des Begriffes KI, da dieses Begriff zu umfangreich und unspezifisch ist. Wir verwenden den Begriff *Machine Learning*, der eine bestimme Klasse von Algorithmen beschreibt. Für eine ausführlichere Diskussion des Themas siehe [3].

©2026

**TAB 1. Übersicht über bestehende Standards, die in Zusammenhang mit der Bewertung von (funktionaler) Sicherheit von ML-basierten Systemen bzw. der Auftrittswahrscheinlichkeiten und dem Schweregrad von Fehlfunktionen stehen.**

| Standard | Beschreibung |
|---|---|
| SAE ARP4754 [1] | Grundlegende Bestimmung der Notwendigkeit der Funktionalen Gefahrenanalyse im Zertifizierungskontext für Luftfahrtanwendungen |
| SAE ARP4761 [2] | Vollständige Prozessdefinition der Funktionalen Gefahrenanalyse |
| RTCA DO178C [12] | Definition des *Software Levels* zur Bewertung der Kritikalität einer Fehlfunktion eines Softwaresystems im Luftfahrtkontext. |
| AVSI AFE 87 [5] | Bericht zur Problematik von ML-basierten Systemen in sicherheitskritischen Luftfahrtanwendungen mit einer ausführlichen Liste an Empfehlungen zu verschiedenen Aspekten der Nutzung und Implementierung von ML-basierten Systemen |
| SAE AIR6988 [4] | Identifizierung von Lücken in bestehenden Standards und Regulatorien hinsichtlich ML-basierten Systemen in Luftfahrtanwendungen. |
| DIN EN 61508 [13] | Standard zur funktionalen Sicherheit von programmierbaren elektronischen Systemen |
| DIN ISO/IEC/TR 5469 [14] | Etabliert ein simples Bewertungsschema nach Anwendungsebene und Technologieklasse zur funktionalen Sicherheit von ML-basierten Systemen (nicht luftfahrtspezifisch). Erweitert DIN EN 61508 um ML-basierte Systeme. |
| SAE ARP 6983[2] | Zukünftiger Prozessstandard zur Zertifizierung von ML-basierten Systemen in Luftfahrtanwendungen (Unveröffentlicht) |
| MIL-STD 882E [15] | Standard zur Risikobewertung und Analyse von Auftrittswahrscheinlichkeiten von Fehlfunktionen in militärisch genutzten technischen Systemen. |
| ISO/IEC TR 5259 [16] | Bewertung und Verbesserung der Datenqualität für ML-Anwendungen |
| ISO/IEC 24029 [17] | Methoden zur Bewertung der Robustheit von Neuronalen Netzen |

Informationsreport AIR6998 veröffentlicht [4]. Dieser bietet einen umfangreichen Überblick über bestehende Standards zur Sicherheitsanalyse von Luftfahrtsystemen und hebt bestehende Lücken hinsichtlich *intelligenter* bzw. *datengetriebener* Algorithmen hervor. Einige Hauptaspekte hier sind die fehlende Verifikationsmethoden für Trainingsdatensätze und die ungenügende Nachvollziehbarkeit der initialen Anforderungen nach dem Training des Algorithmus. Der Bericht hebt ausdrücklich hervor, dass die beschriebenen Lücken sich nicht nur auf die industrielle Anwendung dieser Algorithmen beziehen, sondern derart fundamental sind, dass bestehende Standards adaptiert werden müssen ([4], p. 29). Einige direkt erwähnte Standards sind SAE ARP4754 [1] und ARP4761 [2] (beide FHA), sowie der DO-178C [12] der RTCA zum Thema funktionale Sicherheit von Software in Luftfahrtsystemen. Die Arbeitsgruppe, die AIR6998 veröffentlicht hat, arbeitet aktuell an einem neuen Standard zum Umgang mit ML-Algorithmen, dem SAE ARP6983[2].

Eine weitere aufschlussreiche Analyse des problematischen Stands der Technik ist der Abschlussbericht AFE87 des AVSI aus dem Jahr 2020 [5]. Im Gegensatz zu AIR6988 fokussiert sich der AFE87 mehr auf die technischen Details des Machine Learnings und die intrinsischen Probleme hinsichtlich Vertrauenswürdigkeit und Interpretierbarkeit. Er beinhaltet eine ausführliche Liste an Vorschlägen für zukünftige Forschung zu Machine Learning in sicherheitskritischen Bereichen. Insbesondere fordert der Bericht Standards zu (1) der Kuration von Trainingsdatensätzen, (2) dem Training von ML-Algorithmen, (3) der Demonstration von Fällen von erfolgreich gewährleisteter Sicherheit und (4) der Simulation und Modellierung zur Sicherheitsgewährleistung von ML-Systemen ([5], p. 54).

Es existieren zwei weitere Normen mit Relevanz für diese Thematik, insbesondere hinsichtlich der funktionalen Sicherheit von Softwaresystemen. Dies sind die DIN EN 61508 [13] und die DIN ISO/IEC/TR 5469 [14]. Letztere ist von besonderem Interesse, da sie einen Ansatz zur Klassifizierung von ML-Systemen in sechs sog. Anwendungsebenen empfiehlt, welche die Kritikalität der Anwendung beschreiben sowie die Systeme hinsichtlich dieser in drei sog. Technologieklassen einteilt. Die Anwendungsebenen reichen von vollautonomen Entscheidungssystemen in sicherheitskritischen Umgebungen (Ebene A1) bis hin zu Systemen, die nicht Teil einer Sicherheitsfunktion sind und keinerlei sicherheitsrelevanten Entscheidungen treffen (Ebene D). Für die Technologieklassen, die dieser Standard weiterhin definiert, ist der bestimmende Faktor, ob standardisierte Konzepte bestehen, die zur Risikominimierung und damit der Erhöhung der funktionalen Sicherheit des untersuchten Systems verwendet werden können, oder ob die Konzepte zumindest adaptiert werden können. ([14], Tabelle 1).
Der Standard baut stark auf der DIN EN 61508 auf, der sich mit der funktionalen Sicherheit von programmierbaren elektronischen Systemen beschäftigt. Der DIN 5469 Standard vergleicht sein eigenes Vorgehen zum Umgang mit ML-basierten Systemen mit dem aus der DIN EN 61508 für verschiedene Aktivitäten, wie z.B. Test, Evaluation, Verifikation und Validierung, und gibt für jede Aktivität eine Interpretation ihrer Anwendbarkeit auf ML-Systeme ([14], Anhang A) an. Diese Liste ist umfangreich und eine gute Quelle zum Planen des Design-Prozesses für ein ML-System.
Eine beispielhafte Implementierung von Gefahrenklassifikationen in der Praxis stammt vom U.S.-

[2] *Work in Progress*, siehe: https://www.sae.org/standards/arp6983-process-standard-development-certification-approval-aeronautical-safety-related-products-implementing-ai (aufgerufen am 07.08.2026)

Amerikanischen Verteidigungsministerium[3]. Im Standard MIL-STD-882E [15] werden die möglichen Auswirkungen von Fehlfunktionen mit *Negligable*, *Marginal*, *Critical* und *Catastrophic* angegeben, sowie qualitative Wahrscheinlichkeitsgrenzwerte für Fehlfunktionen beschrieben. Eine kritische Konsequenz ist definiert als:

> *Could result in one or more of the following: permanent partial disability, injuries or occupational illness that may result in hospitalization of at least three personnel, reversible significant environmental impact, or monetary loss equal to or exceeding $1M but less than $10M.* ([15], S. 12)

Der MIL-STD-882E definiert sechs Wahrscheinlichkeitsebenen für Fehlerhäufigkeiten: *Frequent*, *Probable*, *Occasional*, *Remote*, *Improbable* und *Eliminated*. Diese bilden jedoch keine festen Grenzen hinsichtlich ihrer Auftrittshäufigkeit. Der Standard beschreibt, dass unwahrscheinliche (*improbable*) Ereignisse üblicherweise mit einer Wahrscheinlichkeit von weniger als $10^{-6}$ auftreten ([15], p, 12). Während der FHA werden diese Wahrscheinlichkeitsklassen dem Schweregrad der Fehlfunktion gegenübergestellt und resultierten in einer sog Risikomatrix, in der das korrespondierende Feld den Einsatz dieser Funktion mit den Risikoklassen ”Hoch” bis ”Niedrig” bewertet, abhängig vom existierenden Risiko und den möglichen Konsequenzen. Diese Matrix ist in höchstem Maße kontextabhängig und muss für den entsprechenden Anwendungsfall im Voraus definiert sein. Ein Beispiel einer solchen Risikomatrix ist in Tabelle 2 dargestellt.

Um Machine-Learning Algorithmen in dieses bestehende Schema einordnen zu können, ist es notwendig, diese hinsichtlich ihrer funktionalen Sicherheit bewerten zu können, um begründete Risikoabschätzungen treffen zu können. Eine wissenschaftliche Veröffentlichung, die sich unter anderem mit der Bewertung von ML-Algorithmen hinsichtlich der drei Faktoren *Simulierbarkeit*, *Zerlegbarkeit* und *Algorithmischer Transparenz* beschäftigt, stammt von Barredo Arrieta et al. [11]. Sie enthält eine Übersicht verschiedener Algorithmen zu diesen Eigenschaften, aus welchem Schlussfolgerungen für ihr Potential hinsichtlich der funktionalen Sicherheit abgeleitet werden können. Hierauf wird in Kapitel 4 näher eingegangen.
Wie aus den genannten Quellen erkennbar ist, erfährt die Regulierung von Machine Learning in sicherheitskritischen Anwendungen eine zunehmende Aufmerksamkeit. Allerdings gibt es bisher keine präzisen Ergebnisse, die anwendungsreife Bewertungskriterien und/oder Grenzwerte vorschreiben würden. Darüber hinaus, besonders im Luftverteidigungskontext, muss ein weiterer Aspekt berücksichtigt werden: Die ethischen Konsequenzen des Handelns automatischer oder sogar autonomer Systeme. Diese Problematik betrifft nicht exklusiv den Militärbereich, allerdings zeigt sie hier ihr volles Potential für schädliche und unethische Konsequenzen. Diese Problematik wird in Kapitel 5 näher ausgeführt.

[3]Es sei darauf hingewiesen, dass hier kein anderer Name für das U.S.-Amerikanische Verteidungsministerium genutzt wird, solange keine Namensänderung durch den U.S. Kongress formal beschlossen wurde.

