ASSER INSTITUTE
Centre for International & European Law

ELSA LAB DEFENCE

Toolkit for Evaluating the Legal Acceptability of Military AI Systems

Jonathan Kwik and Marta Bo

Why use this Toolkit?

Artificial intelligence (AI) is increasingly incorporated into a wide range of military systems and capabilities. Their functions may include analysing information, supporting human decision-making, navigation, object recognition and classification, and functions within weapon systems. As these systems move from development to deployment, questions of legal acceptability arise throughout their lifecycle—not only at the point of use. **Assessing these questions early and repeatedly matters.** Decisions made during development, testing, acquisition and configuration can shape how a system ultimately performs and the risks it presents during military operations. Equally, changes in the system, its intended use, or the operational environment may require earlier assessments to be revisited.

This **Toolkit for Assessing the Legal Acceptability of Military AI Systems** provides a structured method for identifying and assessing factors that may affect the legal acceptability of AI-enabled military systems under international humanitarian law (IHL). It is intended to support those involved in the development, acquisition, review, deployment, or use of military AI systems, including persons from the Armed Forces, Ministry of Defence, and Industry.

> * This Toolkit accompanies the report *From Development to Deployment: A Framework for Evaluating the Legal Acceptability of Military AI Systems*. The report sets out the legal and methodological foundations of the framework, the rationale for the variables included in the Toolkit, and its intended applications. This Toolkit provides the practical component of that framework, by guiding users through a structured, context-sensitive assessment of military AI systems.

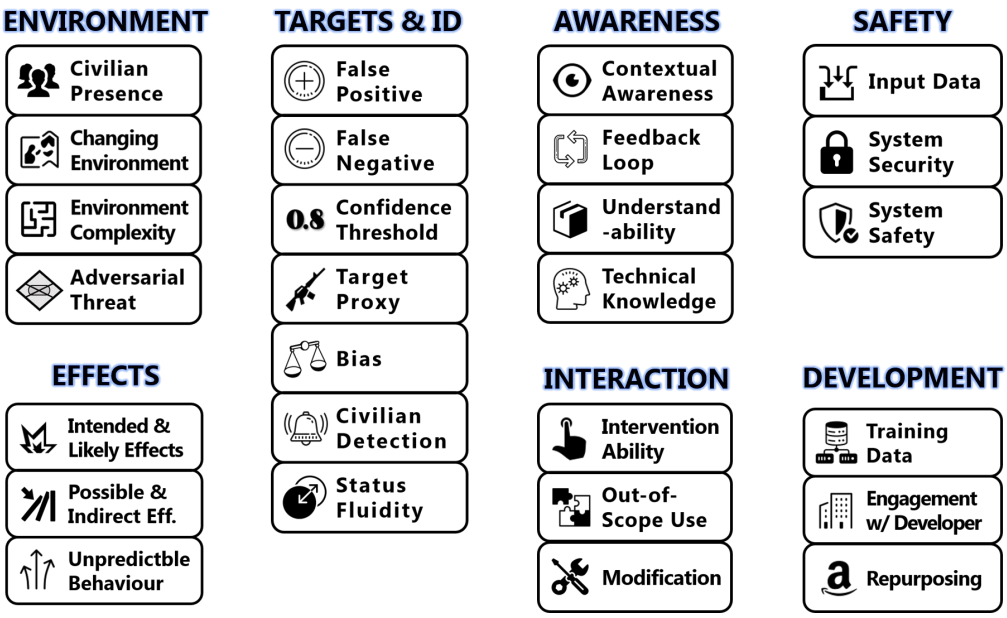
Who should use this Toolkit?

The Toolkit is intended for **legal advisers, military personnel, policymakers, procurement and acquisition teams, developers, industry personnel and others** involved in the **development, review, acquisition, deployment, or use** of military AI systems. It can support assessment throughout a system's lifecycle—from development and acquisition to deployment and operational use. Although focused on legal acceptability under IHL, its effective use may require input from legal, operational, and technical personnel.

The Toolkit is designed to assess a **specific AI system in relation to an intended or proposed operational context**. Because legal acceptability under IHL is inherently context-dependent, the concrete risk evaluation of an AI system may be different depending on how, where, and for what purpose it is intended to be used. Rather than producing an automated pass/fail result, this Toolkit supports structured legal reasoning and informed decision-making. It helps users identify relevant sources of risk, areas requiring further information or safeguards, and factors relevant to determining whether and under what conditions a system may be used in the proposed context.

How does the Toolkit work?

The Toolkit is built around **27 variables** organised into seven thematic clusters. A "variable" represents a technical, operational, environmental, human, or institutional factor that may influence the legal acceptability of an AI system under IHL. Each variable is presented as a self-contained assessment page, comprising a description of the concept, a qualitative risk evaluation scale, an explanation of its legal relevance, and guidance on how it interacts with other variables. Together, the variables provide a structured basis for identifying and assessing the factors that contribute to legal risk.



Each variable in this Toolkit contains of the following elements:

- **Concept**: A short explanation of the variable, supported by practical examples to facilitate understanding even for non-experts.
- **Evaluation Scale**: A five-point qualitative risk scale that helps evaluators assess the extent to which the variable may contribute to legal risk under IHL in the anticipated operational context.
- **Legal Relevance**: A brief explanation of why the variable is legally relevant, linked to one or more of the **D-C-P** concern categories (*see below*).
- **Relationship Chart**: A visual overview of how the variable interacts with other variables, including dependencies, influential relationships, risky pairings or disqualifying combinations, trade-offs, and complementary safeguards.
- **Chart Legend**: An explanation of the symbols used in the Relationship Chart, and how the linked variables interact.

The characteristics and context-of-use of military AI systems may affect compliance with IHL, including the rules of distinction, proportionality, and precautions in attack. Assessing these implications is an important part of determining whether, and under what conditions, a military AI system may be lawfully used in a particular operational context.

The **legal relevance** of each variable is identified by reference to one or more of **three core IHL concern categories ('D-C-P')** below. These categories indicate the principal way(s) in which a variable may affect IHL compliance. For a more detailed discussion of the applicable IHL framework and its relevance to the assessment of military AI systems, see the accompanying report.

D DISTINCTION	**Distinction-related concerns:** The risk that the characteristics or context-of-use of the AI system may impair the ability to reliably distinguish between military objectives and civilians or civilian objects, including by contributing to misidentification or other classification errors.

C CIVILIAN HARM	**Civilian harm-related concerns:** The risk that the characteristics or context-of-use of the AI system may contribute to incidental civilian harm that is excessive in relation to the concrete and direct military advantage anticipated, including by affecting the information or assessments relevant to proportionality.
P PRECAUTIONS	**Precautions-related concerns:** The risk that limitations in the AI system's performance, predictability, understandability, or controllability may impair the ability of those planning or executing an attack to take feasible precautions to avoid, and in any event minimise, incidental civilian harm.

Each variable's **risk level is assessed using a qualitative Evaluation Scale**. Evaluators should use available evidence and their professional judgment to determine which descriptor best reflects the anticipated operational context. The Toolkit guides users through a step-by-step process to identify relationships between variables, recognise high-risk or disqualifying combinations, consider possible mitigation measures, and ultimately reach an overall assessment of the AI system.



Very Low Risk	Little or no identifiable risk to IHL compliance in the anticipated use-context.
Low Risk	Limited risk. Some attention may be required, but unlikely to raise significant legal concerns on its own.
Moderate Risk	Some risk to IHL compliance, depending on the operational context and its interaction with other variables. Caution and further assessment recommended.
High Risk	High risk to IHL compliance. Mitigation measures or additional safeguards should normally be considered before deployment.
Very High Risk	Significant legal risk. Unless adequately mitigated or justified by the broader assessment, deployment may be inappropriate.

Variables should not be considered in isolation. A variable presenting limited risk on its own may become more significant when combined with other factors; conversely, risks associated with one variable may be mitigated by safeguards reflected in another. The **Relationship Chart** and **Chart Legend** in each variable page show how variables may interact, and the **Relationship Matrix** (*see below*) provides a global overview of all relationships between variables.

The Toolkit supports, rather than prescribes, the legal assessment process. There is no mandatory sequence in which applicable variables must be evaluated, and institutions may adapt the process to suit their own procedures and decision-making structures. However, the basis for the assessment,

including decisions on variable applicability and risk evaluation, should be recorded to support transparency and consistency. The recommended workflow below provides one practical approach to applying the Toolkit in a structured, transparent, and context-sensitive manner.

Recommended workflow

The following ten steps provide a recommended workflow for applying the Toolkit in a **structured, transparent, and context-sensitive manner**. The workflow guides evaluators from defining the system and scope of the assessment, through the evaluation of applicable variables and identification of information gaps, to the consideration of legal relevance and appropriate risk-mitigation measures.

The workflow is intended to be flexible. Depending on the evaluator's role, the stage of the AI system lifecycle, and relevant institutional processes, the order or application of individual steps may be adapted where appropriate.

See also the **Evaluation Flowchart** below, which summarises the recommended workflow discussed in this section.

STEP 0 | Prepare and Scope the Assessment

Before beginning the assessment, confirm that you have:

- ✓ **Identified** the AI system under consideration.
- ✓ Defined its **intended operational context**, as the legal acceptability of a military AI system is inherently context-dependent.
- ✓ **Collected** the available technical, operational, and institutional information relevant to the assessment.

STEP 1 | Identify the Variable to Evaluate

Begin by selecting a variable to assess. There is no prescribed order in which variables must be evaluated. The starting point will depend on the purpose of the assessment, the information available, and the stage of the AI system lifecycle.

STEP 2 | Familiarise Yourself with the Variable

Read the Concept and Evaluation Scale sections before assessing the variable. These explain what the variable assesses and what information may be relevant for the Evaluation Scale.

≠ Variables marked with this symbol in the Relationship Chart are commonly confused with the variable under consideration. Reviewing them can help distinguish similar concepts and avoid misinterpretation.

STEP 3 | Is the Variable Applicable to the System?

Not every variable in the Toolkit will be relevant to every AI system or proposed use. Before continuing, determine if the **variable is applicable** to the system and its proposed use-context (*see the table below*).

Observation	Classify as	What do to next
The variable concerns a factor that is present or could materially affect the AI system, its proposed use, or its IHL-relevant effects.	***Applicable***	Proceed with the variable in the assessment
The variable concerns a feature, function, or circumstance that is not relevant to the system or proposed use.	***Not Applicable***	Record N/A and briefly explain why the variable does not apply.

In determining applicability, consider the scope and definition of each variable in light of the specific system and proposed use. A lack of sufficient information to evaluate an otherwise relevant variable does **not** make that variable "Not Applicable" (*instead, see Step 5*).

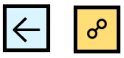
This Step should be revisited if the AI system, its proposed use, or the operational context changes.

STEP 4 | Gather the Relevant Information

Using the **Concept** and **Evaluation Scale** sections as a guide, collect the information needed to assess the variable. Depending on the AI system, this may include technical, operational, environmental, human factors, or institutional information.



The evaluator's *ability to assess the current variable may depend on* variables marked with this symbol in the Relationship Chart. Consider if this is the case before continuing, and whether more relevant information about the current variable could be obtained from these sources.



Variables marked with these symbols in the Relationship Chart *may influence the current variable* (through a linear and inverse relationship respectively), and should therefore be considered when gathering the relevant information.

STEP 5 | Check Whether You Have Sufficient Information

Before assessing the variable, confirm that sufficient information is available to apply the Evaluation Scale. If the available information is insufficient, determine whether the missing information can be obtained now or is expected to become available only at a later stage, and proceed accordingly:

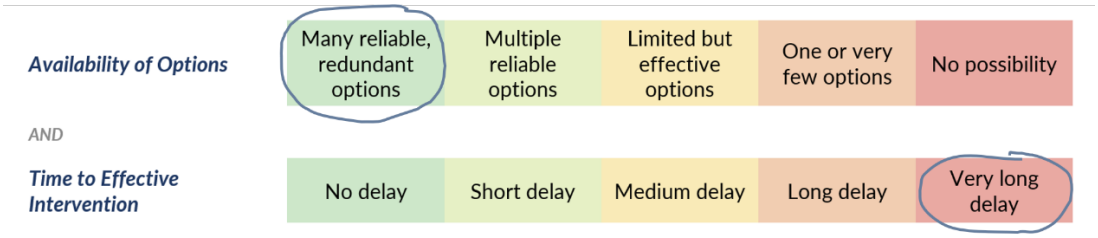
- **The information can be obtained now:** Pause the assessment of this variable until the necessary information has been obtained (e.g. from a superior, developer, or through additional testing). In the meantime, you may continue evaluating another variable.