## 3. FUNKTIONALE GEFAHRENANALYSE

Um zu beschreiben, wie ethische Aspekte in die Betrachtungen der *FHA* einbezogen werden können, stellen wir zuerst die etablierte Vorgehensweise der *FHA* vor und bauen auf dieser Struktur unser ethisches Bewertungssystem für ML-basierte Systeme auf.

Als Eingangsdaten benötigt die FHA folgende Dokumente: (1) Eine *Funktionale Analyse* auf der Luftfahrzeug- bzw. Systemebene. Diese ist eine vollständige Liste aller gewünschten und erwarteten Funktionen auf der gewählten Ebene. Die Ergebnisse dieser Analyse teilen sich in *Interne Funktionen*, *Ausgetauschte Funktionen* und weiter auf die verschiedenen Flugphasen auf. Ausgetauschte Funktionen (*engl. Exchanged Functions*) sind Funktionen, die Daten mit anderen Systemen oder Flugzeugen als den Betrachteten, austauschen. (2) eine sog. *Umwelt- und Notfallkonfigurationsliste (engl. Environmental and Emergency Configuration List)*. Diese beinhaltet externe Faktoren sowie Zustände, die aus Notsituationen resultieren, die während der Gefahrenanalyse mitbetrachtet werden müssen, bspw. verschiedene Wetterbedingungen oder Druckabfall in der Kabine während des Fluges ( [2] Anhang A 3.2.1). Wenn beide Dokumente vorliegen, wird die Gefahrenanalyse zusammen mit den initialen Designentscheidungen, z.B. Anzahl der Passagiere, durchgeführt. Hierbei werden für jedes System mögliche Fehlfunktionen identifiziert und aufgelistet. Dies inkludiert ebenfalls mehrere, simultane Fehler. Diese Fehlfunktionen werden dann hinsichtlich ihrer Auswirkungen, ihrer Bedingungen sowie der nötigen menschlichen Intervention analysiert. Anschließend werden diese Auswirkungen und Bedingungen hinsichtlich ihrer qualitativen und quantitativen Anforderungen geordnet und mit zusätzlichen Belegen begründet [1] [2].
Dies führt zu einer Liste aller funktionalen Gefahren mit einer Risikoklassifikation und Auftrittswahrscheinlichkeit, welche als Ausgangspunkt zur Minderung dieser Fehlfunktionen mittels eines iterativen Prozesses genutzt wird. Schließlich muss entschieden werden, welche Grenzwerte für die verschiedenen Schweregradklassen der möglichen Fehlfunktionen gelten sollen. Dies wird mittels einer Risikomatrix durchgeführt (vgl. Tab 2).

©2026

**TAB 2. Beispiel einer Risikomatrix mit beispielhaften Häufigkeiten von Systemfehlern nach Anhang A von MIL-STD-882E. Diese sollen eine Intuition über die Verbindungen zwischen qualitativen und quantitativen Systemfehlerraten geben. Die Wahrscheinlichkeiten sind über einen gewissen Expositionszeitraum gegeben, beispielsweise Flugstunden, Betriebsjahre, etc. ([15], p. 92). Auf der horizontalen Achse sind die maximal tolerierbaren Fehlerhäufigkeiten aus SAE ARP4754 pro Schweregradklasse angegeben [2].**

| | Catastrophic | Critical | Marginal | Negligable |
|---|---|---|---|---|
| | $P < 10^{-9}$ | $P < 10^{-7}$ | $P < 10^{-5}$ | |
| Frequent ($> 10^{-1}$) | High | High | Serious | Medium |
| Probable ($< 10^{-1}, \geq 10^{-2}$ | High | High | Serious | Medium |
| Occasional ($< 10^{-2}, \geq 10^{-3}$) | High | Serious | Medium | Low |
| Remote ($> 10^{-1}$) | Serious | Medium | Medium | Low |
| Improbable ($< 10^{-6}$) | Medium | Medium | Medium | Low |
| Eliminated | Eliminated | | | |

## 4. SCHWEREGRADBEWERTUNG UND ALGORITHMUSKLASSEN

Eine Problematik mit der notwendigen Risikoquantifikation von potenitellen ML-basierten Systemen ist in AFE87 angemerkt:

> *The traditional system to item requirement flow down is replaced with requirements that are empirically encoded with training and datasets.* ([5], S.41).

Das bedeutet, funktionale Gefahren können nicht auf bspw. bestimmte Codezeilen zurückverfolgt werden, sondern resultieren in nicht reduzierbaren Black-Box Modellen. Allerdings, da die FHA eine *funktionale* Analyse durchführt, und nicht Implementierung oder Algorithmen analysiert, bleibt die grundlegende Vorgehensweise valide, bringt allerdings die Notwendigkeit mit sich, das Vorgehen zu adaptieren und flexibler zu gestalten, was in [14] hervorgehoben wird. Diese Adaptierung bezieht sich insbesondere auf die Risikoquantifikation. Hinsichtlich der FHA in ARP4754 bemerkt die AIR6988 ebenfalls, dass die Anforderungen hinsichtlich Vollständigkeit und Richtigkeit implizit im Datensatz enthalten sind und dies neue Methoden erfordert, um den Datensatz hinsichtlich dieser Anforderungen untersuchen zu können ([4], p.21).

Die DIN ISO/IEC/TR 5469 [14] gibt einige Hinweise darüber, wie Datensatzqualität bewertet werden kann. Datensatzgenerierungen müssen umsichtig geplant und ausgeführt werden, und statistische und stochastische Verfahren müssen verwendet werden, um die Fairness und Verzerrungsfreiheit der Trainings-, Test- und Validierungsdatensätze zu gewährleisten. Wenn Fehlfunktionen während des Testens auftreten, müssen die Datensätze adaptiert und erweitert werden. Zudem muss sichergestellt werden, dass der Anwendungsbereich von Anfang an vollständig definiert ist. so dass keine Out-of-Distribution, d.h. nicht gelernte und somit nicht generalisierte, Eingaben während des Testens oder während der operationellen Ausführung auftreten. Zwei Konzepte spielen hier eine wichtige Rolle: (1) *Data drift* und (2) *Concept Drift*. Data Drift beschreibt das Phänomen, dass der Bereich der Eingangsdaten des ML-Modells während des Einsatzes nicht der gleiche Bereich wie während des Trainings ist. Das bedeutet, das System wird außerhalb seines Anwendungsbereiches betrieben. Dies benötigt entweder einen neuen Trainingszyklus oder eine Einschränkung des Einsatzbereiches auf den Bereich, welchen die Trainingsdaten abdecken.
Das zweite Phänomen, Concept Drift, beschreibt das Problem von sich verändernden operationellen Gegebenheiten. Selbst wenn die Eingangsdaten und der Prozess identisch bleiben, können sich andere Faktoren, wie das Wetter oder die Einsatzregeln verändern, die der Algorithmus nicht gelernt hat. Dies benötigt wiederum einen neuen Trainingszyklus des Modells. Wie man hieraus erkennen kann, muss der initiale Trainingsprozess gut geplant werden, um alle möglichen zukünftigen Veränderungen des Einsatzbereiches, wenn möglich, vorherzusagen und miteinzubeziehen, damit das System als funktional sicher angesehen werden kann ([14], Ch. 8).
ML-Algorithmen unterscheiden sich zudem stark hinsichtlich ihrer Transparenz[4] und Erklärbarkeit[5]. Intransparenz betrifft bspw. insbesondere die tiefen Methoden, bei denen starke nicht-Linearität und die schiere Menge an Parametern die internen Vorgänge des Systems, selbst mit technischen Hilfsmitteln, nicht mehr nachvollziehbar machen. Insbesondere für solche ML-basierten Systeme bleibt aktuell nur ein Weg zu evaluieren, ob sie funktional sicher sind: die empirische Analyse ihrer Performanz. Jedoch fehlen hierfür aktuell geeignete Bewertungsmetriken und insbesondere Grenzwerte. Es muss deutlich gemacht werden, dass Erklärbarkeit und Transparenz keine binären Metriken sind, sondern ein Spektrum. Barredo Arrieta et al. [11] beschreiben eine Ordnung

[4] Ein transparentes System ist definiert als eines, das relevanten Zielgruppen angemessene Informationen seiner internen Abläufe liefern kann [14, 18]

[5] Ein erklärbares System ist definiert als eines, das auf eine für Menschen verständliche Art wichtige Faktoren seines Entscheidungsfindungsprozesses kommunizieren kann [14, 18]

©2026

**TAB 3. Risikofaktoren in ML-basierten Systemen, die als Grundlage zur Bewertung der funktionalen Sicherheit dienen können (nach DIN iSO 5469 [14], Kapitel 8). Diese betreffen ganze ML-basierte Systeme, nicht direkt die darunterliegenden Algorithmen. Zur Bewertung der Zuverlässigkeit von bestimmten ML-Algorithmen sind insbesondere die Eigenschaften Transparenz, Erklärbarkeit und Resilienz gegen Angriffe relevant und werden im Folgenden als Grundlage verwendet.**

| Bezeichnung | Fragestellung |
|---|---|
| Automatisierungsgrad | Inwieweit funktioniert das System ohne menschlichen Eingriff und Überwachung? |
| **Grad an Transparenz und Erklärbarkeit** | Inwieweit können wichtige Entscheidungsparameter für Menschen verständlich kommuniziert werden (Erklärbarkeit) und inwieweit können Informationen über interne Prozesse für relevante Interessensgruppen bereitgestellt werden (Transparenz)? |
| Komplexität der Umgebung und der Spezifikationen (inkl. Data Drift und Concept Drift) | Inwieweit beeinträchtigt die Komplexität der Umgebung die Transparenz und Erklärbarkeit des Systems? Wie weit liegen die tatsächlichen Eingabedaten entfernt von den Trainingsdaten (Data Drift)? Inwieweit haben sich die Rahmenbedingungen zu den Trainingsdaten verändert (Concept Drift)? |
| Erwägungen zur Systemhardware | Wie zuverlässig ist die Hardware? Inwieweit können Hardwarefehler die funktionale Sicherheit des Systems beeinträchtigen? |
| **Resilienz gegen Angriffe** | Wie anfällig ist das System für Angriffe die (1) die Fehler im Softwareablauf verursachen und (2) absichtlich zu falschen Ausgaben führen, ohne Fehler im Softwareablauf zu verursachen |
| Technologiereifegrad des Systems | Wie groß ist die Fehlerhäufigkeit in einem bestimmten Anwendungskontext? |

von verschiedenen Arten von ML-Algorithmen nach den drei Kriterien Simulierbarkeit, Zerlegbarkeit und Algorithmische Transparenz. Am einen Ende des Spektrums, mit direkt interpretierbaren Vorhersagen, liegen die Lineare und die Logistische Regression, allerdings bereits mit einer herausfordernden Menge von Variableninteraktionen, welche mathematische Analysetools zum Verständnis der komplexen Interaktionen benötigt, die während der Ausführung des entsprechenden Algorithmus auftreten. Am anderen Ende des Spektrums liegen die *Deep Learning* Algorithmen, wie bspw. CNNs und RNNs. Diese haben keinerlei inhärente Transparenz und können ohne unerreichbar hohen Rechenaufwand weder sinnvoll zerlegt noch simuliert werden. In diesem Bereich liegen ebenfalls zwei Algorithmen aus dem klassischen, d.h. nicht-tiefen Machine Learning: SVMs und Random Forests. All diese benötigen *post-hoc* Analysen um ihren Entscheidungsprozess auch nur ansatzweise verstehen zu können [11].