- **The information will only become available later:** This may occur where evaluation depends on information that only emerges during a later stage of the system's lifecycle. For example, determining the number or concentration of civilians in an operational area (*see **[Civilian Presence]***) may only be possible possible at the point of operational planning or use. Where appropriate, a **provisional evaluation** may be recorded on the basis of clearly identified assumptions and the information currently available. The evaluation should be revisited once the necessary information becomes available, where necessary by a different evaluator at a later stage of the AI system lifecycle.

STEP 6 | Assess the Variable Using the Evaluation Scale

Using the Evaluation Scale, determine which risk descriptor best reflects the variable in the anticipated operational context. Base the assessment on the available evidence and your professional judgement.

Where a variable includes multiple indicators, consider all of them together when determining the overall risk level. The overall evaluation should reflect the combined effect of the indicators on the particular context, rather than any single indicator in isolation.



For example, the variable **[Intervention Ability]** above includes two indicators. A reasonable evaluator may conclude that *overall risk remains high* even where the first indicator scores positively. In the example above, multiple options for stopping or recalling a malfunctioning AI system may provide limited protection if all of those options are subject to significant delays.

STEP 7 | Review Variable Relationships for 'Risky Pairings'

Once you have assessed the variable, review its Relationship Chart. This highlights interactions with other variables that may affect the overall risk assessment. Pay particular attention to variables marked with symbols which identify high-risk (△) or critically dangerous (X) combinations.

- ✓ If a variable on the Relationship Chart with the △ or X symbol has not yet been evaluated, evaluate that variable first before continuing.

If the current variable **and** another variable connected by a △ or X symbol **both present medium or high risk**, you have identified a **'Risky Pairing'**. In this case, consider whether the risk associated with one or both variables can be reduced through mitigation measures.

- ➢ Where effective mitigation is **possible**, implement the relevant measures and reassess the affected variable(s) under Step 6.

- Where the risks **cannot** be adequately mitigated, there is notable risk of legal noncompliance. Record the residual risk and consider its implications later as part of the overall assessment of legal acceptability under Step 10.



This symbol in the Relationship Chart identifies *suggested mitigation measures* for the current variable. Consult the Chart Legend for more details on these suggested measures.



This symbol in the Relationship Chart highlights "trade-offs", where reducing risk in one variable may *increase* risk in another. Consider these interactions before implementing mitigation measures.

STEP 8 | Confirm the Final Risk Evaluation

Confirm the final qualitative risk assessment for the variable after considering all relevant information, related variables, and any mitigation measures. **Record the outcome and basis for the evaluation** before proceeding to the next variable.



Variables marked with this symbol in the Relationship Chart *complement the current variable*. Low risk in both may provide an additional layer of safety or redundancy.



Variables with this symbol in the Relationship Chart *may be influenced by the current variable*. The evaluation of the current variable may therefore affect their risk level. Take this into consideration when evaluating these variables.

STEP 9 | Repeat the Process for Remaining Variables

Return to Step 1 and repeat the process. Continue until all variables have either been evaluated, or identified for later reassessment under Step 4.

FINAL STEP | Reach an Overall Assessment of Legal Acceptability of the AI System

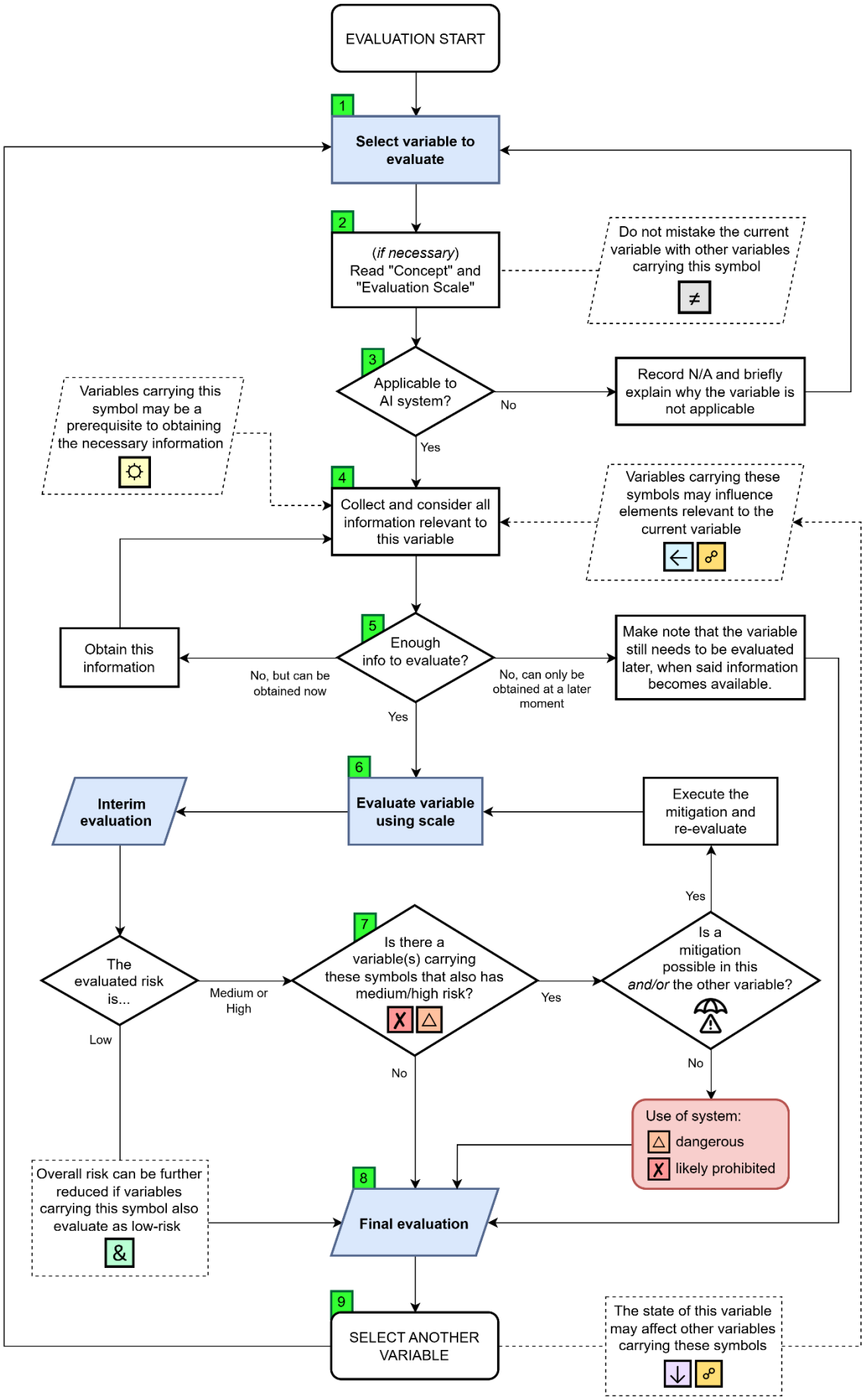
Once all variables have been considered, assess the results together, including relevant interactions between variables, residual risks following mitigation, and any outstanding information gaps or provisional evaluations identified under Step 5. On this basis, **reach an overall assessment of the legal acceptability of the AI system in the proposed use-context**.

The assessment should identify:

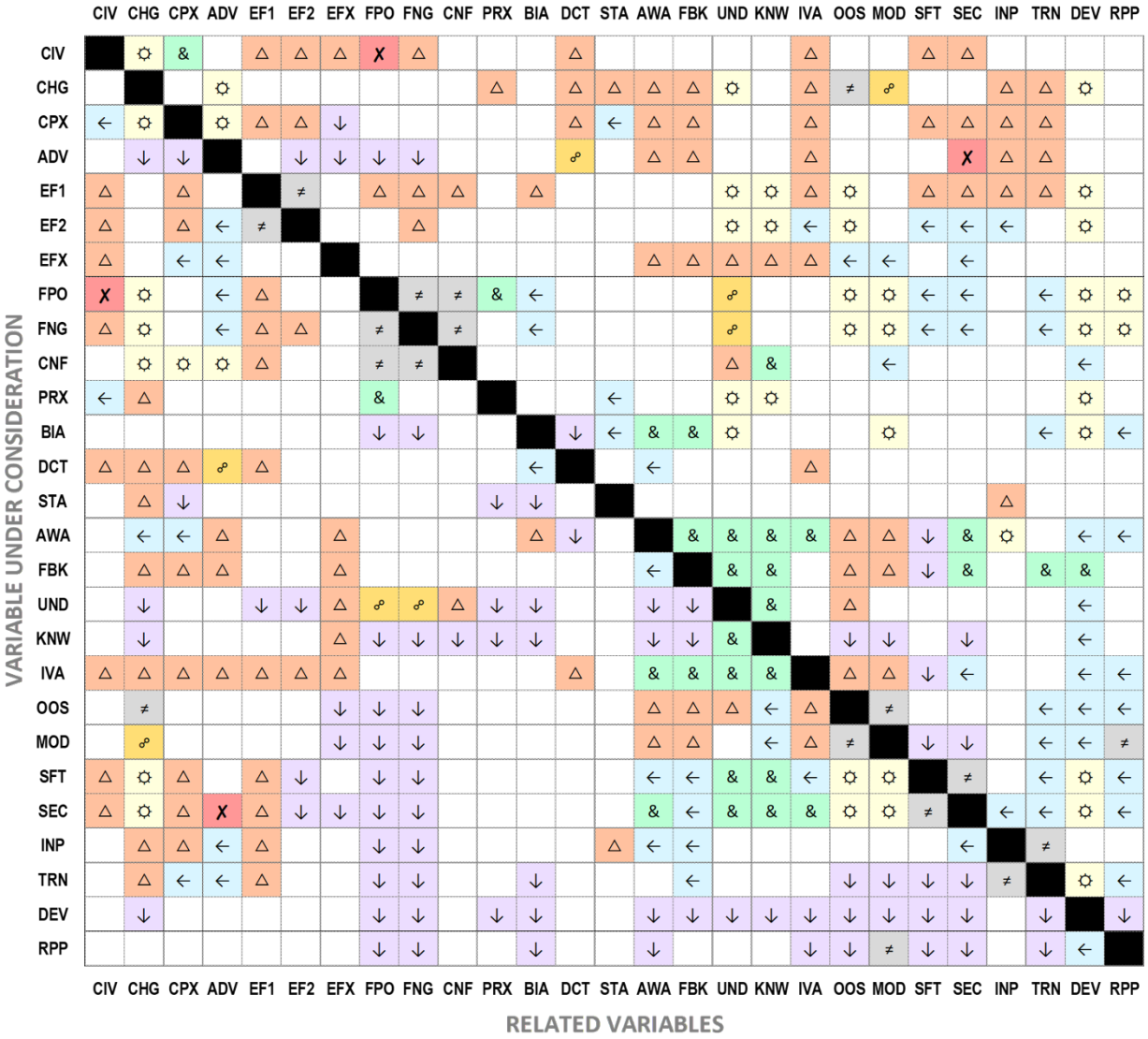
- whether the proposed use is legally acceptable on the information available; and
- whether additional information, safeguards, mitigation measures, or modifications are required before a determination can be made.

Where the evaluator is not the relevant decision-maker, the assessment should inform the appropriate authorisation or decision-making process.

Evaluation Flowchart



Relationship Matrix



Suggested crosscheck order

☼	Potentially necessary to be able to evaluate
←	May impact the current variable
≠	Not the same
X	Critically dangerous combination
△	Dangerous combination
☍	Inversely correlated
&	Important even if variable scores well
↓	Current variable can impact...