Der Übergang zwischen den Methoden, die *de facto* unerklärbar sind und denen, die zumindest in einem gewissen Maße nachträglich interpretiert werden können, liegt zwischen den BNNs, welche statistisch gut begründete und nachvollziehbare Lösungen ausgeben und den Tree Ensembles, welche bereits eine überwältigende Komplexität aufweisen. Das bedeutet, dass die Eigenschaft der Unerklärbarkeit nicht ausschließlich tiefen Lernmethoden zukommt, sondern sich auch komplexeren Methoden des klassischen Machine Learnings hier wiederfinden. Das Problem liegt darin, dass es eine Korrelation zwischen der Durchdringungstiefe der Algorithmen in der Gesellschaft, verursacht durch ihre hohe Performanz insbesondere von Modellen zur Verarbeitung natürlicher Sprache und der Bildgenerierung, und ihrem geringen Grad an Erklärbarkeit gibt. Dies ist der Grund, warum die Verwendung dieser Algorithmen im Kontext der FHA besonders kritisch betrachtet werden muss.

Wir können uns eine Zuweisung zwischen den Hierarchien hinsichtlich Zerlegbarkeit, Transparenz, etc. von Barredo Arrieta et al. [11] und den Technologieklassen aus der DIN ISO 5469 [14] vorstellen. Der DIN ISO 5469 definiert drei voneinander abgegrenzte Technologieklassen für ML Algorithmen, die sich darin unterscheiden, ob die ML-basierten Systeme (im Standard als ”KI-Technologie” bezeichnet) anhand von bestehenden Standards und Normen zur funktionalen Sicherheit entwickelt wurden und, falls dies nicht der Fall sein sollte, zumindest ergänzende Verfahren und Techniken zur Erreichung der gewünschten Eigenschaften und Risikominderung ermittelt und genutzt werden können ([14], Kapitel 6.2). Da nach Wissen der Autoren zum aktuellen Zeitpunkt keine Standards zur funktionalen Sicherheit spezifischer Algorithmen existieren, ist Technologieklasse 1 aktuell außer Reichweite. Klasse 2 ist erreichbar, Standards wie bspw. [16] and [17] definieren Metriken und Methoden für den robusten Gebrauch von Neuronalen Netzwerken sowie für den verantwortungsvollen und sinnvollen Aufbau von Trainingsdatensätzen. Darüber hinaus gibt die DIN 5469 Beispiele für zusätzliche Parameter, um die mögliche Risikominimierung für ein gegebenes, ML-basiertes System zu verifizieren. Die wichtigsten genannten Eigenschaften hierfür sind in Tabelle 3 dargestellt.
Diese Eigenschaften sind nicht universell, d.h. nicht alle sind direkt auf Algorithmen anwendbar, sondern

©2026

**TAB 4. Beispielklassifikation von prominenten Algorithmenklassen nach den fünf Kriterien Transparenz, Erklärbarkeit, Robustheit, Reward Hacking und Eigenständige Unsicherheitsvorhersage. Mithilfe dieser Klassifikation können die Algorithmen nach Zuverlässigkeit sortiert werden. Dies kann später helfen, geeignete Algorithmen für Fehlfunktionen mit bestimmten Schweregraden auszuwählen. Diese Ordnung beruht auf der Erfahrung der Autoren und teilweise auf Barredo Arrieta et al. [11] und ist stark subjektiv geprägt. Konkrete Metriken und Grenzwerte stehen noch aus.**

| | GP | LR | BNN | RBS | DT | SVM | RF | CNN | RNN | RL | LLM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Transparenz | ✓ | ✓ | ✗ | O | ✓ | ✗ | ✗ | O | ✗ | ✗ | ✗ |
| Erklärbarkeit | O | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | O | ✗ | ✗ | ✗ |
| Robustheit | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | O | ✗ | ✗ | ✗ | ✗ |
| Kein Reward Hacking | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| Unsicherheitsvorhersage | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Klassifizierung | I | I | I | I-II | II | II | III-IV | III-IV | IV | IV | IV |

beziehen sich auf ML-basierte Systeme, wie z.B. der Technologiereifegrad. Wir wollen uns hier auf grundlegende Kriterien zur Bewertung von Algorithmen beschränken. Beispielsweise sind Data und Concept Drift eine Frage von gutem Datensatzdesign und des Wahrens des definierten Einsatzbereiches. Allerdings existiert, z.B. im Reinforcement Learning, der Datensatz nur implizit in Form von Simulationsumgebungen, die zum Training genutzt werden. Darüber hinaus bezieht sich auch der Grad an Automatisierung nicht auf die Eigenschaften des Algorithmus, sondern darauf, wie die Ausgaben des Algorithmus genutzt werden. Deshalb fokussieren wir uns zur weiteren Untersuchung auf die Aspekte *Transparenz*, *Erklärbarkeit* und *Robustheit*, womit Resilienz gegen Angriffe gemeint ist, aber auch die natürlich vorkommende Schwankungen in den Eingabedaten, die zum Beispiel bei der Nutzung von physikalischen Sensoren auftreten.

In dieser Arbeit sollen vorerst keine konkreten Bewertungsmaßstäbe für diese Parameter dargestellt werden, allerdings ein Schema vorgestellt werden, wonach sich verschiedene Typen von Algorithmen grob in *Zuverlässigkeitsklassen* einordnen lassen, die später für die Risikobewertung verwendet werden können. Dies soll in vier Klassen I-IV geschehen, wobei die Klasse I das höchste und die Klasse IV das niedrigste Maß an Zuverlässigkeit darstellt. Dies baut auf den drei genannten Parametern *Transparenz*, *Erklärbarkeit* und *Robustheit* aus der DIN 5469 auf, wird allerdings noch um zwei Parameter erweitert: (1) Eigenständige Unsicherheitsprädiktion seiner Ausgabewerte und (2) das zu erwartende Reward Hacking. Reward Hacking bezieht sich im Wortsinne direkt auf das Reinforcement Learning und beschreibt das Finden von unzulässigen Lösungen durch fehlerhaften Aufbau der Reward Struktur. Der Algorithmus findet ungewollte ”Abkürzungen”. Reward Hacking kann aber auch im Sinne des *Overfitting* verstanden werden, also das *Auswendiglernen* der Trainingsdaten ohne Verallgemeinerung der zugrundeliegenden Strukturen, was ein fundamentales Problem im Machine Learning darstellt.

Algorithmen werden schließlich anhand dieser fünf Eigenschaften in die jeweiligen Klassen eingeteilt(vgl. Tab. 4). Da nach Kenntnisstand der Autoren keine ML-Algorithmen existieren, die alle Eigenschaften erfüllen, fordern wir vier von fünf erfüllte Eigenschaften für Klasse I, drei von fünf für Klasse II, zwei von fünf für Klasse III und weniger als zwei für Klasse IV. Es steht weiterhin zur Diskussion, die Eigenschaft der eigenständigen Unsicherheitsprädiktion zu einer Bedingung für Klasse I zu machen, da dies eine Fehlerbaumanalyse ermöglichen würde. Dies ist aber noch nicht abschließend entschieden.

Die Klassen I und II beschreiben somit Klassen von Algorithmen, die voraussichtlich an die vorgegebenen Anforderungen an die funktionale Sicherheit angepasst werden könnten. Diese entsprechen der Technologieklasse II aus der DIN 5469 [14] mit geringem bis mittlerem Aufwand. Die Klasse III enthält Algorithmen, welche die Garantien aus den Klassen I und II nicht besitzt, für die es aber etablierte post-hoc Methoden zur Analyse von Unsicherheit und Robustheit geben kann. Klasse III könnte bspw. Algorithmen enthalten, für die praktisch erprobte Werkzeuge zur Visualisierung ihrer Entscheidungsfindung existieren. Klasse IV enthält schließlich die Algorithmen, für die solche Werkzeuge nicht existieren, und die keine inhärenten Erklärungen ihrer Entscheidungsfindung bieten. Beispiele für Algorithmen, die Klasse I erreichen könnten sind die BNNs oder die GPs. Beide sind stochastische Verfahren, die mathematisch konsistente und nachvollziehbare Vorhersagen inklusive Unsicherheitsquantitifzierungen bieten. Beispiele für Algorithmen, die potentiell in Klasse IV eingeordnet würden, sind LLMs und Reinforcement Learning, die zwar performant und weit verbreitet sind, aber keinerlei Garantien hinsichtlich der Richtigkeit ihrer Vorhersagen geben (vgl. Bild 1).