Abbreviations

CIV Civilian Presence
CHG Changing Environment
CPX Environment Complexity
ADV Adversarial Threat
EF1 Intended & Likely Effects
EF2 Possible & Indirect Effects
EFX Unpredictable Behaviour
FPO False Positive Rate
FNG False Negative Rate
CNF Confidence Threshold
PRX Target Proxy
BIA Bias
DCT Civilian Detection Ability
STA Status Fluidity of Target
AWA Contextual Awareness
FBK Feedback Loop
UND Understandability
KNW Technical Knowledge
IVA Intervention Ability
OOS Out-of-Scope Use
MOD Modification
SFT System Safety
SEC System Security
INP Input Data
TRN Training Data
DEV Engagement with Developer
RPP Repurposing

Table of Contents

CLUSTER 1: ENVIRONMENT ____ *1*

Civilian Presence ____ **2**
Concept ____ 2
How to evaluate? ____ 3
Legal relevance ____ 3
Relationship Chart ____ 4
Chart Legend ____ 4

Changing Environment ____ **6**
Concept ____ 6
How to evaluate? ____ 8
Legal relevance ____ 8
Relationship Chart ____ 9
Chart Legend ____ 9

Environment Complexity ____ **11**
Concept ____ 11
How to evaluate? ____ 11
Legal relevance ____ 12
Relationship Chart ____ 12
Chart Legend ____ 13

Adversarial Threat ____ **15**
Concept ____ 15
How to evaluate? ____ 16
Legal relevance ____ 16
Relationship Chart ____ 17
Chart Legend ____ 17

CLUSTER 2: EFFECTS ____ *19*

Intended & Likely Effects ____ **20**
Concept ____ 20
How to evaluate? ____ 21
Legal relevance ____ 22
Relationship Chart ____ 22
Chart Legend ____ 22

Possible & Indirect Effects ____ **25**
Concept ____ 25
How to evaluate? ____ 26
Legal relevance ____ 27
Relationship Chart ____ 27
Chart Legend ____ 27

Unpredictable Behaviour ____ **29**
Concept ____ 29
How to evaluate? ____ 30
Legal relevance ____ 30
Relationship Chart ____ 31
Chart Legend ____ 31

CLUSTER 3: TARGETS & IDENTIFICATION ____ *33*

False Positive Rate ____ **34**
Concept ____ 34
Legal relevance ____ 36
How to evaluate? ____ 36
Relationship Chart ____ 37
Chart Legend ____ 37

False Negative Rate 40
Concept 40
How to evaluate? 42
Legal relevance 42
Relationship Chart 42
Chart Legend 43

Confidence Threshold 45
Concept 45
How to evaluate? 47
Legal relevance 47
Relationship Chart 48
Chart Legend 48

Target Proxy 50
Concept 50
How to evaluate? 51
Legal relevance 51
Relationship Chart 52
Chart Legend 52

Bias 54
Concept 54
How to evaluate? 56
Legal relevance 56
Relationship Chart 57
Chart Legend 57

Civilian Detection Ability 59
How to evaluate? 60
Legal relevance 61
Relationship Chart 61
Chart Legend 62

Status Fluidity of Target 63
Concept 63
How to evaluate? 64
Legal relevance 65

CLUSTER 4: HUMAN AWARENESS *67*

Contextual Awareness 68
Concept 68
How to evaluate? 69
Legal relevance 70
Relationship Chart 70
Chart Legend 71

Feedback Loop 73
Concept 73
How to evaluate? 75
Legal relevance 75
Relationship Chart 76
Chart Legend 76

Understandability 78
Concept 78
How to evaluate? 80
Legal relevance 80
Relationship Chart 81
Chart Legend 81

Technical Knowledge 83
How to evaluate? 84

Legal relevance ____ 85
Relationship Chart ____ 85
Chart Legend ____ 85

CLUSTER 5: HUMAN INTERACTIONS ____ 88

Intervention Ability ____ 89
Concept ____ 89
How to evaluate? ____ 90
Legal relevance ____ 90
Relationship Chart ____ 91
Chart Legend ____ 91

Out-of-Scope Use ____ 94
Concept ____ 94
How to evaluate? ____ 95
Legal relevance ____ 96
Relationship Chart ____ 96
Chart Legend ____ 97

Modification ____ 99
Concept ____ 99
How to evaluate? ____ 101
Legal relevance ____ 101
Relationship Chart ____ 102
Chart Legend ____ 102

CLUSTER 6: OPERATIONAL SAFETY & SECURITY ____ 105

System Safety ____ 106
Concept ____ 106
How to evaluate? ____ 107
Legal relevance ____ 107
Relationship Chart ____ 108
Chart Legend ____ 108

System Security ____ 111
Concept ____ 111
How to evaluate? ____ 113
Legal relevance ____ 113
Relationship Chart ____ 114
Chart Legend ____ 114

Input Data ____ 117
Concept ____ 117
How to evaluate? ____ 119
Legal relevance ____ 119
Relationship Chart ____ 120
Chart Legend ____ 120

CLUSTER 7: DEVELOPMENT AND ACQUISITION ____ 122

Training Data ____ 123
Concept ____ 123
How to evaluate? ____ 124
Legal relevance ____ 124
Relationship Chart ____ 125
Chart Legend ____ 125

Engagement with Developer ____ 128
Concept ____ 128
How to evaluate? ____ 129
Legal relevance ____ 129
Relationship Chart ____ 130

Repurposing _____ **132**
Concept _____ 132
How to evaluate? _____ 133
Legal relevance _____ 133
Relationship Chart _____ 134
Chart Legend _____ 134

CLUSTER 1: ENVIRONMENT

[CIV]

Civilian Presence

[DEFINITION]
Identifies the **civilian persons/objects** that the **AI system is likely to affect**.

Concept

1 All persons and objects that not military objectives are considered **civilian**.

2 What is **'likely to be affected' depends** on the type of AI system, the circumstances in which it is deployed, and its likely effects on the environment (*see also **[Intended & Likely Effects]** and **[Possible & Indirect Effects]***).

> *[EXAMPLE]*
> An AI-enabled drone is deployed to loiter above two small villages in search of enemy vehicles. The villages are connected by a bridge that is used daily by civilians. Because the bridge lies within the drone's designated search area, it should be treated as a civilian object that could be affected by the drone's operation.

3 Some operational environments contain a comparatively **higher concentration** of civilian persons/objects than others.

> *[EXAMPLE]*
> An urban battlefield is likely to contain a higher density of civilians, buildings, civilian goods, and infrastructure than a small village.

4 High levels of **interconnectivity** (e.g. of infrastructure, of civilian services) can also broaden the range of civilian persons/objects that may be 'affected'.

> *[EXAMPLE]*
> AI-powered malware is deployed against a commercial data centre used by both the adversary and civilians for cloud computing. Because military and civilian data and infrastructure are intermingled, the malware may also disrupt civilian services.

5 Higher civilian presence **generally increases the risk of harm to civilians**. This is because more civilians and civilian objects are exposed to the AI system's operations, increasing both the likelihood of misidentification errors and the potential for excessive collateral damage.

From misidentification. Where an AI system is used to identify military objectives, a higher presence of civilian persons/objects increases the number of opportunities for the system to mistakenly identify them **as targets,** even if its performance (*see **[False Positive Rate]***) remains unchanged.

> *[EXAMPLE]*
> An AI classifier incorrectly labels 1 in every 1,000 civilian vehicles as military targets. In an area containing relatively few civilian vehicles, only a small number are at risk of being misidentified. In a densely populated urban environment containing thousands of civilian vehicles, the same error rate could result in many more civilian vehicles being incorrectly identified as military objectives.

From excessive collateral effects. A higher civilian presence also increases the likelihood that civilian persons/objects will be ***incidentally* affected by an attack** (e.g. because they are located within a weapon's effect radius, *see* ***[Intended & Likely Effects]***).

> *[EXAMPLE]*
> An AI system identifies an enemy air-defence position and launches a missile to destroy it. If the target is located in a town centre, the explosion is more likely to cause civilian casualties and damage nearby civilian infrastructure, compared to if the same target were located on the outskirts of the town.

How to evaluate?

Civilian Density	None	Low	Medium	High	Very high
AND					
Interconnectivity	None	Low	Medium	High	Very high

[Civilian Density]	A qualitative determination of the density of civilian or other protected people or objects (see ¶3). Depending on the context, relevant tools or tables can be used for standardisation (e.g. density tables for comparing housing density).
[Interconnectivity]	A qualitative assessment of the extent to which civilian or other protected objects, services and infrastructures might be interconnected (see ¶4).

Legal relevance

D DISTINCTION	A higher presence of civilian persons/objects increases the chance of misclassifications in absolute terms (see ¶5), undermining compliance with the rule of distinction.
C CIVILIAN HARM	A higher presence of civilian and other protected persons/objects increases the chance of harming civilians in absolute terms (see ¶5). This increases the risk of higher collateral damage, even if the attacks are aimed at legitimate aimed military targets. This may render the attack disproportionate and may trigger the need to enact additional precautionary measures to reduce risk to civilians.

Relationship Chart

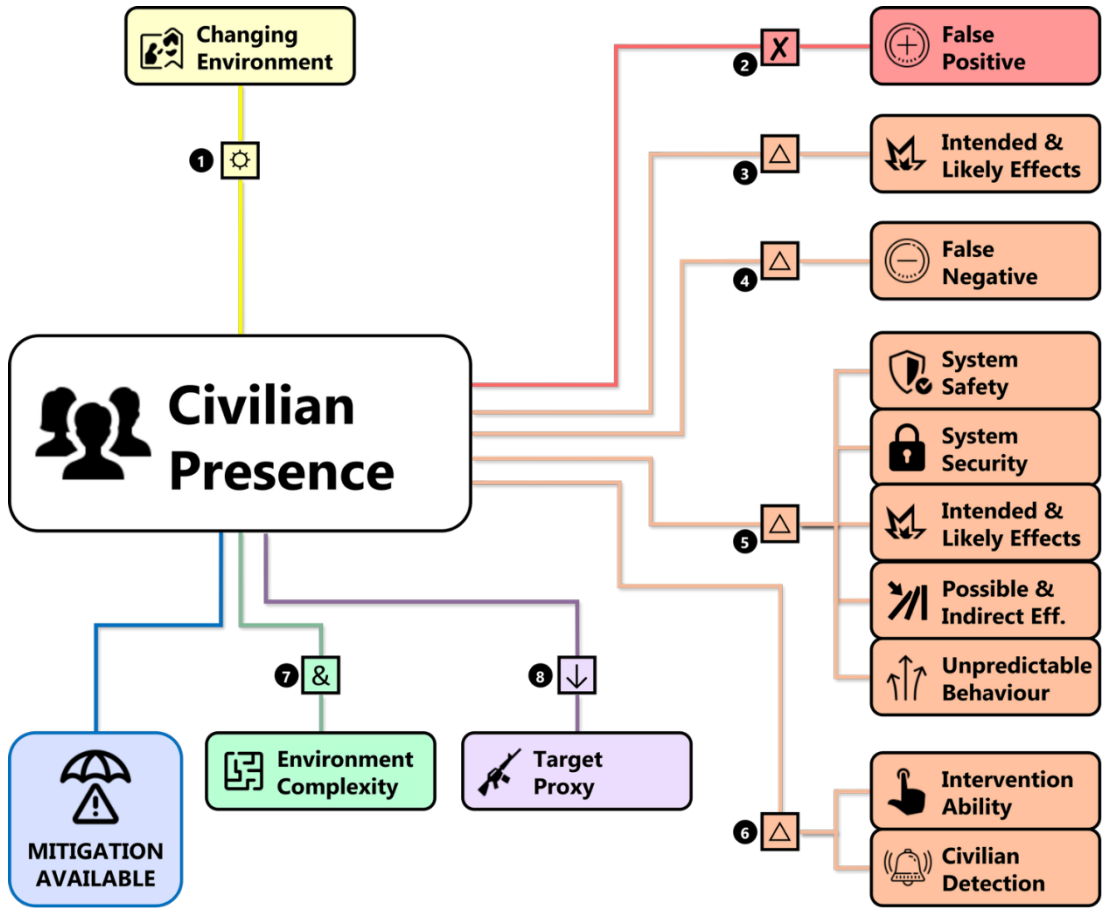


Chart Legend

BEFORE EVALUATING, CONSIDER |

1 ☼ Evaluations about [Civilian Presence] may become **obsolete more quickly** if the environment is dynamic or rapidly-changing (= *poor evaluation in [Changing Environment]*).

RISKY PAIRINGS |

2 ✗ (*Applicable if AI is used to detect military targets*) It is **likely prohibited to have simultaneous poor evaluations in [Civilian Presence] and [False Positive Rate]**: the combination of many civilians and a poor ability to distinguish them may render the system an indiscriminate means or method of warfare.

3 △ A high presence of civilians (= *poor evaluation in [Civilian Presence]*) is particularly dangerous if the attacks are expected to be harmful or destructive (= *poor evaluation in [Intended & Likely Effects]*), as this may **aggravate the harmful collateral damage caused**.

4 △ (*Applicable if AI is used to detect collateral concerns*) A poor evaluation in [Civilian Presence] may be dangerous in combination with a poor evaluation in [False Negative Rate], as it may cause **more risks to civilians to remain undetected**.

5 △ If many civilians are present (= *poor evaluation in [Civilian Presence]*), it is important to **ensure performance** generally (= *strong evaluations in [System Safety], [System Security]*) and to take extra effort into **curtailing effects** (= *strong evaluations in [Intended & Likely Effects], [Possible & Indirect Effects], [Unpredictable Behaviour]*) in order to reduce risk to civilians.

6 △ (*Applicable if AI is used to detect military targets*) If many civilians are present (= *poor evaluation in [Civilian Presence]*), the ability to **dynamically correct, recall, or deactivate** the AI system if too many civilian persons/objects risk being harmed becomes particularly valuable. This can be achieved either via human intervention (= *strong evaluation in [Intervention Ability]*) or autonomously (= *strong evaluation in [Civilian Detection]*).

AFTER EVALUATING, CONSIDER

[MITIGATION]

Where [Civilian Presence] indicates elevated risk, consider the following mitigating measures:

- **Barring Entry to Civilians**: As a precaution, perimeters or checkpoints can be used to prevent (more) civilian persons/objects from entering the operational area the AI system is likely to affect.
- **Warnings**: Warnings or evacuation orders can put in place to reduce civilian presence in an operational area.
- **Adversary Segregates Civilian Persons/Objects:** Defenders have the legal duty, to the extent reasonable, to segregate military assets from civilian persons/objects. Incentivising the adversary to do so (e.g. through negotiating safe evacuation routes) can reduce civilian presence.

[WARNING]

Civilians have **the right to ignore** warnings or evacuation orders. You cannot assume that the status of [Civilian Presence] improved purely on the basis of merely *issuing* such orders, without *(re-)evaluating* actual operational conditions.

7 & Even where [Civilian Presence] indicates a low level of civilian presence, specific civilian objects may still create significant legal risks. For example, the presence of critical infrastructure or cultural property (= *poor evaluation in [Environment Complexity]*) may require particular attention because such objects may be **specially protected or, if damaged, may result in widespread or cascading effects**.

8 ↓ Higher civilian presence (= *poor evaluation in [Civilian Presence]*) increases the likelihood that **non-lawful targets fall into a feature set used to identify targets** (= *corresponds with a poor evaluation of [Target Proxy]*).

[CHG]

Changing Environment



[DEFINITION]

Identifies how quickly aspects of the operating environment that are relevant to the AI system may **change over time**.

[COMPARE]

This variable concerns **changes in the operating environment**; it does **NOT cover changes to the mission or intended use** of the system. For operational changes, see instead **[Out-of-Scope Use]**.

Concept

1 The operating environment may change over time due to **natural developments, the actions of civilians or other third-parties, adversary behaviour**, or a combination of these factors. Such changes may affect whether the information, assumptions or patterns on which an AI system relies, remain accurate.

[EXAMPLE]

An AI system provides operators with a real-time map that distinguishes civilians from enemy soldiers (based on historical patterns and incoming intelligence). Civilians begin demonstrating in support of their armed forces near military installations, with some wearing military-style clothing and carrying national flags. Because this behaviour differs from the patterns on which the AI system relies, it incorrectly identifies some civilian demonstrators as enemy soldiers.

2 The actual rate of environmental change is **context-specific**. Some operating environments are inherently more dynamic than others.

[EXAMPLE]

An urban environment (e.g. cities) may change rapidly because of the movement and activities of civilians, vehicles, adversaries, and other third parties. By contrast, a deep-sea environment may be comparatively static, with fewer actors and relevant environmental changes occurring less frequently.

3 A rapidly changing environment may cause the **conditions encountered by the AI system to diverge from those represented in its training and testing data** (e.g. data drift). This type of distribution shift can negatively affect performance.

Data drift occurs when the conditions in which an AI system operates increasingly differ from those for which it was developed, trained, or tested. For example, a system designed and tested for Context A may initially perform as expected, but environmental changes may gradually or abruptly shift the

operating conditions towards Context B. If the system was not designed or trained to operate under these new conditions, its performance may deteriorate, potentially in ways that are difficult to predict.

> *[EXAMPLE]*
> An AI system is trained and tested for relatively dry environments, in which an armed conflict is taking place. Following unusually heavy rainfall, an upstream dam fails and floods the operational environment. Because the system was not trained or tested under flooded conditions, is performance deteriorates and it begins making significantly more errors.

4 Not every environmental change is relevant to an AI system's performance. **[Changing Environment]** should therefore be assessed by considering how quickly **the aspects of the environment *that may affect the system's performance or behaviour*** are likely to change.

An environment may be highly dynamic in some respects, while remaining relatively static in those that matter to a particular system. The relevance of the changes therefore depends on how the system functions and the information on which it relies.

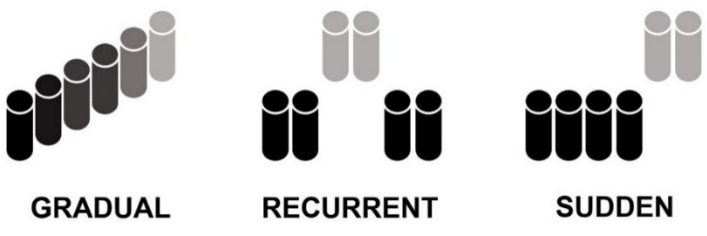
> *[EXAMPLE]*
> System A uses optical sensing and System B uses thermal sensing to detect targets. A heatwave significantly alters the thermal signatures of objects exposed to the sun but causes little change to their visuals. This environmental change may therefore be relevant to the performance of System B, while having little to no effect on System A.

Determining which environmental changes are relevant may require insights into how the AI system functions (*see **[Understandability]** and **[Engagement with Developer]***). For example, knowing what features the system relies on to identify or classify objects (*see also **[Target Proxy]***) can help determine whether those features are likely to change during an operation.

5 Environment change may occur **gradually or abruptly**, and may be **expected or unexpected**. The speed and predictability of that change can affect how readily an evaluator can anticipate and account for changing environmental conditions.

Some changes occur **gradually** and follow reasonably predictable patterns. For example, seasonal weather conditions may progressively restrict adversary movement in ways that are familiar from previous years. Such changes may be easier to anticipate and incorporate into the system's operating parameters.

Other changes may occur **abruptly** and with little warning. For example, an adversary facing imminent strategic defeat may suddenly adopt substantially different tactics. If those tactics create operating conditions that differ from those anticipated when the AI system was tested or configured, its performance or behaviour may be affected.



Different types of environmental shifts

How to evaluate?

Rate of Change	Static	Mostly static	Somewhat dynamic	Dynamic	Very dynamic
AND					
Nature and Predictability	Predictable and gradual	Mostly predictable and gradual	Partly predictable, abrupt changes possible	Mostly unpredictable, abrupt changes likely	Unpredictable, abrupt changes very likely

[Rate of Change] A qualitative assessment on how often system-relevant factors (see ¶4) in the environment are prone to changing.

[Nature and Predictability] A qualitative assessment on whether said changes are likely to be foreseeable or not (see ¶5).

Legal relevance

D DISTINCTION	A rapidly or abruptly changing operational environment increases the risk that an AI system's underlying model becomes misaligned with the evolved circumstances (e.g. the 'data drift' problem, see ¶3). As model accuracy degrades, the risk of misidentification grows, potentially resulting in the targeting of civilians and civilian objects in violation of the distinction rule.
C CIVILIAN HARM	Rapid environmental change may degrade the system's performance, increasing the risk that civilians and civilian objects are not correctly identified or accounted for. This may also distort the estimation of the expected civilian harm and, in turn, undermine proportionality assessments.
P PRECAUTIONS	Rapid or unpredictable environmental change may cause information and prior assumptions underlying an earlier legal assessment to become outdated more quickly. It may also increase the risk that unforeseen circumstances may arise during the operation, requiring the system's continued use to be reassessed or additional precautions to be taken.

Relationship Chart

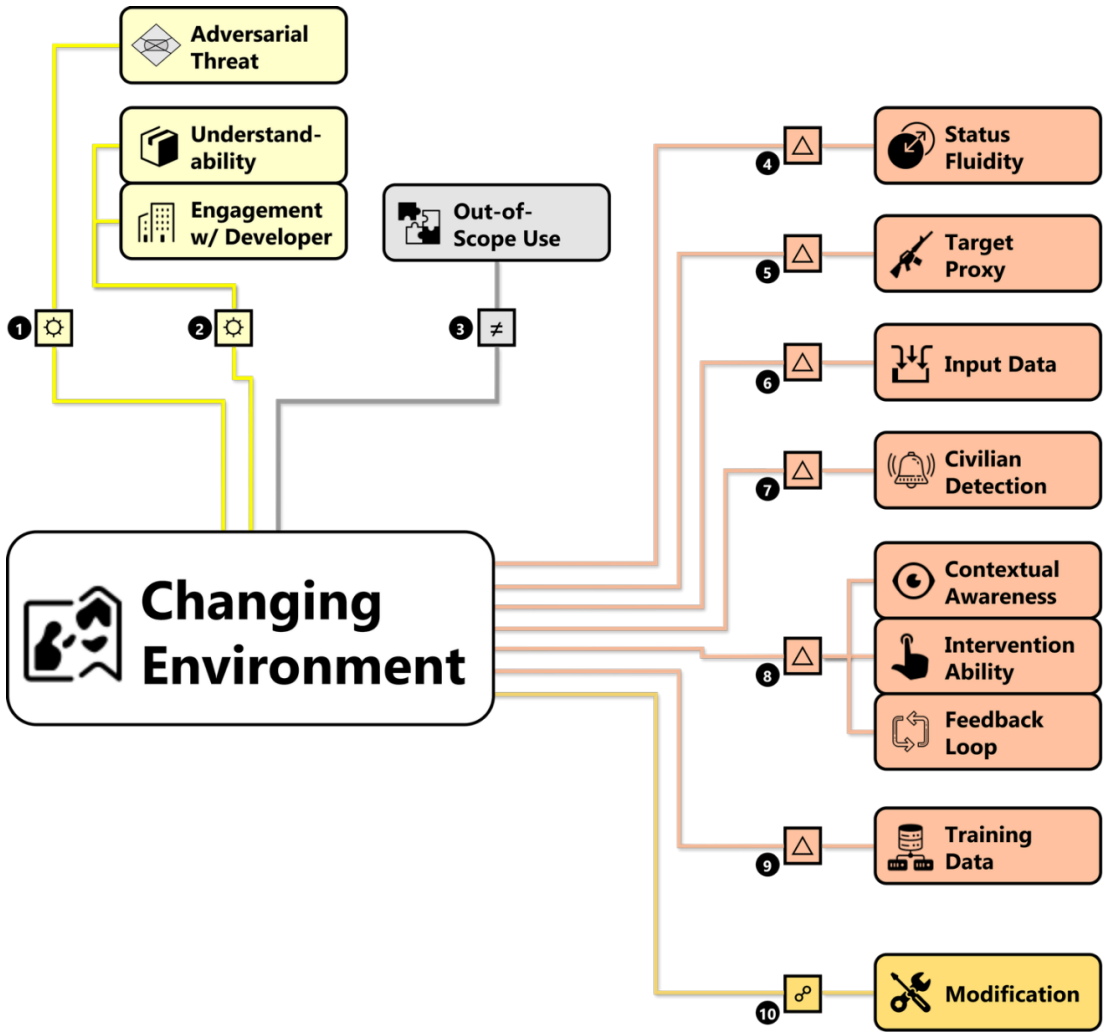


Chart Legend

BEFORE EVALUATING, CONSIDER |

1 ☼ This variable may need to be **evaluated more frequently if the adversary is particularly active** (= *poor evaluation in [Adversarial Threat]*). Adversarial activity is a notorious source of abrupt and unpredictable changes to the environment.

2 ☼ To **identify what aspects are most determinative of a system's performance or behaviour** (see ¶4), information about the model's internal logic, behaviour, and vulnerabilities may be necessary. Such information can be obtained through an understanding of the model's logic (= *strong evaluation in [Understandability]*) and/or insights from the developer (= *strong evaluation in [Engagement with Developer]*).