Es sei darauf verwiesen, dass diese Klasseneinteilung noch vage ist, dass wir auf dieser jedoch aufbauen möchten um präzisere Metriken zur Algorithmusklassifikation zur Bewertung der funktionalen Sicherheit zu entwickeln. Dieses Schema wird in Zukunft um Metriken erweitert werden, die sich auch auf zukünftige Standards zur funktionalen Sicherheit

©2026

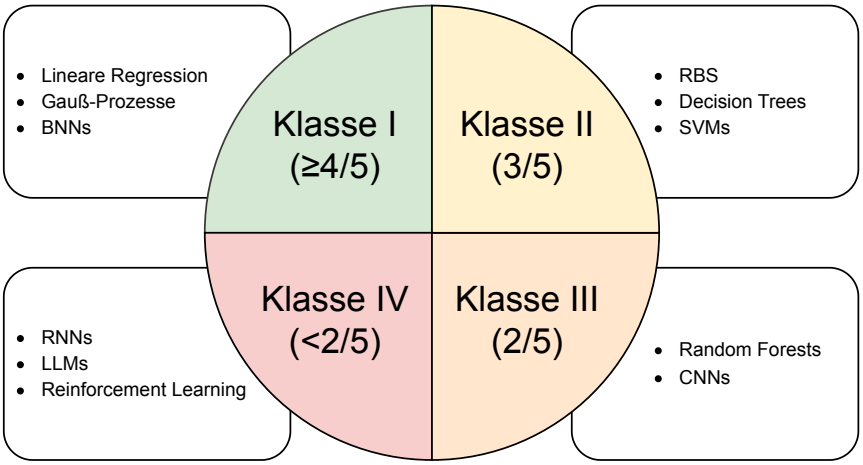



**BILD 1. Übersicht der beispielhaften Klassifizierung verschiedener gängiger Algorithmen anhand der vorgestellten Zuverlässigkeitsmetrik mittels der Anzahl der erfüllten Kriterien. Durch diese Klassen kann eine Zuordnung zu den Schweregradklassen der ethischen und technischen Auswirkungen getroffen werden, siehe Tabelle 8.**

von ML-basierten Systemen stützen sollen.
Während der Durchführung der FHA werden die ML-basierten Funktionen von jenen Funktionen getrennt, die kein Machine Learning nutzen, für welche das FHA in seiner klassischen Form angewendet werden kann. Basierend auf dem oben genannten Klassifikationsschema können im Anschluss verschiedene Maßnahmen getroffen werden, um das Risiko der entsprechenden ML-basierten Funktion zu bewerten, und zu elaborieren, welche Methode geeignet sind, das Risiko zu mindern. Je höher die Gefahrenklasse, desto umfangreicher und potenter müssen dieser Methoden sein.

## 5. ETHISCHE GEFAHRENANALYSE

Technische Systeme können nicht nur Gefahren hinsichtlich physischer oder finanzieller Schäden aufweisen, sonder zusätzlich *normative* Auswirkungen haben. Gerade im Kontext von Luftverteidigungssystemen treten diese zutage, denn hier bewegen sie sich im Rahmen des humanitären Völkerrechts. Dieses setzt einen rechtlichen Rahmen darüber, welche Rechte und Pflichten die verschiedenen beteiligten Gruppen im Konflikt haben. Normative Auswirkungen leiten sich jedoch nicht ausschließlich aus gesetzlichen Bestimmungen ab. Darüber hinaus gibt es ethisch-moralische und gesellschaftliche Normen, die von einem technischen System gefährdet, geschwächt oder gebrochen werden können, beispielsweise das Vertrauen in die Zuverlässigkeit eines Systems oder der Gleichbehandlungsgrundsatz. Für weitere Beispiele siehe Tabelle 5. Somit wird deutlich, dass zusätzlich zur technischen Betrachtung der Auswirkung von Fehlfunktionen eine normative Betrachtung ergänzt werden muss.

Hinsichtlich der ethischen Analyse eines zu untersuchenden technischen Systems präsentieren wir zusätzlich zur Zuverlässigkeitsklassifikation der ML-Algorithmen zwei neue Erweiterungen zum FHA. Zuerst wird eine neue Art von Auswirkungen, die *normativen* Auswirkungen hinzugefügt. Diese beinhalten zwei Unterklassen: (1) normative Auswirkungen aufgrund von technischen Fehlfunktionen und (2) normative Auswirkungen, die ohne eine technische Fehlfunktionen auftreten. Diese normativen Auswirkungen können parallel zu den technischen Fehlfunktionen auftreten und lassen sich auf ML-basierte Funktionen, aber auch auf klassische, nicht ML-basierte Funktionen anwenden.
Ein Beispiel für eine normative Gefahr ohne technische Fehlfunktion wäre ein Algorithmus, der feindliche Einheiten mittels eines Computer Vision Systems richtigerweise als feindlich klassifiziert. Dies funktioniert allerdings über die Zeit so gut, dass der Mensch, der diese Information weitergibt, irgendwann aufhört, die Entscheidung selbst zu überprüfen, und über Zeit verlernt, wie er selbst die Entscheidung treffen könnte (sog. *Deskilling*)[6]. Ein Beispiel für eine normative Auswirkung durch eine technische Fehlfunktion wäre in diesem Kontext der Angriff auf eine verbündete Einheit, die durch den Algorithmus zuvor fälschlicherweise als gegnerisch identifiziert wurde.
Die zweite Erweiterung ist die *Betroffene Gruppe*, d.h. die Gruppe, die von den Auswirkungen einer Fehlfunktion, betroffen ist. Da die gesamte Analyse

[6] Für eine generelle Einführung in das Theme, siehe [26] und [27], die beschreiben, wie die Automatisierung von Prozessen dem Nutzer Schwierigkeiten machen kann und die Produktivität senken kann. Für Literaturrecherchen, siehe [28] und [29]. Für eine praktische Fallanalyse, siehe [30]. Für ein empirische Studie zum Thema, siehe [31]. Kim et al. [32] zeigen die psychologischen Konsequenzen auf und Vallor [33] geht auf das sogenannte *moral deskilling* ein.

©2026

**TAB 5. Beispiele für weit verbreitete rechtliche, ethische und gesellschaftliche Normen. Diese korrespondieren nicht notwendigerweise zueinander, bauen aber teilweise aufeinander auf. Viele ethische Normen sind in Gesetzen gespiegelt. Das Völkerrecht ist als besonderer Rechtsfall im Kontext von bewaffneten militärischen Konflikten zu sehen und verbietet im Gegensatz zum zivilen Strafrecht bspw. die Körperverletzung nicht pauschal, sondern nur an illegitimen Zielen. (*GC: Geneva Convention, AP: Additional Protocol*)**

| **Völkerrechtliche Norm** [19–23] | **Strafrechtliche Norm [24] [25]** | **Ethisch-Moralische Norm** | **Soziale/Gesellschaftliche Norm** |
|---|---|---|---|
| Verbot des Zielens auf Zivilisten oder zivile Objekte (AP I Art. 51.2) | Verbot von Mord, Totschlag etc. (§211, §212 StGB) | Man soll nicht töten | Ein technisches System soll die Funktionen zuverlässig erfüllen, für die es entwickelt wurde |
| Verbot der Misshandlung von Kriegsgefangenen (GC III Art. 13, 14) | Verbot des Diebstahls und Raubes (§242, §249 StGB) | Man soll nicht stehlen | Ein technisches System soll missbräuchlicher Nutzung vorbeugen |
| Verbot des Angriffes auf militärische Ziele mit exzessiven (überproportionalen) zivilen Opfern (AP I Art. 51.5.b) | Verbot von Diskriminierung (Art. 3 GG) | Man soll alle Menschen gleich behandeln | Man soll nicht neidisch sein |
| Pflicht zur Inobhutnahme und Versorgung von Verwundeten (GC I Art. 3.2, 12; AP I Art. 10) | Verbot der Unterlassenen Hilfeleistung (§323c StGB) | Man soll Menschen in Not helfen | Man soll nicht verschwenderisch sein |
| Verbot des Angriffs auf mit einem geschützten Symbol gekennzeichneten Einheiten (GC I Art. 19, 24, 42; AP I Art. 12, 18) | Verbot der Körperverletzung (§223 StGB) | Man soll keinem Lebewesen Gewalt antun | Man soll nicht gierig sein |
| Verbot der missbräuchlichen Nutzung von geschützten Symbolen (GC I Art. 44: AP I, Art.38) | Wahrheitspflicht vor Gericht (§153 StGB) | Man soll nicht lügen | Man soll nicht schlecht über andere reden |
| Verbot von Perfidität, also einer feindlichen Handlung unter der Vortäuschung von rechtlichem Schutz (AP I Art. 37) | Verbot der Brandstiftung (§306 StGB) | Man soll nicht fremdes Eigentum zerstören | Man soll Mitleid mit Benachteiligten haben |
| Verbot der Deportation von Zivilisten (GC IV Art. 49) | Verbot der falschen Verdächtigung (§164 StGB) | Man soll die Privatsphäre des Einzelnen achten | Man soll keine ungefragten Bild- oder Tonaufnahmen anderer Menschen erstellen |
| Verbot des Angriffs auf Gebäude mit allgemeingefährlichem Inhalt, wie zum Beispiel Atomkraftwerke (AP I Art. 56) | Verbot der Bestechung und der Bestechlichkeit (§332, §334 StGB) | Man soll die Umwelt bewahren | Man soll anderen Menschen mit Respekt entgegentreten |

dieser Arbeit im Kontext von Luftverteidigungssystemen steht, berufen wir uns hier auf die Prinzipien des humanitären Völkerrechts [19–23][7], um sinnvolle Unterscheidungen treffen zu können. Diese sind in Tabelle 6 genauer erläutert. Hierauf aufbauend sind die möglichen betroffenen Gruppen definiert, die in Tabelle 7 dargestellt sind.

Für das Anwendungsbeispiel, dass in Kapitel 6 beschrieben ist, werden Kriterien zur Klassifizierung von ethischen und technischen Auswirkungen von Fehlfunktionen definiert. Diese sind aus Platzgründen in Anhang A aufgeführt. Diese Definitionen beschreiben möglichst allgemein Kriterien zur Einordnung der ethischen und technischen Folgen von Fehlfunktion in die Klassen *Vernachlässigbar* bis *Katastrophal* u.A. anhand des Ausmaßes des Schadens, der Zahl der betroffenen Personen oder Objekte und der Regelungen des internationalen humanitären Völkerrechts (vgl. RTCA DO178C [12]). Die Definitionen der technischen Auswirkungen sind größtenteils aus dem MIL-STD 882E übernommen [15].

Ein wichtiger Faktor zur Beachtung während der EHA ist die funktionale Rolle des System im Gesamtsystemkontext. Ein ML-Algorithmus zur Rauschunterdrückung im Sensorsystem eines Luftfahrzeugs kann selbst im Worst-Case Scenario nicht über das Luftfahrzeug hinaus wirken. Im Gegensatz dazu potenziert sich das Risiko bspw. bei Missionsplanungsalgorithmen. Darüber hinaus ist relevant, ob es sich um eine *assistierende* oder eine *delegierte* Komponente handelt, d.h., der Grad an Autonomie wird entscheidend (vgl. [14]).

Auch wenn dieses Schema auf Basis geltenden Rechts und grundlegender ethischer Prinzipien entwickelt wurde, soll es nicht als **korrekte** Lösung dargestellt werden; es ist ein vielmehr beispielhaftes Schema, um den Prozess des vorgestellten EHA zu verdeutlichen. Wenn diese Struktur in der Zukunft durch Organisationen, Firmen oder Regulierungsbehörden zur Definition von spezifischen Klassifikationen genutzt werden wird, muss dies aufgrund der ethischen Problemstellungen und der potentiellen Gefahren durch einen elaborierten Prozess, der interdisziplinäre Kommissionen mit Experten aus Ethik, Recht, Militär und dem Ingenieurwesen miteinbezieht, geschehen. Aufgrund der Menge an Einflüssen und nötigen Kompromissen wird es hierfür keine objektiv eindeutige Lösung geben; allerdings wird ein möglichst umfangreicher Konsens die Durchdringung in verschiedene Fachbereiche dieses wichtigen Themas zur funktionalen Sicherheit erhöhen.