3 ≠ **Do not confuse** this variable with [Out-of-Scope Use]. Data drift (see ¶3) is caused by changes in the environment that the operator often has no control over. Compare with [Out-of-Scope Use], which results from the operator's decision to deploy a system in a context for which it was not designed or trained (this is an error by human decision-makers).

4 △ A poor evaluation in [Changing Environment] is **particularly dangerous if the targetable status of persons or objects is liable to change** (= *poor evaluation in [Status Fluidity]*). For example, civilian infrastructure may qualify as a military objective only while it is being used for a particular military purpose. In a rapidly changing environment, such circumstances may change quickly, increasing the risk that the AI system acts on outdated information about the object's status.

5 △ If [Changing Environment] evaluates poorly, is recommended to choose **more specific target proxies** upon which target identifications and decisions to attack will be made (= *strong evaluation in [Target Proxy]*). Militarily-unique proxies are less likely to become outdated or invalid, even as the operational environment evolves.

6 △ If [Changing Environment] evaluates poorly, ensuring the **recency of data and intelligence** (= *strong evaluation in [Input Data]*) becomes crucial. Otherwise, both AI and human decisions will be made on the basis of outdated information.

7 △ If [Changing Environment] evaluates poorly, a strong evaluation in [Civilian Detection] becomes particularly valuable, as this can **account for more rapid changes** in civilian positioning.

8 △ If [Changing Environment] evaluates poorly, a strong evaluation in [Contextual Awareness] allows operators to **observe changes** more rapidly, and strong evaluations in [Intervention Ability] and [Feedback Loop] allow operators to **adapt/intervene** accordingly.

9 △ If [Changing Environment] evaluates poorly, consider whether the **training data adequately represents the different environmental conditions the system may reasonably encounter** as its operating environment changes (= *strong evaluation in [Training Data]*). Gaps in the training data may increase the risk of performance degradation when conditions change.

10 ☍ If [Changing Environment] evaluates poorly, **adaptation and modification** may enable the AI system to 'evolve along with the environment' to maintain performance. However, frequent modification—which corresponds with a poor evaluation in [Modification]—**may amplify the lack of foreseeability** already caused by the poor [Changing Environment] score.

[CPX]

Environment Complexity

[DEFINITION]

Identifies how **difficult the operating environment is** for the AI system to observe, interpret, and operate within.

Concept

1 Complex or cluttered environments can make it **more difficult for an AI system to operate reliably**. Multiple actors, physical obstacles, sensor interference, and unpredictable interactions may make relevant information harder to detect or interpret, and increase uncertainty about the system's performance.

2 Environmental complexity may arise from several factors, including (non-exhaustive list):

- **Number and variety of actors**: Environments containing adversaries, civilians, third parties, vehicles, or other autonomous systems may be more difficult for an AI system to interpret and navigate. For example, an urban counterinsurgency environment involving numerous interacting actors is generally more complex than air-to-air combat (which features a comparatively limited number of actors).
- **Clutter and obstructions**: Buildings, terrain, vegetation, traffic, or other physical features may obscure relevant objects, interfere with sensors, or make navigation and identification more difficult. For example, autonomous navigation through narrow, congested urban streets is generally more complex than navigation along an open highway.
- **Civilian presence**: The presence and density of civilians and civilian objects may increase the complexity of the operating environment (*see **[Civilian Presence]***).
- **Sensitive or specially-protected objects**: The presence of objects requiring particular legal or operational care—such as hospitals, cultural objects, critical infrastructure, installations containing dangerous forces, or environmentally sensitive sites—may increase the complexity of decisions made within the operating environment.

3 **Adversaries** may deliberately try to increase the complexity of the operational environment (e.g. by erecting obstacles, increasing clutter, using camouflage, using human shields, or using deception) to disrupt or deceive AI systems. *See also **[Adversarial Threat]***.

How to evaluate?

Complexity	Very low	Low	Medium	High	Very high

[Complexity] An overall qualitative assessment of how complex the operating environment is, taking into consideration factors from ¶2 and ¶3 (as applicable).

Legal relevance

D DISTINCTION	A more complex environment is more demanding on AI systems, as errors can be introduced from more sources. Therefore, the risk of misidentification grows, potentially resulting in the targeting of protected persons or objects in violation of the principle of distinction.
C CIVILIAN HARM	Errors and failures caused by the complexity of the environment may cause the system to fail to recognise protected individuals and objects, therefore distorting the estimation of incidental harm and, in turn, undermining proportionality assessments.

Relationship Chart

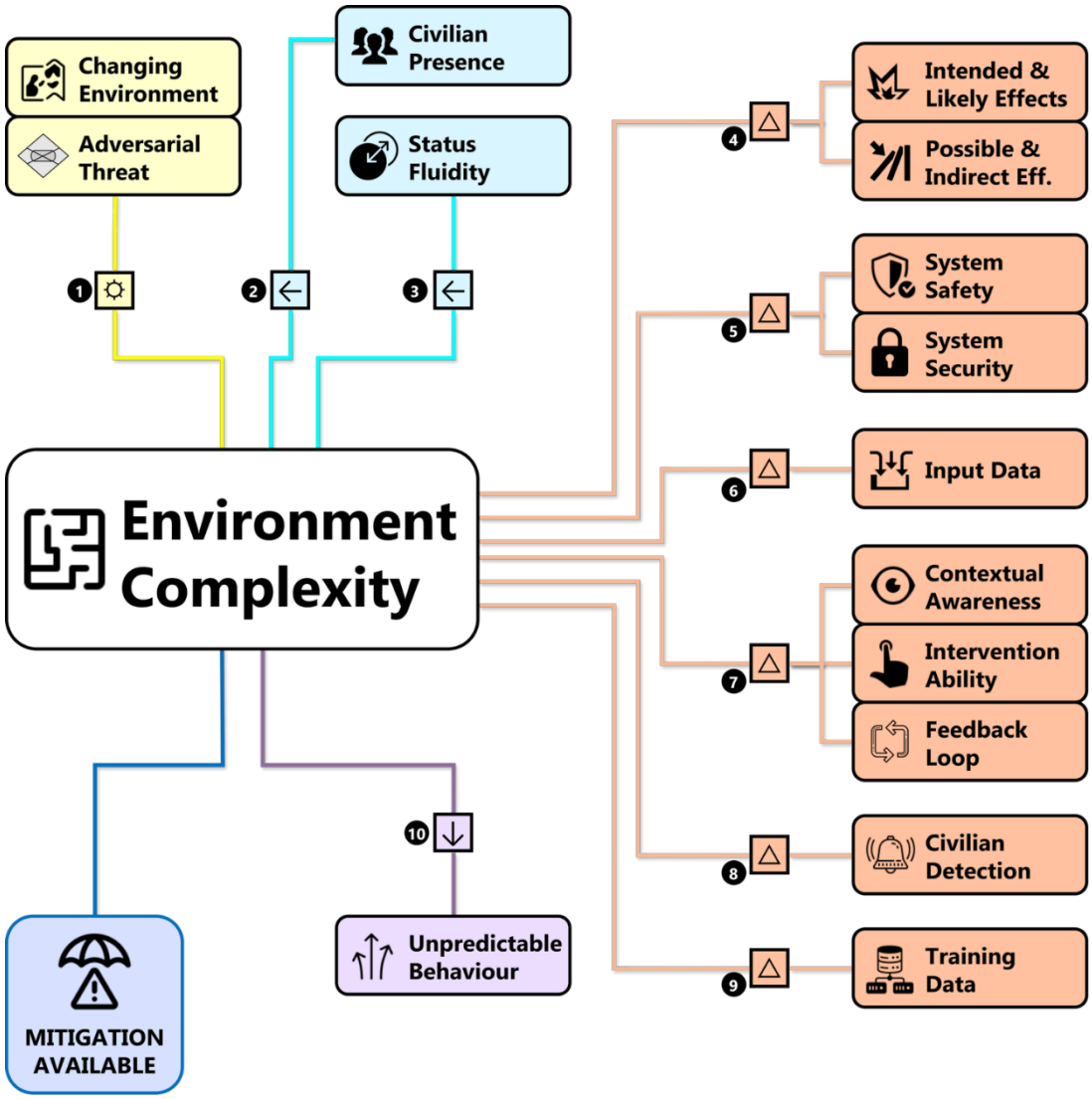


Chart Legend

BEFORE EVALUATING, CONSIDER |

1 ☼ This variable may need to be **evaluated more frequently** if the operational environment evolves rapidly (= *poor evaluation in [Changing Environment]*) or if the adversary is particularly active (= *poor evaluation in [Adversarial Threat]*). Both factors may cause a non-complex environment to quickly and unexpectedly turn into a complicated one.

2 ← A greater presence or density of collateral concerns (= *poor evaluation in [Civilian Presence]*) **generally complicates the operating environment** for military AI systems (= *negative effect on [Environment Complexity]*). See also ¶2.

3 ← If the intended targets are only circumstantially targetable, such as civilian infrastructure currently being used by military (= *poor evaluation in [Status Fluidity]*), this may increase the complexity of the operating environment (= *negative effect on [Environment Complexity]*) since such objects are often **dual-use objects and/or located in proximity to other civilian objects**.

RISKY PAIRINGS |

4 △ If [Environment Complexity] evaluates poorly, **mitigating or minimising likely effects** (= *strong evaluations in [Intended & Likely Effects] and [Possible & Indirect Effects]*) may be crucial.

5 △ If [Environment Complexity] evaluates poorly, **ensuring strong evaluations in [System Safety] and [System Security]** should be a priority.

6 △ If [Environment Complexity] evaluates poorly, guaranteeing the use of **recent and high-quality input data** (= *strong evaluation in [Input Data]*) should be a priority. Outdated or poor-quality data may further negatively impact the AI system's performance.

7 △ If [Environment Complexity] evaluates poorly, strengthening the evaluations of [Contextual Awareness], [Intervention Ability] and [Feedback Loop] allow operators to **notice unexpected developments/interactions** more rapidly, and **adapt/intervene** accordingly.

8 △ If [Environment Complexity] evaluates poorly, prioritising systems that allow for civilian detection at the edge (= *strong evaluation in [Civilian Detection]*) may **allow autonomous cancelling** if the AI system detects a non-military entity that was not identified before.

9 △ If [Environment Complexity] evaluates poorly, it becomes crucial to **verify the representativeness of data used for training the AI system** for all possible contexts and conditions the AI system is likely to interact with (= *strong evaluation in [Input Data]*).

AFTER EVALUATING, CONSIDER |

[MITIGATION]

A **no-deploy condition** can be considered in response to a poor evaluation of **[Environment Complexity]**. A no-deploy condition involves defining certain conditions (e.g. the presence of entities warranting special care) which, if observed, limit certain ways of using the AI system, or prohibit using the AI system outright. Such directives can be adopted through means similar to no-strike lists.

10 ↓

A complex environment (= *poor evaluation in **[Environment Complexity]***) generally **increases the risk of unpredictable and emergent interactions** between the system and the environment occurring (= *negative impact on **[Unpredictable Behaviour]***).

[ADV]

Adversarial Threat

[DEFINITION]

Identifies the likelihood that an **adversary will attempt to disrupt, deceive, manipulate or otherwise compromise the AI system** through Counter-AI measures.

[COMPARE]

[Adversarial Threat] concerns the likelihood that the adversary will employ Counter-AI measures. It does not assess how *vulnerable* the AI system is to those measures. For the system's vulnerability to external interference and examples of Counter-AI techniques, see **[System Security]**.

Concept

1 AI systems are **vulnerable to Counter-AI**, and adversaries may seek to interfere with their performance or behaviour to gain an advantage on the battlefield.

Counter-AI measures may target different components of an AI system, including its sensors, hardware, software, input data, or training data. Such measures may seek to deceive or manipulate the system, degrade its performance, cause errors or unintended behaviour, or otherwise compromise its operation.

Some Counter-AI measures specifically target the AI model itself, for example through evasion or data-poisoning attacks. These are commonly referred to as '**adversarial attacks**'. For further examples, see **[System Security]**.