## 6. ANWENDUNGSBEISPIEL FAHRZEUGKLASSIFIZIERUNG

Im Folgenden soll die Funktionsweise des EHA an einem Beispiel verdeutlicht werden. Bei dem Beispielsystem handelt es sich um ein hypothetisches *Intelligence, Surveillance & Reconnaissance* System, bestehend aus einer Aufklärungsdrohne, ähnlich einer Global Hawk-Drohne mit Kamerasystem und

[7] Die gültigen Sprachen dieser völkerrechtlichen Abkommen sind unterschiedlich. Bei den Genfer Konventionen sind die offiziellen Sprachen Englisch und Französisch. Beim ersten Zusatzprotokoll sind es zusätzlich Arabisch, Chinesisch, Russisch und Spanisch. Eine offizielle deutsche Version existiert nicht.

©2026

**TAB 6. Auflistung und Erläuterung der vier Prinzipien des humanitären Völkerrechts: Notwendigkeit, Unterscheidung, Proportionalität und Menschlichkeit, abgeleitet aus dem 1. Zusatzprotokoll (AP I) zur Genfer Konvention [23].**

| Völkerrechtliches Prinzip | Erklärung |
|---|---|
| Militärische Notwendigkeit | • Es ist nur so viel Gewalt anzuwenden, wie nötig ist, um das Ziel zu erreichen.<br>• Es ist verboten, Waffen, Projektile sowie Materialien und Methoden der Kriegsführung zu verwenden, die geeignet sind, überflüssige Verletzungen oder unnötiges Leid zu verursachen (AP I, Art. 35.2).<br>• Bei der Planung und Durchführung von Angriffen muss man "bei der Wahl der Angriffsmittel und -methoden alle möglichen Vorkehrungen treffen, um zufällige Verluste an Menschenleben, Verletzungen von Zivilisten und Schäden an zivilen Objekten zu vermeiden und auf jeden Fall zu minimieren"(AP I, Art. 57.2.a.ii). |
| Unterscheidung | • Kombattanten müssen zwischen Zivilisten und Kombattanten unterscheiden und dürfen nur letztere direkt angreifen. (AP I, Art. 48,51)<br>• Wahllose Angriffe sind verboten (AP I, Art. 51).<br>• Eine Person, die sich nicht im Kampf befindet, oder als außer Gefecht gesetzt gelten sollte, darf nicht Gegenstand eines Angriffes werden (AP I, Art. 41). |
| Proportionalität | • Stellt sicher, dass Gewalt nicht in einem Missverhältnis zu den eigenen Zielen steht.<br>• Angriffe gelten als wahllos, wenn erwartbar ist, dass sie zu zufälligen Todesfällen unter der Zivilbevölkerung, Verletzungen von Zivilisten, Schäden an zivilen Objekten oder einer Kombination davon führen, die im Verhältnis zum erwarteten konkreten und direkten militärischen Vorteil übermäßig hoch wären (AP I, Art 51.5.b). |
| Menschlichkeit | • Zivilisten und Kombattanten bleiben unter dem Schutz und der Geltung der Grundsätze des Völkerrechts, wie sie sich aus feststehenden Gewohnheiten, aus den Grundsätzen der Menschlichkeit und aus den Forderungen des öffentlichen Gewissens ergeben (AP I, Art. 1.2).<br>• Die Würde feindlicher Kombattanten und aller Zivilisten ist zu achten.<br>• Von einem Feind begangene Verletzungen der Würde rechtfertigen keine Repressalien. |

**TAB 7. Auflistung der aus dem Völkerrecht entlehnten *betroffene Gruppen*, die von einer technischen Fehlfunktion oder einer Normverletzung betroffen sein können.**

| Betroffene Gruppen |
|---|
| Eigene Einheiten |
| Alliierte Einheiten |
| Neutrale Einheiten |
| Feindliche Einheiten |
| Hors de Combat (Außer Gefecht) |
| Zivilisten |
| Speziell geschützte Einheiten |

einem ML-basiertem Objektklassifikationsalgorithmus. Die Signalkette ist in Bild 3 dargestellt. Das Gesamtsystem besteht aus mehreren Funktionsebenen, in diesem Fall der Sensorebene, d.h. dem Kamerasystem auf dem die Bilder physikalisch registriert werden, dem Verarbeitungssystem in Form eines DSPs sowie der Drohne als fliegender Plattform. Jede dieser Ebenen übernimmt eine bestimmte Funktion im Gesamtsystem und auf jeder dieser Ebenen könnte ein dedizierter ML-Algorithmus arbeiten und Funktionen bereitstellen. Auf der Sensorebene könnte bspw. eine ML-basierte Rauschunterdrückung ausgeführt werden. Auf der Verarbeitungsebene arbeitet ein Objektklassifikationsalgorithmus und auf der Plattformebene könnte eine ML-gestützte Flugpfadoptimierung bspw. eine optimale Flughöhe für gut aufgelöste Kamerabilder bestimmen. Je nach Trainingsgrad und -güte ergibt sich ein besseres oder schlechteres Ergebnis im jeweiligen Prozessschritt. Für dieses Anwendungsbeispiel nehmen wir an, dass nur eine ML-basierte Objektklassifikation auf der Verarbeitungsebene stattfindet.

Die Aufgabenstellung für den Algorithmus besteht darin, Fahrzeuge in drei mögliche Klassen einzuteilen: (1) URAL LKW, (2) URAL mit Schutzsymbol (3) BMP-1 Schützenpanzer (vgl. Bild 2). Im Allgemeinen läuft ein Klassifikationsprozess folgendermaßen ab: Zuerst muss ein Objekt detektiert werden, d.h. es muss als etwas erkannt werden, dass nicht zur Umgebung gehört. Dies geschieht in der Detektionsphase. An dieser Stelle hat man noch keinerlei Informationen über das Objekt. In der Klassifikationsphase wird anhand von bestimmten visuellen Eigenschaften (Räder oder Ketten, Form, etc.) das Objekt einer Klasse zugewiesen (LKW, Panzer, Soldat, etc). Bekommt man schließlich ein noch klareres Bild des Objektes, kann dieses schließlich identifiziert und somit einem konkrete Modell/Typ zugewiesen werden. Dieser Prozess ist in Bild 4 illustriert. Wie groß ein Objekt im Bild zu sehen sein muss, um detektiert, klassifiziert bzw. identifiziert zu werden, lässt sich mittels der Johnson Kriterien bestimmen [37].

©2026



(a) URAL-43206, ein weit verbreiteter Logistik-LKW [34].

(b) URAL-43206 als Sanitätsfahrzeug mit dem vorgeschriebenen Schutzsymbol nach Artikel 38 der Genfer Konvention (I) von 1949 [19] [35].

(c) Schützenpanzer BMP-1 als Beispiel für ein bewaffnetes Fahrzeug mit Kettenantrieb, im Gegensatz zum Radantrieb des URAL [36].

**BILD 2. Übersicht über die drei möglichen Fahrzeugklassen, in die der Algorithmus Objekte einklassifizieren soll. Dies soll insbesondere Unterschiede zwischen legitimen und nicht-legitimen Zielen sowie unterschiedliches Gefährdungspotential aufzeigen ([34], bearbeitet).**

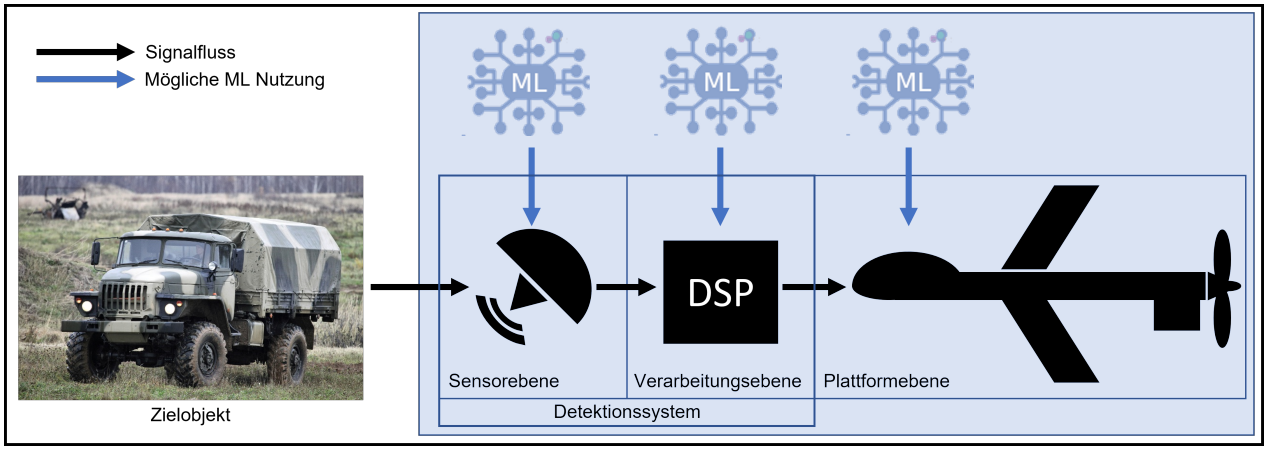



**BILD 3. Signalfluss und mögliche Einsatzebenen von ML-Algorithmen. In diesem Beispiel nimmt ein Kamerasensor das gesucht Objekt auf, leitet das Signal an einen DSP weiter. Beide Systeme bilden das Detektionssystem, dass in die Drohnenplattform integriert ist. Auf jeder dieser drei funktionalen Ebenen kann ML eingesetzt werden. In diesem Beispiel beschränken wir uns auf die Verarbeitungsebene.**

Auch wenn diese Klassifikation für einen Menschen trivial scheint, gibt es einige mögliche Fehlerquellen bei der automatisierten Klassifikation, die wir mit der EHA identifizieren wollen. Es wird angenommen, dass das Kamerasystem der Drohne fehlerfrei funktioniert, d.h. die Eingangsdaten für den Algorithmus kommen nicht als Fehlerquelle in Betracht. In einer vollständigen FHA müsste dies als mögliche Fehlerquelle jedoch mitbetrachtet werden.