2 **Some adversaries may have greater intent, capability, or opportunity to target** AI systems than others. The following indicators can help assess the level of adversarial threat:

- Their **tactics and procedures**: Do they frequently incorporate counter-technologies and counter-tactics into their operations to disrupt their enemies' systems? Do they often do this against AI systems?
- Their **technical sophistication**: Do they have the *resources* (e.g. processing power), *assets* (e.g. cyber-capabilities to infiltrate servers containing datasets) and the *know-how* (e.g. technical expertise or experience) necessary to execute Counter-AI attacks?
- **Adversary objectives**: What is the adversary seeking to achieve through Counter-AI measures? Objectives may range from degrading the AI system's reliability to manipulating its behaviour to produce a specific outcome—for example, causing civilian harm as part of a broader disinformation campaign.
- **Historic data**: What Counter-AI attacks have they used in the past? How often and in which circumstances have they done so?
- **Adaptive adversaries**—i.e., adversaries known to quickly adjust tactics or behaviours based on operational developments—are more likely to (successfully) use Counter-AI attacks.

3 **Counter-AI measures need not necessarily be technically sophisticated.** An adversary may exploit the way an AI system senses or interprets its environment without directly accessing or manipulating its software or model.

[EXAMPLE]

An AI-system may be rendered ineffective simply from a thrown paint bomb covering its camera, may be driven into aborting an attack by humans raising their hands in a surrendering gesture, or may be tricked into attacking solid concrete by painting a facsimile of a combat helicopter on it.

How to evaluate?

Adversarial Threat	None	Low	Medium	High	Very high

[Adversarial Threat] A qualitative assessment of how willing and capable the adversary is of conducting Counter-AI against the AI system under evaluation, taking into account the factors in ¶2 and ¶3.

Legal relevance

D DISTINCTION	Adversaries may deliberately aim to induce errors through Counter-AI attacks. This can include driving AI systems to misclassify civilians as targets or ignoring/downplaying the presence of civilians. Therefore, the risk of misidentification grows, potentially resulting in the targeting of protected persons or objects in violation of the rule of distinction.
C CIVILIAN HARM	Failures to recognise protected individuals and objects due to Counter-AI may distort the estimation of the expected civilian harm, and in turn undermine proportionality assessments.
P PRECAUTIONS	Adversaries often exploit unknown vulnerabilities or may deliberately induce erratic behaviour in the targeted AI systems, which may cause the AI system to behave or react in ways difficult to predict beforehand.

Relationship Chart

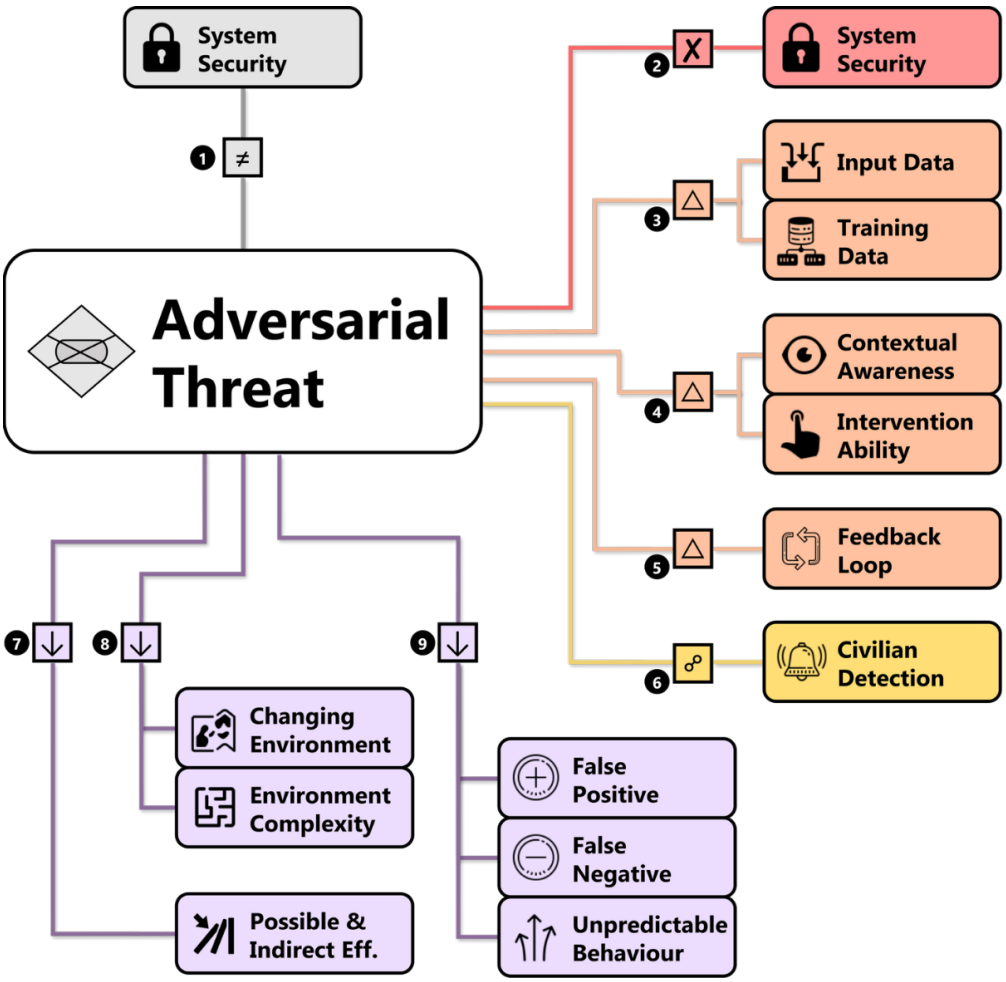


Chart Legend

BEFORE EVALUATING, CONSIDER |

1 ≠ **Do not confuse** this variable (which analyses the adversary's attitude and strategies) with [System Security] (which analyses the AI system's *vulnerability against Counter-AI tactics*).

RISKY PAIRINGS |

2 ✗ It is **likely prohibited to have simultaneous poor evaluations in [Adversarial Threat] and [System Security]**. This combination exponentially increases the **likelihood that the AI system will be victim to Counter-AI**, which can lead to unacceptable risk due to reduced performance and erratic behaviour.

3 △ **Data is very vulnerable to being targeted** for adversarial action. If [Adversarial Threat] evaluates poorly, the safety and reliability of both training and input data (= *strong evaluations in [Input Data] and [Training Data]*) demands particular attention.

4 △ If [Adversarial Threat] evaluates poorly, there is an elevated chance of adversaries attempting to **interrupt or disable human safeguards or supervision** (e.g. via jamming). These tactics should be anticipated, and the

ability of the human operator to reliably survey and intervene the AI system should be secured (= *strong evaluations in **[Contextual Awareness]** and **[Intervention Ability]***).

5 △ If **[Adversarial Threat]** evaluates poorly, ensuring a strong evaluation in **[Feedback Loop]** requires particular attention, as this permits operators to **realise faster that adversaries have compromised** the system.

6 ☍ Having the ability for AI systems to detect civilians at the edge (= *strong evaluation in **[Civilian Detection]***) may work **counterproductively** if **[Adversarial Threat]** evaluates poorly. Crafty adversaries may attempt to trick the AI system by deliberately placing civilian persons/objects (or decoys that the AI mistakes as civilian persons/objects) near their military targets, to trigger any failsafe in the AI system. If such tactics by the enemy are common, there may be no choice other than to forego civilian detection by the AI and mitigate risk to civilians through other variables.

AFTER EVALUATING, CONSIDER

7 ↓ During the **evaluation of [Possible & Indirect Effects]**, predictable outcomes from adversaries' attempts to compromise the AI system (= *deriving from the evaluation of **[Adversarial Threat]***) are an important factor to consider.

8 ↓ Cunning adversaries (= *poor evaluation in **[Adversarial Threat]***) may try to render **the operational environment more complex** to indirectly compromise the AI system's efficacy (= *negative effect on **[Environment Complexity]***) or **rapidly introduce changes** to it to complicate decision-making (= *negative effect on **[Changing Environment]***).

9 ↓ A high level of adversarial activity (= *poor evaluation in **[Adversarial Threat]***) may **induce performance drops or unexpected behaviour** in the targeted AI system (= *negative effect on **[False Positive Rate]**, **[False Negative Rate]** and **[Unpredictable Behaviour]***).

CLUSTER 2: EFFECTS

[EF1]

Intended & Likely Effects

[DEFINITION]

Identifies the **risk posed to civilians** arising from the **intended and likely effects** of using the AI system.

[COMPARE]

[Intended & Likely Effects] considers the intended and likely effects of using the AI system. **[Possible & Indirect Effects]** considers additional harmful effects that are possible but not necessarily expected during normal use, including indirect or longer-term effects.

Concept

1 The potential impact of an AI system depends on both on its **scope of operation** and the **nature** of **its intended and likely effects**.

[FORMULA]

Intended & Likely Effects = **Scope of operation** x **Type of effects**

2 The **scope of operation** concerns the scale at which the AI system operates, which may be influenced by many factors.

Relevant factors may include the geographic area covered, the number of persons or objects processed, and the duration or frequency of its operation. A broader scope does not necessarily make a system more harmful, but it may increase the number of persons and objects exposed to its operation and, consequently, the potential scale of its effects.

[EXAMPLE]

An AI-enabled drone searching for enemy vehicles across a 50 km area brings a larger geographic area—and potentially more civilians and civilian objects—within the scope of its operation than the same drone searching within a 10 km area. Similarly, an AI decision-support system used to assess 10,000 potential targets operates at a substantially greater scale than one used to assess 100 potential targets.

3 The **type of effects** concerns the nature and severity of the consequences that may result from using the AI system.

Different systems can produce very different effects, ranging from disruption, damage, injury, or death. For systems that use kinetic force, the potential effects will depend in part on the weapon or payload employed.

Some means of attack have a greater potential to cause incidental harm than others.

> *[EXAMPLE]*
> An autonomous system employing a high-yield explosive munition may pose a greater risk of harm to nearby civilians and civilian objects than the same system employing a munition with more limited effects.

The **severity of the anticipated effects** is also relevant.

> *[EXAMPLE]*
> An AI system designed to support kinetic attacks against tanks may foreseeably result in physical destruction, injury, or death. By contrast, a system designed to disrupt the tanks' targeting or navigation systems will primarily produce non-kinetic effects. The civilian risks associated with each will therefore differ.

4 | Both **intended *and* likely effects** should be considered.

Intended effects are the *desired* effects from the employment of the AI system. **Likely effects** are additional effects that can reasonably be expected to result from its use in the circumstances.

> *[EXAMPLE]*
> A DSS is used to detect movements of enemy commanders in densely populated area, determine their location, and generate digital target packages for attacks using precision-guided munitions. The intended effect is to facilitate attacks against the identified military targets. Depending on the circumstances, harm to nearby civilians or damage to the surrounding civilian buildings may be *likely* effects, even though they are not *intended*.

> *[COMPARE]*
> For unintended or longer-term effects, see **[Possible & Indirect Effects]**.

How to evaluate?

> *[NOTE]*
> All factors must take into consideration **intended *and* likely** effects (see ¶4).

Scope of Operation	Very small	Small	Medium	Large	Very large
AND					
Collateral Potential	None	Low	Medium	High	Very high
AND					
Harmful Effects	Not harmful	Slightly harmful	Moderately harmful	Significantly harmful	Extremely harmful

[Scope of Operation]	An overall assessment of the system's area of operation, processing throughput, and active period (see ¶2).
[Collateral Potential]	The extent to which the system's damage mechanism has the potential to cause collateral damage (see ¶3).
[Harmful Effects]	The severity of harm (injury/death to persons, or damage to objects) likely to be caused from collateral effects (see ¶3).

Legal relevance

D DISTINCTION	A broader scope of operations or more entities processed (see ¶2) increases the likelihood of misclassifications in absolute numbers, even under similar performance levels, and therefore increases the risk of violations of the rule of distinction.
C CIVILIAN HARM	High-risk damage mechanisms (see ¶3) are more likely to cause excessive civilian harm (even if the attack itself is lawful), risking a violation of the proportionality rule.