Wir wollen nun das zuvor eingeführte EHA auf dieses Beispiel anwenden. Hierfür werden die in Anhang A gegebenen Definitionen für die ethischen und technischen Auswirkungen pro betroffener Klasse benötigt. Diese sind dort für alle im humanitären Völkerrecht definierten Klassen angegeben, die in Tab. 7 aufgeführt sind. Anschließend müssen alle Funktionen des Systems definiert werden, und darauf aufbauend alle möglichen Fehlfunktionen aufgelistet werden. Die gewünschten Funktionen sind in diesem Fall: (1) Feindlichen URAL-LKW korrekt klassifizieren, (2) Feindlichen URAL-LKW mit Schutzsymbol korrekt klassifizieren, (3) Schützenpanzer BMP-1 korrekt klassifizieren. Die möglichen Fehlfunktionen ergeben sich als Klassifikationen der jeweils zwei anderen Klassen. Nach der Definition der Fehlfunktionen muss jede Fehlfunktion ethisch und technisch bewertet werden. Dies geschieht unter Zuhilfenahme der Definitionen aus Anhang A sowie der Betrachtung des Worst-Case Szenarios. Beispielsweise kann das Nicht-Angreifen eines feindlichen Logistik LKW nachträglich negative Folgen für eigene Einheiten haben, da Lieferungen dieses LKWs den Gegner stärken. Ein weiteres aufschlussreiches Beispiel ist der Fall der Fehlklassifikation eines LKWs als ein Schützenpanzer. Dies könnte zur Folge haben, dass der LKW mit einer panzerbrechenden Waffe angegriffen wird, was einer Verschwendung von Munition gleichkommt und somit einen relativ hohen finanziellen Schaden nach sich zieht. Auf diese Art und Weise werden alle Fälle betrachtet und hinsichtlich ihrer Kritikalität notwendigen Zuverlässigkeitsklassen der Algorithmen zugewiesen. Die notwendige Klasse für einen Algorithmus, der die Funktionen des Systems bereitstellt wird schließlich durch die anspruchsvollste Klasse gegeben, die bei wenigstens einer der Fehlfunktionen benötigt wird. Die vollständige Bewertungsmatrix ist in Tab. 8 dargestellt.

©2026




**BILD 4. Identifikationsprozess eines Fahrzeugs mittels der Schritte Detektion, Klassifikation und Identifikation. In der Detektionsphase wird ein Objekt erkannt, in der Klassifikationsphase wird es anhand von Merkmalen einer Klasse zugewiesen, z.B. Mensch, LKW, Panzer. In der Identifikationsphase wird das konkrete Objekt je nach Wissensstand bestmöglich identifiziert ([34], bearbeitet).**

## 7. ZUSAMMENFASSUNG

In dieser Arbeit wurde eine Erweiterung der weit verbreiteten Funktionalen Gefahrenanalyse vorgestellt, die sogenannte Ethische Gefahrenanalyse. Diese erweitert die FHA um zwei Aspekte: (1) Die Bewertung von ML-Algorithmen hinsichtlich ihren Zuverlässigkeit zum Zweck des Einsatzes in sicherheitskritischen Umgebungen. Dies baut auf Empfehlungen aus der DIN 5469 auf, um Kriterien zur Evaluation von verschiedenen Arten von Machine-Learning Algorithmen abzuleiten. Klassen zur Unterscheidung von Algorithmen hinsichtlich ihrer erwartbaren Zuverlässigkeit zur funktionalen Sicherheit wurden definiert. Für diese müssen in Zukunft genauere Abgrenzungskriterien definiert werden, um eine möglichst genaue und sichere Zuordnung zu den Fehlfunktionen vornehmen zu können.
Zweitens wurde eine ethische Dimension zum FHA hinzugefügt, welche die Analyse um normative Auswirkungen erweitert und sogenannte *Betroffene Gruppe* als relevanten Faktor hinzufügt. Da aktuell nur für technische Fehlfunktionen Klassifikationsschemata zur Feststellung des Schweregrades existieren (siehe bspw. MIL-STD-882E [15]), wurden neue Definitionen entwickelt, die zur Bewertung des Schweregrads von *normativen* Auswirkung dienen und auf dem internationalen humanitären Völkerrecht beruhen. Diese sind jedoch vorläufig und stehen zur weiteren Diskussion. Mithilfe dieser Definitionen wurde ein Anwendungsbeispiel aufgezeigt, dass in der Lage ist, normative und technische Auswirkungen von Fehlfunktionen in einem einfachen ML-basiertem Objektklassifikationssystem zu bewerten und notwendige Zuverlässigkeitsanforderungen zu begründen. Dies geschieht schlussendlich durch die Kombination der genannten Definitionen und den oben genannten Zuverlässigkeitsklassen für ML-Algorithmen.

Ein Punkt, der in dieser Arbeit nicht behandelt wurde, allerdings eine hohe Relevanz für die funktionale Sicherheit von ML-Algorithmen hat, ist die Datenqualität der einzelnen Systeme, die einen kaum zu überschätzenden Einfluss auf die Leistung des fertigen Systems besitzt, ähnlich der menschlichen Erfahrung in einem manuellen Entscheidungsfindungsprozess. Die Trainingsdatenqualität sollte deshalb in Zukunft als ein weiterer Faktor in die Zuverlässigkeitsbewertung der Algorithmen einfließen. Er wurde hier fortgelassen, da die Datenqualität eine höchst kontext- und systemspezifische Größe ist, und somit nicht auf Algorithmusebene bewertet werden kann.
Die exakten Kriterien zur Abgrenzung der Schwere von Fehlfunktionen sind noch nicht festgelegt, da dies eine Aufgabe mit großer gesellschaftlicher Tragweite ist und nicht vorschnell angegangen werden sollte. Diese Arbeit soll als Stein des Anstoßes dienen und Diskussionen über die Entwicklung einer umfangreicheren Funktionalen Gefahrenanalyse voranbringen, damit ML-Algorithmen zuverlässig in sicherheitskritischen Umgebungen angewendet werden können.

## 8. FÖRDERHINWEIS

Das Projekt „MilEth“ (Militärische Verteidigungstechnologien und Ethik) wird vom Bundesministerium für Forschung, Technologie und Raumfahrt gefördert (01UG2409A). Die Autoren sind für den Inhalt dieser Veröffentlichung verantwortlich.

**Kontaktadresse:**

hauke.budig@tuhh.com

Gefördert durch:

Bundesministerium
für Forschung, Technologie
und Raumfahrt

©2026

**TAB 8. Analyse des Beispielszenarios mittels der Ethischen Gefahrenanalyse. Von links nach rechts sind die intendierten Funktionen, die Fehlfunktionen sowie deren ethische und technische Auswirkungen inklusive Begründungen aufgeführt. Nach der Schweregradbewertung ist es hier möglich, eine Zuordnung zu den Zuverlässigkeitsklassen der ML-Algorithmen vorzunehmen, um zu bestimmen, welche Algorithmen geeignet sein könnten, das System funktional sicher zu machen.**

| Funktion | ID | Fehlerfall | Ethische Klasse | Technische Klasse | Ethische Begründung | Technische Begründung | Algorithmusklasse |
|---|---|---|---|---|---|---|---|
| **Feindliches URAL mit Schutzsymbol korrekt klassifizieren** | 1 | Klassifiziert als feindlicher, militärischer URAL | Katastrophal | Katastrophal | Angriff auf völkerrechtlich geschützte Einheit | Verursacht ungerechtfertigten, hohen Schaden bei Mensch und Material | I |
| | 2 | Klassifiziert als feindlicher Schützenpanzer | Katastrophal | Katastrophal | Angriff auf völkerrechtlich geschützte Einheit | Verursacht ungerechtfertigten, hohen Schaden bei Mensch und Material | I |
| **Feindliches URAL korrekt klassifizieren** | 3 | Klassifiziert als feindlicher URAL mit Schutzsymbol | Marginal | Unerheblich | Einheit wird nicht angegriffen, kann also den Feind logistisch unterstützen, was negative Auswirkungen auf unsere Einheiten hat. | Keine Auswirkung | II-III |
| | 4 | Klassifiziert als feindlicher Schützenpanzer | Gering | Marginal | Einheit wird rechtmäßig angegriffen, aber Ausrüstung sollte zuverlässig funktionieren. Durch Fehlfunktionen kann das Vertrauen in das System leiden | Verwendung falscher Munition, d.h. hoher finanzieller Verlust | II-III |
| **Feindlicher Schützenpanzer korrekt klassifizieren** | 5 | Klassifiziert als feindlicher, militärischer URAL | Katastrophal | Katastrophal | Bei Angriff mit falscher Munition: Schutz der eigenen Truppen wird nicht gewährleistet, feindliche Einheit könnte verbündete Truppen angreifen und beliebig viel Schaden anrichten. | Angriff mit falscher Munition wahrscheinlich ohne Schaden am Schützenpanzer. Damit kommen eigene Einheiten in Gefahr. | I |
| | 6 | Klassifiziert als feindliches URAL mit Schutzsymbol | Katastrophal | Katastrophal | Schutz der eigenen Truppe wird nicht gewährleistet, feindliche Einheit könnte verbündete Truppen angreifen und beliebig viel Schaden anrichten. | Hoher Verlust an Mensch und Material, falls Schützenpanzer angreift. | I |
| | | | | | | **Gesamtbewertung:** | I |

## Literatur


[1] Society of Automotive Engineers, Inc. Guidelines for Development of Civil Aircraft and Systems. Recommended Practice SAE ARP4754, November 1996.

[2] Society of Automotive Engineers, Inc. Guidelines and Methods for Conducting the Safety Assessment Process on Civli Airborne Systems and Equipment. Recommended Practice SAE ARP4761, Dezember 1996.

[3] Nathan Gabriel Wood, Scott Robbins, Eduardo Zegarra Berodt, Anton Graf von Westerholt, Michelle Behrndt, Hauke Budig, and Daniel Kloock-Schreiber. Stop Saying "AI". *Philosophy & Technology*, 39(2):106, Juni 2026. DOI: 10.1007/s13347-026-01114-4.

[4] SAE International. Artificial Intelligence in Aeronautical Systems: Statement of Concerns. Technical Report SAE AIR6988, April 2021.

[5] Aerospace Vehicle Systems Institute. AFE 87 - Machine Learning Final Report. Final Report AFE87, College Station, TX, United States, 2020.

[6] Andrew P Williams. Defining Autonomy in Systems: Challenges and Solutions. In *Autonomous Systems - Issues for Defence Policymakers*, pages 27–62. NATO Communications and Information Agency, The Hague, Netherlands, 2015.

[7] Jean-François Caron. Defining semi-autonomous, automated and autonomous weapon systems in order to understand their ethical challenges. *Digital War*, 1(1):173–177, Dezember 2020. DOI: 10.1057/s42984-020-00028-5.

[8] Nathan Gabriel Wood. Autonomous Weapon Systems: A Clarification. *Journal of Military Ethics*, 22(1):18–32, Januar 2023. DOI: 10.1080/15027570.2023.2214402.

[9] Center for War Studies. The AutoPractices Project. Project Report, Odense, September 2025.

[10] The Global Commission on Responsible Artificial Intelligence in the Military Domain and Mary O'Connell. Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain. September 2025.


©2026

[11] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, Juni 2020. DOI: 10.1016/j.inffus.2019.12.012.

[12] RTCA Radio Technical Commission for Aeronautics, Inc. Software Considerations in Airborne Systems and Equipment Certification. Technical Report RTCA DO-178C, Dezember 2011.

[13] DIN Deutsches Institut für Normung. Funktionale Sicherheit sicherheitsbezogener elektrischer/elektronischer/programmierbarer elektronischer Systeme. Standard DIN EN 61508, 2011. Normenreihe, Teile 1–7.

[14] DIN Deutsches Institut für Normung. Künstliche Intelligenz -Funktionale Sicherheit und KI-Systeme. Standard DIN ISO/IEC/TR 5469 (VDE 0803-5469):2025-07, Juli 2025.

[15] U.S. Department of Defense (DoD). Department of Defense Standard System Safety. Standard Practice MIL-STD-882 Revision E Change 1, September 2023.

[16] DIN Deutsches Institut für Normung. Künstliche Intelligenz - Datenqualität für Analytik und maschinelles Lernen (ML). Standard DIN EN ISO/IEC 5259, März 2025. Normenreihe, Teile 1–4.

[17] ISO International Organization for Standardization. Artificial Intelligence (AI) - Assessment of the robustness of neural networks. Technical Report ISO/IEC TR 24029, März 2021. Normenreihe, Teile 1–3.

[18] International Standardization Organization. Information technololgy - Artificial Intelligence - Overview of Trustworthiness in Artificial Intelligence. Technical Report ISO/IEC TR 24028:2020, Mai 2020.

[19] Geneva Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field of 12 August 1949, August 1949.

[20] Geneva Convention (II) for the Amelioration of the Condition of Wounded,Sick and Shipwrecked Members of Armed Forces at Sea of 12 August 1949, August 1949.

[21] Geneva Convention (III) Relative to the Treatment of Prisoners of War of 12 August 1949, August 1949.

[22] Geneva Convention (IV) Relative to the Protection of Civilian Persons in Time of War of 12 August 1949, August 1949.

[23] Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977.

[24] Strafgesetzbuch. In der Fassung der Bekanntmachung vom 13. November 1998 (BGBl. I S. 3322), zuletzt geändert durch Artikel 1 des Gesetzes vom 20. März 2026 (BGBl. 2026 I Nr. 95).

[25] Grundgesetz für die Bundesrepublik Deutschland. In der im Bundesgesetzblatt Teil III, Gliederungsnummer 100-1, veröffentlichten bereinigten Fassung, zuletzt geändert durch Artikel 1 des Gesetzes vom 22. März 2025 (BGBl. 2025 I Nr. 94).

[26] Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, November 1983. DOI: 10.1016/0005-1098(83)90046-8.

[27] Auste Simkute, Lev Tankelevitch, Viktor Kewenig, Ava Elizabeth Scott, Abigail Sellen, and Sean Rintel. Ironies of Generative AI: Understanding and Mitigating Productivity Loss in Human-AI Interaction. *International Journal of Human–Computer Interaction*, 41(5):2898–2919, März 2025. DOI: 10.1080/10447318.2024.2405782.

[28] Winfred Arthur Jr., Winston Bennett Jr., Pamela L. Stanush, and Theresa L. McNelly. Factors That Influence Skill Decay and Retention: A Quantitative Review and Analysis. *Human Performance*, 11(1):57–101, März 1998. DOI: 10.1207/s15327043hup1101_3.

[29] Marina Klostermann, Stephanie Conein, Thomas Felkl, and Annette Kluge. Factors Influencing Attenuating Skill Decay in High-Risk Industries: A Scoping Review. *Safety*, 8(2):22, Juni 2022. DOI: 10.3390/safety8020022.

[30] Tapani Rinta-Kahila, The University of Queensland Business School, Esko Penttinen, Aalto University School of Business, Antti Salovaara, Aalto University / University of Helsinki, Wael Soliman, University of Agder, Norway / University of Jyväskylä, Joona Ruissalo, and Aalto University School of Business. The Vicious Circles of Skill Erosion: A Case Study of Cognitive Automation. *Journal of the Association for Information Systems*, 24(5):1378–1412, 2023. DOI: 10.17705/1jais.00829.

[31] Michael Gerlich. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, 15(1):6, Januar 2025. DOI: 10.3390/soc15010006.

©2026

[32] Tae Woo Kim, Umair Usman, Aaron Garvey, and Adam Duhachek. From algorithm aversion to AI dependence: Deskilling, upskilling, and emerging addictions in the GenAI age. *Consumer Psychology Review*, 9(1):142–164, 2026. DOI: 10.1002/arcp.70008.

[33] Shannon Vallor. Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. *Philosophy & Technology*, 28(1):107–124, März 2015. DOI: 10.1007/s13347-014-0156-9.

[34] Vitaly V. Kuzmin. Ural-43206, Oktober 2012. CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0, via Wikimedia Commons.

[35] Vitaly V. Kuzmin. Ambulance vehicle AS-4350.1. Designed for evacuation and transportation of the wounded and sick on stretchers and sitting. Equipped with stretcher holders., 17 August 2013, 18:36:26. CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0, via Wikimedia Commons.

[36] Don S. Montgomery. A left front view of a Soviet BMP-1 mechanized infantry vehicle. The vehicle is on display at Bolling Air Force Base., 21. Januar 2009, 13:41. Public domain, via Wikimedia Commons.

[37] John Johnson. Analysis of Image Forming Systems. In *Image Intensifier Symposium*, pages 249–274, Fort Belvoir, VA, USA, Oktober 1958.

©2026

## A. ANHANG: DEFINITIONEN ZUR ETHISCHEN GEFAHRENANALYSE

In diesem Anhang sind die genutzten Definitionen für die Unterscheidung des Schweregrads einer technische oder ethischen Auswirkung einer Fehlfunktion zu finden. Diese sind nach betroffenen Gruppen im Sinne des Völkerrechts sortiert. Es sei darauf verwiesen, dass diese nur beispielhafte Definitionen darstellen und keineswegs einen Vorschlag, der von den Autoren zur Nutzung empfohlen wird. Die Erarbeitung passender Definitionen muss zu gegebener Zeit von den zuständigen Stellen durchgeführt werden. Diese Definition basieren weitestgehend auf die Regeln des internationalen humanitären Völkerrechtes und auf den Prinzipien der Unterscheidung, der Proportionalität, der Notwendigkeit und der Menschlichkeit.

Die Definitionen für die technischen Auswirkungen sind größtenteils aus dem MIL-STD-882E [15] entnommen, da dieser viele verschiedene Aspekte (Verletzung, Umweltschaden, finanzieller Verlust) abdeckt. Es wurde allerdings eine Klasse ”Gering” hinzugefügt, da es sich insbesondere für die Unterscheidung ethischer Auswirkungen anbot. Deshalb wurde die Höhe des finanziellen Schadens für die Klassen ”Gering” und ”Vernachlässigbar” angepasst.

©2026

**TAB 9. Bewertungsdefinitionen zur Ethischen Gefahrenanalyse für Angehörige verschiedener militärischer Konfliktteilnehmer, Teil I**

| Schweregrad | Auswirkungen auf Kombattanten (Eigene, Alliierte, Neutrale) | | Auswirkungen auf feindliche Einheiten | |
|---|---|---|---|---|
| **Effekt** | **Technisch** | **Normativ** | **Technisch** | **Normativ** |
| **Unerheblich/ Beabsichtigt** | Könnte in einem der Folgenden resultieren: Verletzung oder berufsbedingte Krankheit ohne einen verlorenen Arbeitstag, minimaler Umweltschaden oder finanzieller Verlust geringer als $5K. | Keine oder sehr abstrakte Möglichkeit der Verletzung einer ethischen, sozialen oder anderen Norm, und kein Potential das Vertrauen in eine solche Norm zu mindern oder die Hemmschwelle zu senken, eine solche Norm zu brechen. | Legitime Gewalt gegen Kombattanten im Sinne des IVR, beabsichtigt oder unbeabsichtigt, welche zu Schäden bzw. Verletzung an/von legitimen Zielen (Kombattanten, Einheiten, Gebäuden o.Ä.) führt. | Legitime Gewalt gegen Kombattanten im Sinne des IVR, solange die Einsatzmittel nach dem IVR konventionell und in ihren Auswirkungen nicht wahllos sind. |
| **Gering** | Könnte in einem oder mehreren der Folgenden resultieren: Verletzung oder berufsbedingte Krankheit ohne einen verlorenen Arbeitstag, minimaler Umweltschaden oder finanzieller Verlust geringer als $100K. | Keine direkte Verletzung einer ethischen, sozialer oder anderen Norm, allerdings hat die Auswirkung das Potential, das Vertrauen in Normen zu verringern und subjektiv die Hemmschwelle zu senken, eine solche Norm zu brechen. | Ungerechtfertigte Auswirkung im Sinne des IVR, welche zu einem oder mehreren der Folgenden führen könnte: Verletzung oder berufsbedingte Krankheit ohne einen verlorenen Arbeitstag, minimaler Umweltschaden, oder finanzieller Verlust geringer als $100K. | Keine direkte Verletzung einer ethischen, sozialer oder anderen Norm, allerdings hat die Auswirkung das Potential, das Vertrauen in Normen zu verringern und subjektiv die Hemmschwelle zu senken, eine solche Norm zu brechen. |
| **Marginal** | Könnte in einem oder mehreren der Folgenden resultieren: Verletzung oder berufsbedingte Krankheit verbunden mit einem oder mehreren verlorenen Arbeitstagen, umkehrbarer, moderater Umweltschaden oder finanzieller Verlust über $100K, aber geringer als $1M. | Offensichtliche Verletzung einer ethischen, sozialen oder anderen Norm ohne direkte negative Konsequenzen für eine bestimmte Person, allerdings auf Prozesse und die Durchführung der Operation. | Ungerechtfertigte Auswirkung im Sinne des IVR, welche in einem oder mehreren der Folgenden resultieren könnte: Verletzung oder berufsbedingte Krankheit verbunden mit einem oder mehreren Arbeitstagen, umkehrbarer, moderater Umweltschaden oder finanzieller Verlust über $100K, aber geringer als $1M. | N/A |
| **Kritisch** | Könnte in einem oder mehreren der Folgenden resultieren: Permanente, teilweise Behinderung oder berufsbedingte Krankheit und Krankenhausaufenthalt von mindestens drei Personen, umkehrbarer signifikanter Umweltschaden oder finanzieller Verlust über $1M aber geringer als $10M. | Direkte und offensichtliche Verletzung einer ethischen, sozialen, oder anderen Norm mit direkten, negativen Konsequenzen auf eine Person, die durch diese Norm verhindert werden sollte. | Ungerechtfertigte Auswirkung im Sinne des IVR, welche in einem oder mehreren der Folgenden resultieren könnte: Permanente, teilweise Behinderung oder berufsbedingte Krankheit und Krankenhausaufenthalt von mindestens drei Personen, umkehrbarer signifikanter Umweltschaden oder finanzieller Verlust über $1M aber geringer als $10M. | Direkte und offensichtlicher Verletzung einer ethischen, sozialen, oder anderen Norm mit direkten negativen Auswirkungen auf eine Person durch nicht legitime Auswirkungen (bspw. verbotene Waffen, überflüssige Gewalt, Niederträchtigkeit, o.Ä.), die durch diese Norm verhindert werden sollte. |
| **Katastrophal** | Könnte in einem oder mehreren der Folgenden resultieren: Tod, permanente vollständige Behinderung, irreversibler signifikanter Umweltschaden oder finanzieller Verlust gleich oder über $10M. | Direkte und offensichtliche Verletzung einer ethischen, sozialen oder anderen Norm mit direkten negativer Auswirkung auf mehr als eine Person oder auf eines oder mehrere Objekte, die durch diese Norm geschützt werden sollten. | Ungerechtfertigte Auswirkung im Sinne des IVR, welches in einem oder mehreren der Folgenden resultieren könnte: Tod, permanente vollständige Behinderung, irreversibler signifikanter Umweltschaden oder finanzieller Verlust gleich oder über $10M. | Direkte und offensichtliche Verletzung einer ethischen, sozialen oder anderen Norm mit direkten negativen Auswirkungen auf mehr als eine Person oder auf ein oder mehrere Objekte auf nicht legitime Art und Weise (z.B, durch verbotene Waffen, überflüssige Gewalt, Niederträchtigkeit, o.Ä.) die durch diese Norm verhindert werden sollte. |