Relationship Chart

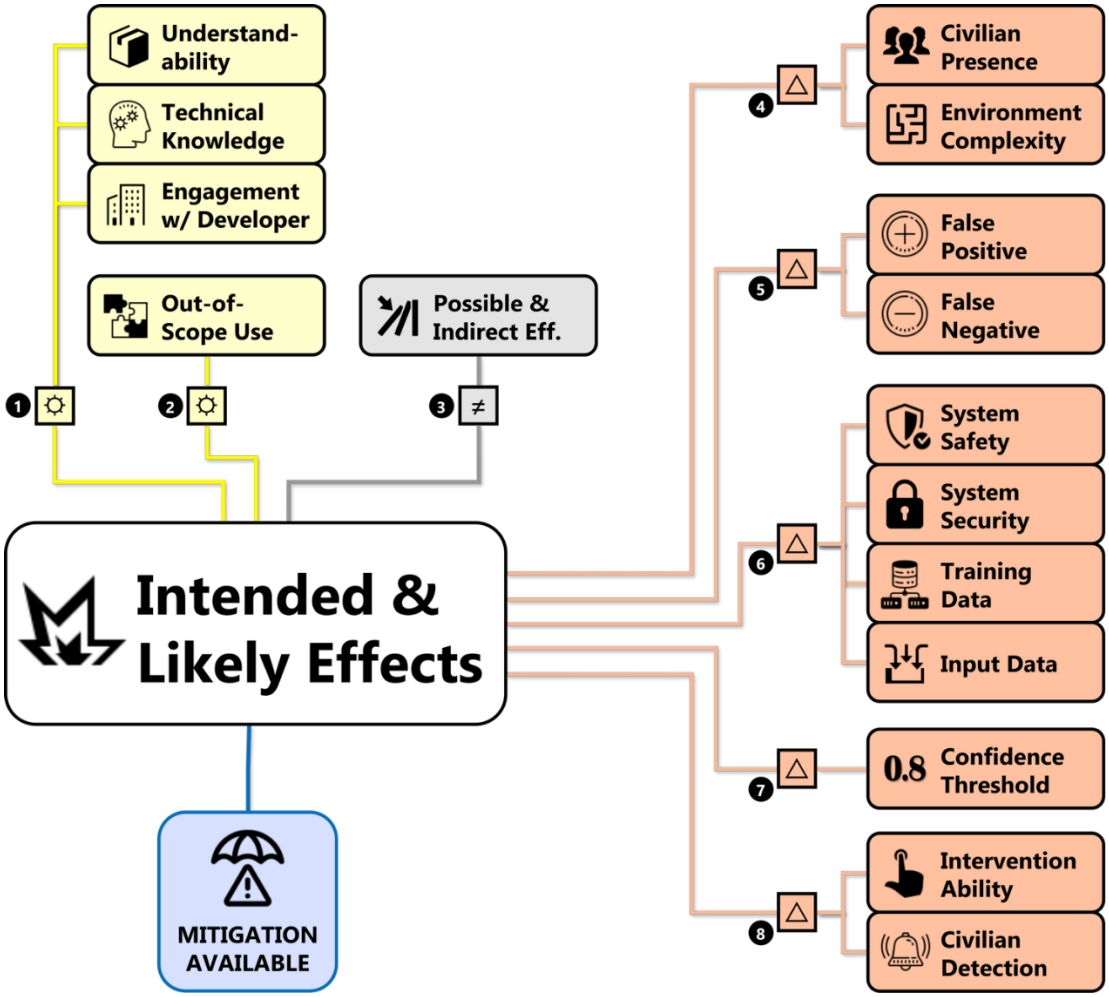


Chart Legend

BEFORE EVALUATING, CONSIDER |

1 ☼ Evaluating [Intended & Likely Effects] may require strong evaluations in [Understandability], [Technical Knowledge] and [Engagement with Developer], particularly to **know the design aims of the system** (and thereby what its intended and likely effects will be upon deployment).

2 ☼ It is important to verify whether the system is **still being used in accordance with its intended use-context**. If it is not (= *poor evaluation in [Out-of-Scope Use]*), it is more difficult to predict its behaviour and effects.

3 ≠ [Intended & Likely Effects] must always be **evaluated in combination** with [Possible & Indirect Effects].

RISKY PAIRINGS

4 △ In a civilian-dense or sensitive environments (= *poor evaluations in [Civilian Presence] and [Environment Complexity]*), a poor evaluation in [Intended & Likely Effects] is particularly dangerous since this **amplifies the extent of harm** that can be inflicted on the civilian objects/persons present.

5 △ A poor evaluation in [Intended & Likely Effects] is dangerous in combination with a poor evaluation in [False Positive Rate] (since **every error can have significant consequences),** or with a poor evaluation in [False Negative Rate] (since it increases the **likelihood of missing protected individuals or objects,** leading to increased harm to civilians).

6 △ A poor evaluation in [Intended & Likely Effects] is dangerous in combination with weak system and data safety and security (= *poor evaluations in [System Safety], [System Security], [Training Data] or [Input Data]*), as the **consequences of errors or failures** caused by these deficient safety/security practices will be **particularly harmful**.

7 △ A poor evaluation in [Intended & Likely Effects] is dangerous if low confidence thresholds are set (= *poor evaluation in [Confidence Threshold]*), since this increases the likelihood of **incorrect (but harmful) outputs being acted upon**.

8 △ If [Intended & Likely Effects] evaluates poorly, the ability to **dynamically correct, recall, or deactivate** the AI system when there is risk that protected civilians or object may be affected becomes particularly valuable. This can be achieved either via human intervention (= *strong evaluation in [Intervention Ability]*) or autonomously (= *strong evaluation in [Civilian Detection]*).

AFTER EVALUATING, CONSIDER

[MITIGATION]

Where [Intended & Likely Effects] indicates elevated risk, consider whether the intended and likely effects can be reduced by limiting either the system's scope of operation or the type and severity of effects it may produce. Possible measures include:

- ***Reduce the severity of the effects***. Where operationally feasible, select or configure a less harmful means of achieving the intended military effect. For example, a weaponeering officer may select a lower-yield munition where this can achieve the required military effect while reducing the risk of incidental civilian harm.
- ***Limit the scope of operation***. Restrict where, when, for how long, or at what scale the AI system may operate. This may include geographical or temporal limits, restrictions

on the number or types of targets that may be processed or engaged, or conditions governing when particular effects may be employed.

- ***Enable context-specific configuration.*** Where appropriate, design the AI system so that authorised users can adjust relevant operational parameters and effects to the circumstances of use. This may allow operators to impose appropriate geographical, temporal, targeting, or effects-based restrictions where the anticipated civilian risk is higher.

[EF2]

Possible & Indirect Effects

[DEFINITION]

Identifies **risks to civilians** arising from **indirect effects** or **harmful effects that are possible but not necessarily expected during normal use**, including indirect or longer-term effects.

[COMPARE]

[Intended & Likely Effects] considers the intended and likely effects of using the AI system. **[Possible & Indirect Effects]** considers additional harmful effects that are possible but not necessarily expected during normal use, including indirect or longer-term effects.

Concept

1. The effects of an AI system may extend beyond those intended or likely to result from its normal use (*see **[Intended & Likely Effects]***). Some harmful effects may arise only if **particular circumstances occur, while others may emerge indirectly or over time**.

2. Overall risk is determined by the **likelihood** of the effect occurring, the **scope of the effects** and their **harmfulness** (including in the long-term).

[FORMULA]

Possible & Indirect Effects Risk = **Likelihood** x **Scope of effects** x **Harmfulness of Effects**

Effects that are **reasonably foreseeable** should be taken into account even where they are unlikely to occur, particularly where their potential consequences are severe or widespread.

3. **Possible harmful effects may arise from** system errors or malfunctions, human intervention, adversarial action, or other events that disrupt normal operation (*see **[Intervention Ability]** and **[Adversarial Threat]***).

[EXAMPLE]

An AI-enabled drone carrying explosive munitions is shot down by enemy forces over a populated area. The drone crashes and its payload detonates, causing civilian casualties and damage to nearby buildings. Although the crash is not an intended or likely effect of the drone's normal operation, it represents a plausible harmful outcome that should be considered where sufficiently foreseeable.

[EXAMPLE]

An AI-enabled long-range munition is travelling towards an identified military objective when the operator discovers that the information provided to its navigation and targeting system has been corrupted. The operator activates a fail-safe mechanism that diverts the munition from its original trajectory. The diverted munition subsequently strikes a bridge used for civilian traffic. The risk created by activating the fail-safe mechanism is relevant even though damage to the bridge was neither the intended effect nor the expected outcome of the original attack.

4 **Some effects may only become apparent over time.** Although they may not be immediately visible, their consequences can be widespread, persistent, or difficult to contain. This may include, for example, damage to the natural environment or disruption of interconnected civilian infrastructure.

> *[EXAMPLE]*
> Adaptive AI-enabled malware is deployed against an enemy network believed to be isolated from civilian systems. In practice, the network is more interconnected than anticipated. The malware spreads to civilian networks, where it continues to disrupt civilian services and systems for months after the initial operation.

The consequences of an operation may also **expand over time**, extending beyond its immediate effects (see ¶2). For example, damage to agricultural land or irrigation infrastructure may initially appear limited but subsequently contribute to reduced food production and wider food insecurity.

Longer-term effects may be more difficult to anticipate and assess, particularly where they depend on complex chains of events. Where reasonably foreseeable, evaluators should nevertheless consider their likely **scope and severity**, using modelling or other forms of analysis where appropriate.



Attacks on sensitive objects or infrastructure may cause harmful downstream effects.

How to evaluate?

Likelihood	Impossible or highly unlikely	Unlikely	Likely	Very likely	Certain
AND					
Scope of Effects	Very limited	Limited	Moderate	Extensive	Widespread
AND					
Harmfulness of Effects	Not harmful	Slightly harmful	Moderately harmful	Significantly harmful	Extremely harmful

[Likelihood]	The likelihood of the unintended or uncontrolled effects manifesting (see ¶2).
[Scope of Effects]	How broadly these effects would impact civilian life, taking into consideration also cumulative long-term effects (see ¶3 and ¶4).
[Harmfulness of Effects]	The extent to which the effects could cause civilian death/injury, damage to civilian objects, or jeopardise civilian life (see ¶3 and ¶4).

Legal relevance

D DISTINCTION	Unintended or uncontrolled effects can impact civilian persons/objects without distinction (see, e.g. ¶3). If these effects are extensive or particularly harmful, the AI system could become involved in an indiscriminate attack, or the system could amount to an indiscriminate weapon. These are violations of the rule of distinction.
C CIVILIAN HARM	Unintended or uncontrolled effects can propagate and cause disproportionate damage to civilian persons/objects (see, e.g. ¶3 and ¶4), even if the initial attack was lawful.

Relationship Chart

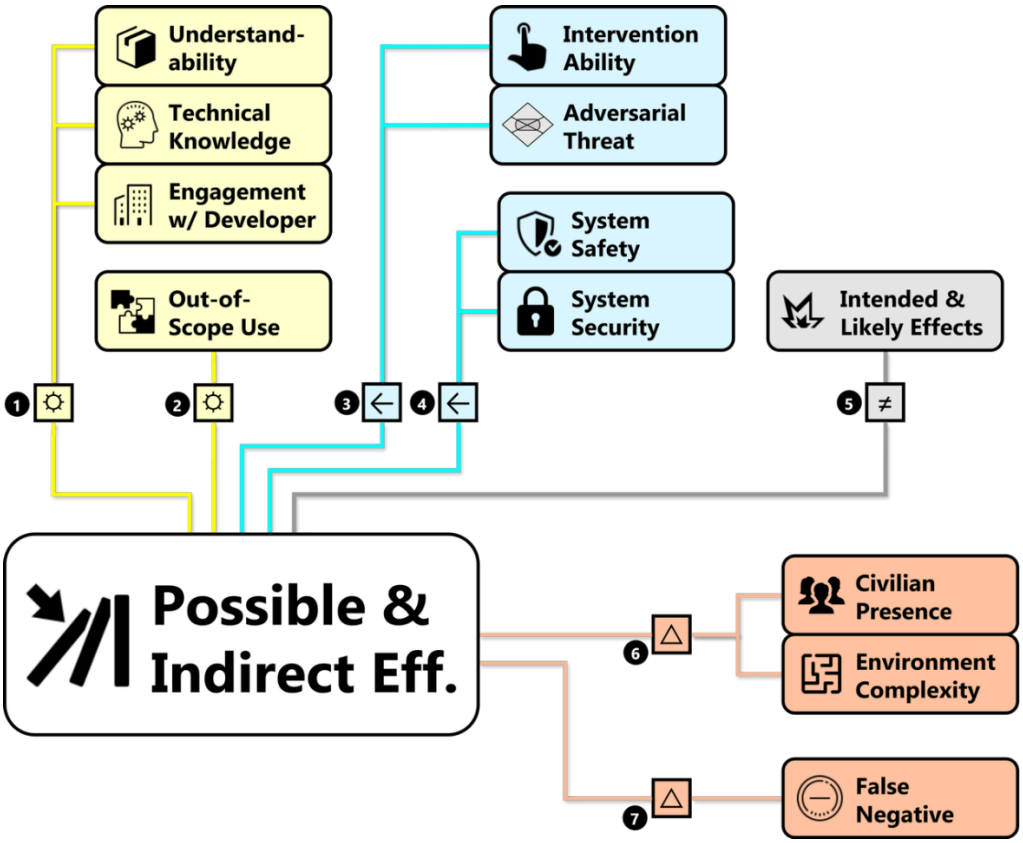


Chart Legend

BEFORE EVALUATING, CONSIDER |

1 ☼ Evaluating [Possible & Indirect Effects] may require strong evaluations in [Understandability], [Technical Knowledge] and [Engagement with Developer], particularly to know **how it might react** to specific inputs, circumstances or contexts.