©2026

**TAB 10. Bewertungsdefinitionen zur Ethischen Gefahrenanalyse für Angehörige verschiedener militärischer Konfliktteilnehmer, Teil II**

| Schweregrad | Auswirkungen auf außer Gefecht gesetzte Einheiten (Kriegsgefangene, Schiffbrüchige, Kranke oder Verwundete) | | Auswirkungen auf Zivilisten | |
|---|---|---|---|---|
| **Effekt** | **Technisch** | **Normativ** | **Technisch** | **Normativ** |
| **Unerheblich/ Beabsichtigt** | Könnte in einem der Folgenden resultieren: Verletzung oder berufsbedingte Krankheit ohne einen verlorenen Arbeitstag, minimaler Umweltschaden oder finanzieller Verlust geringer als $5K. | Keine oder sehr abstrakte Möglichkeit der Verletzung einer ethischen, sozialen oder anderen Norm, und kein Potential das Vertrauen in eine solche Norm zu mindern oder die Hemmschwelle zu senken, eine solche Norm zu brechen. | Könnte in einem der Folgenden resultieren: Verletzung oder berufsbedingte Krankheit ohne einen verlorenen Arbeitstag, minimaler Umweltschaden oder finanzieller Verlust geringer als $5K. | Keine oder sehr abstrakte Möglichkeit der Verletzung einer ethischen, sozialen oder anderen Norm, und kein Potential das Vertrauen in eine solche Norm zu mindern oder die Hemmschwelle zu senken, eine solche Norm zu brechen. |
| **Gering** | Könnte in einem oder mehreren der Folgenden resultieren: Verletzung oder berufsbedingte Krankheit ohne einen verlorenen Arbeitstag, minimaler Umweltschaden oder finanzieller Verlust geringer als $100K. | N/A | Könnte in einem oder mehreren der Folgenden resultieren: Verletzung oder berufsbedingte Krankheit ohne einen verlorenen Arbeitstag, minimaler Umweltschaden oder finanzieller Verlust geringer als $100K. | Keine direkte Verletzung einer ethischen, sozialer oder anderen Norm, allerdings hat die Auswirkung das Potential, das Vertrauen in Normen zu verringern und subjektiv die Hemmschwelle zu senken, eine solche Norm zu brechen. |
| **Marginal** | Könnte in einem oder mehreren der Folgenden resultieren: Verletzung oder berufsbedingte Krankheit verbunden mit einem oder mehreren Arbeitstagen, umkehrbarer, moderater Umweltschaden oder finanzieller Verlust über $100K, aber geringer als $1M. | Keine direkte Verletzung einer ethischen, sozialer oder anderen Norm, allerdings hat die Auswirkung das Potential, das Vertrauen in Normen zu verringern und subjektiv die Hemmschwelle zu senken, eine solche Norm zu brechen. | Könnte in einem oder mehreren der Folgenden resultieren: Verletzung oder berufsbedingte Krankheit verbunden mit einem oder mehreren Arbeitstagen, umkehrbarer, moderater Umweltschaden oder finanzieller Verlust über $100K, aber geringer als $1M. | Offensichtliche Verletzung einer ethischen, sozialen oder anderen Norm ohne direkte negative Konsequenzen für eine bestimmte Person, allerdings auf Prozesse und die Durchführung der Operation. |
| **Kritisch** | Könnte in einem oder mehreren der Folgenden resultieren: Permanente, teilweise Behinderung oder berufsbedingte Krankheit und Krankenhausaufenthalt von mindestens drei Personen, umkehrbarer signifikanter Umweltschaden oder finanzieller Verlust über $1M aber geringer als $10M. | Offensichtliche Verletzung einer ethischen, sozialen oder anderen Norm ohne direkte negative Konsequenzen für eine bestimmte Person, allerdings auf Prozesse und die Durchführung der Operation. | Könnte in einem oder mehreren der Folgenden resultieren: Permanente, teilweise Behinderung oder berufsbedingte Krankheit und Krankenhausaufenthalt von mindestens drei Personen, umkehrbarer signifikanter Umweltschaden oder finanzieller Verlust über $1M aber geringer als $10M. | Direkte und offensichtliche Verletzung einer ethischen, sozialen oder anderen Norm mit direkten negativen Auswirkungen auf proportional viele Personen, die durch diese Norm verhindert werden sollten. |
| **Katastrophal** | Könnte in einem oder mehreren der Folgenden resultieren: Tod, permanente vollständige Behinderung, irreversibler signifikanter Umweltschaden oder finanzieller Verlust gleich oder über $10M. | Direkte und offensichtliche Verletzung einer ethischen, sozialen oder anderen Norm mit direkten negativer Auswirkung auf eine oder mehrere Personen die durch diese Norm geschützt werden sollten. | Könnte in einem oder mehreren der Folgenden resultieren: Tod, permanente vollständige Behinderung, irreversibler signifikanter Umweltschaden oder finanzieller Verlust gleich oder über $10M. | Direkte und offensichtliche Verletzung einer ethischen, sozialen oder anderen Norm mit direkten negativen Auswirkungen auf überproportional viele Personen oder ein oder mehrere Objekte, die kein legitimes Ziel sind, die durch diese Norm verhindert werden sollten. |

©2026

**TAB 11. Bewertungsdefinitionen zur Ethischen Gefahrenanalyse für Angehörige verschiedener militärischer Konfliktteilnehmer, Teil III**

| Schweregrad | Effekte auf besonders geschützte Einheiten (Medizinisches Personal, inkl. militärisch-medizinisches Personal, Rotes Kreuz und andere humanitäre Organisationen) | |
|---|---|---|
| **Effekt** | **Technisch** | **Normativ** |
| **Unerheblich/ Beabsichtigt** | Könnte in einem der Folgenden resultieren: Verletzung oder berufsbedingte Krankheit ohne einen verlorenen Arbeitstag, minimaler Umweltschaden oder finanzieller Verlust geringer als $5K. | Keine oder sehr abstrakte Möglichkeit der Verletzung einer ethischen, sozialen oder anderen Norm, und kein Potential das Vertrauen in eine solche Norm zu mindern oder die Hemmschwelle zu senken, eine solche Norm zu brechen. |
| **Gering** | Könnte in einem oder mehreren der Folgenden resultieren: Verletzung oder berufsbedingte Krankheit ohne einen verlorenen Arbeitstag, minimaler Umweltschaden oder finanzieller Verlust geringer als $100K. | N/A |
| **Marginal** | Könnte in einem oder mehreren der Folgenden resultieren: Verletzung oder berufsbedingte Krankheit verbunden mit einem oder mehreren Arbeitstagen, umkehrbarer, moderater Umweltschaden oder finanzieller Verlust über $100K, aber geringer als $1M. | Keine direkte Verletzung einer ethischen, sozialer oder anderen Norm, allerdings hat die Auswirkung das Potential, das Vertrauen in Normen zu verringern und subjektiv die Hemmschwelle zu senken, eine solche Norm zu brechen. |
| **Kritisch** | Könnte in einem oder mehreren der Folgenden resultieren: Permanente, teilweise Behinderung oder berufsbedingte Krankheit und Krankenhausaufenthalt von mindestens drei Personen, umkehrbarer signifikanter Umweltschaden oder finanzieller Verlust über $1M aber geringer als $10M. | Offensichtliche Verletzung einer ethischen, sozialen oder anderen Norm ohne direkte negative Konsequenzen für eine bestimmte Person, allerdings auf Prozesse und die Durchführung der Operation. |
| **Katastrophal** | Könnte in einem oder mehreren der Folgenden resultieren: Tod, permanente vollständige Behinderung, irreversibler signifikanter Umweltschaden oder finanzieller Verlust gleich oder über $10M. | Direkte und offensichtliche Verletzung einer ethischen, sozialen, oder anderen Norm mit direkten, negativen Konsequenzen auf eine oder mehrere Personen oder Objekte (insbesondere Krankenhäuser und andere medizinische Einrichtungen), die durch diese Norm verhindert werden sollte. |

©2026