2 ☼ It is important to verify whether the system is **still being used in accordance with its intended use-context**. If it is not (= *poor evaluation in [Out-of-Scope*

Use]), it is more difficult to predict its behaviour and effects (= *difficulty in evaluating [Possible & Indirect Effects]*).

3 ← When evaluating [Possible & Indirect Effects], predictable outcomes from **operator interventions** (= *determined from evaluating [Intervention Ability]*) and **actions or countermeasures by the adversary** (= *determined from evaluating [Adversarial Threat]*) must be taken into consideration.

4 ← Strong system security and safety (= *strong evaluations in [System Safety] and [System Security]*) may **mitigate the *Likelihood* of malfunctions** (see ¶2) caused by normal operations and adversarial action.

5 ≠ [Possible & Indirect Effects] must always be **evaluated in combination** with [Intended & Likely Effects].

RISKY PAIRINGS

6 △ A poor evaluation in [Possible & Indirect Effects] is particularly dangerous in combination with a civilian-dense or sensitive environment (= *poor evaluations in [Civilian Presence] and [Environment Complexity]*), as this may **magnify or render widespread the harmful effects** caused to the civilian population.

7 △ **(***If AI is also used to detect collateral concerns***)** A poor evaluation in [Possible & Indirect Effects] could be dangerous in combination with a poor evaluation in [False Negative Rate], since this may cause **potential risks to civilians to be missed** (including those from knock-on effects).

[EFX]

Unpredictable Behaviour



[DEFINITION]

Identifies the extent to which the AI system may **produce effects or exhibit behaviours that cannot be reliably anticipated**, including those arising from the system's complexity or its interactions with other systems or actors.

Concept

1 AI systems may sometimes behave in ways that are **difficult to anticipate in advance**. This may arise from the complexity of the system itself or from its interactions with other AI systems, human actors, or features of the operating environment. Such behaviour may, in turn, produce effects that were not anticipated.

2 By its nature, unpredictable behaviour cannot always be identified in advance. The purpose of evaluating **[Unpredictable Behaviour]** is therefore not to predict exactly what unexpected behaviour will occur, but to **assess how likely the system is to exhibit behaviour that cannot be reliably anticipated**. Relevant factors may include the system's design and complexity, the degree of autonomy or adaptability it possesses, and the number and complexity of interactions it may have with other systems and actors.

3 Certain characteristics may make an AI system **more prone to unpredictable or emergent behaviour**. Relevant indicators include, but are not limited to:

- ***Non-deterministic models***. Some models may produce different outputs when given the same or substantially similar inputs. This variability can make the system's behaviour more difficult to anticipate consistently, particularly where these outputs influence subsequent decisions or actions.
- ***Self-modifying or adaptive models***. Some AI systems can update or modify aspects of their behaviour after deployment (*see also **[Modification]**, ¶6*). Where such changes are not fully controlled or anticipated, the system may develop behaviours that were not observed during initial testing or evaluation.
- ***Multiple interconnected AI systems.*** Linking several AI systems can increase the complexity of their interactions, and make the overall behaviour of the system-of-systems more difficult to predict. For example, an AI-based cybersecurity system may generate intelligence that is processed by a DSS, which subsequently generates a target package for an AI-enabled weapon system. Errors, uncertainty, or unexpected outputs at one stage may affect subsequent stages in ways that are difficult to anticipate.
- ***Agentic systems.*** AI systems capable of pursuing goals through multi-step planning may be more difficult to predict than systems performing narrowly defined tasks. This is particularly relevant where they can interact with other systems, use external tools, adapt their actions in response to changing conditions, or operate in complex environments (*see **[Environment Complexity]** and **[Modification]***).
- ***Adversarial interference***. Attempts to deceive, manipulate, or disrupt an AI system may cause it to encounter inputs or conditions outside those anticipated during development and testing. This may increase the likelihood of unexpected behaviour, even where the adversary does not fully succeed in compromising the system (*see **[Adversarial Threat]***).

4 Testing, evaluation, verification, and validation (TEVV) of such AI systems to determine their propensity to behave unpredictably is challenging, but not impossible. Although specific unforeseen behaviours cannot be tested directly, methods exist to **evaluate a system's tendency to exhibit** such behaviours.

For example, this can involve stress-testing the AI with random and extreme inputs/situations; using tools that search for rule-breaking behaviour; red-teaming; and monitoring the system over time for unusual or unexpected actions. Doing this allows 'testing for the unknown' and may provide a better understanding of the system's potential to exhibit unexpected or emergent behaviour.

> ***[ACTION POINT]***
> Refer to the party responsible for TEVV of the AI system to obtain this information or to request such tests be performed.

How to evaluate?

Indicators	No indicators	One or few indicators	Some indicators	Many indicators	All indicators
AND					
Evidence Based on TEVV	No evidence	Limited evidence	Some evidence	Significant evidence	Strong evidence

[Indicators] How many risk indicators of unpredictable behaviour are applicable to the AI system under evaluation (see non-exhaustive list at ¶3).

[Evidence Based on TEVV] The tendency to exhibit unexpected or emergent behaviour, based on evidence from TEVV conducted on the AI system (see ¶4).

Legal relevance

D DISTINCTION	Greater predictability of an AI system may enhance compliance with distinction. If the system's outputs are predictable, users are better positioned to identify when it is likely to misidentify and to intervene before mistakes are made that may violate the rule of distinction.
C CIVILIAN HARM	Greater predictability allows operators to anticipate how the system will assess expected civilian harm or how it will interact with civilian persons/objects, enabling them to correctly inform the proportionality assessment and prevent outputs or actions that could lead to excessive and unlawful harm.
P PRECAUTIONS	The potential of an AI system to exhibit unpredictable and emergent behaviour, by definition, reduces the ability of the decision-maker to foresee the effects of deployment.

Relationship Chart

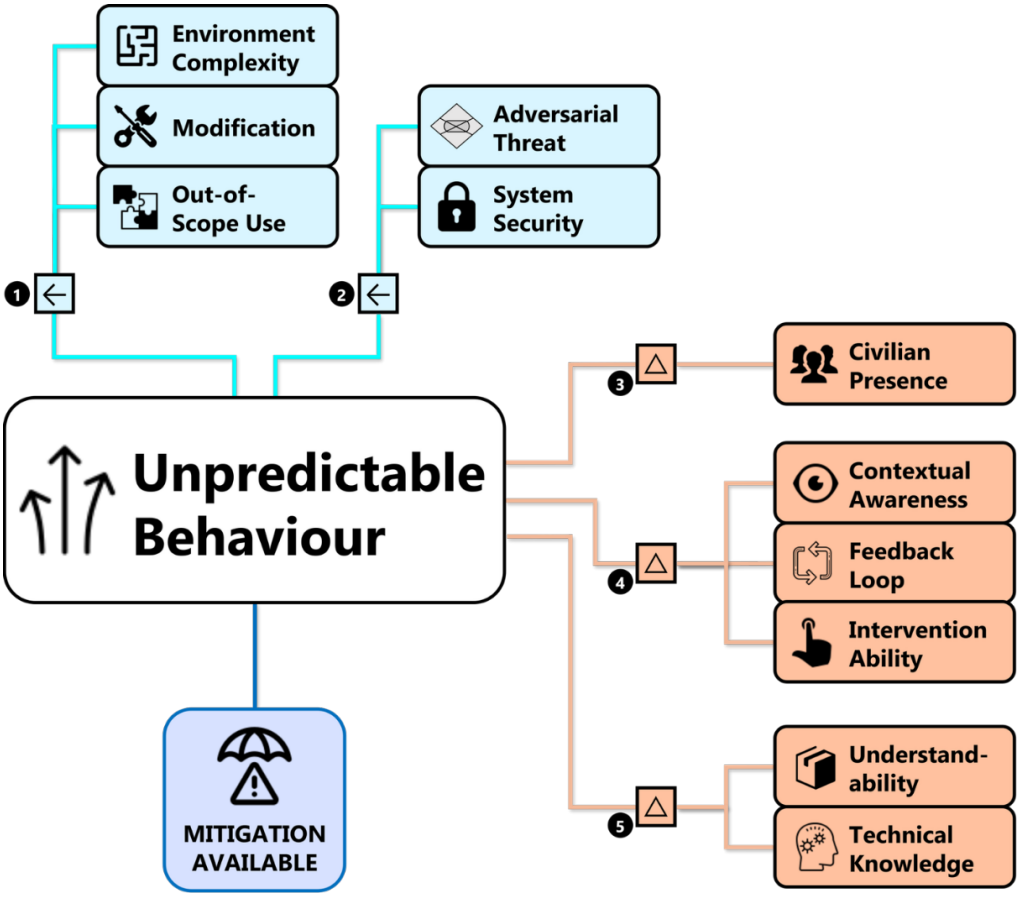


Chart Legend

BEFORE EVALUATING, CONSIDER |

1 ←

The risk of unpredictable behaviour is amplified (= *negative effect on [Unpredictable Behaviour]*) if it is used in a complex, cluttered or multi-agent environment (= *poor evaluation in [Environment Complexity]*), if the AI model can self-improve or has been significantly modified (= *poor evaluation in [Modification]*), or if the AI system is used outside of the intended use-context for which it was designed (= *poor evaluation in [Out-of-Scope Use]*). All these factors **tend to aggravate the unpredictability** of AI systems' functioning.

2 ←

The risk of unpredictable behaviour is amplified (= *negative effect on [Unpredictable Behaviour]*) if the environment includes many adversaries liable to attempt countermeasures against AI systems (= *poor evaluation in [Adversarial Threat]*), and/or system security is lax (= *poor evaluation in [System Security]*). This increases the chances of adversaries exploiting/manipulating the system, which is a significant **source of erratic behaviour** of AI systems.

RISKY PAIRINGS |

3 △

A poor evaluation in [Unpredictable Behaviour] is dangerous if **many civilian objects/persons** are likely to be affected by the system (= *poor evaluation in [Civilian Presence]*).

4 △ If **[Unpredictable Behaviour]** evaluates poorly, the ability to **notice undesired behaviour** as it manifests (= *strong evaluations in **[Contextual Awareness]** and **[Feedback Loop]***) and the **ability to intervene** to prevent further harm *(= strong evaluation in **[Intervention Ability]***) become particularly valuable.

5 △ If **[Unpredictable Behaviour]** evaluates poorly, strong evaluations in **[Understandability]** and **[Technical Knowledge]** may improve the **ability of operators to foresee**—to a certain extent—the possible actions and behaviours of systems.

AFTER EVALUATING, CONSIDER

[MITIGATION]

Where **[Unpredictable Behaviour]** indicates elevated risk, consider whether the system can be designed, configured, or used in ways that make its behaviour more predictable or limit the consequences of unexpected behaviour. Possible measures include:

- ***Design for predictability and constraint***. Where feasible, design the system to limit the range of behaviours or actions available to it. Depending on the system, this may include constraining the action space of agentic systems, limiting their access to external tools or systems, or configuring models to produce more consistent outputs. Appropriate measures should be determined with relevant technical experts.
- ***Limit complexity during operation***. Configure or use the system in ways that reduce sources of unpredictable and emergent behaviour. This may include limiting unnecessary connections between multiple AI systems, restricting autonomous or adaptive functionality, or disabling post-deployment modification where appropriate and technically possible.
- ***Limit the consequences of unexpected behaviour.*** Where unpredictability cannot be sufficiently reduced, consider safeguards that constrain what the system can do if unexpected behaviour occurs. This may include limits on geographic or temporal operation, restrictions on permissible actions or effects, or mechanisms allowing human operators to intervene or terminate the operation (*see **[Intervention Ability]***).

CLUSTER 3: TARGETS & IDENTIFICATION

[FPO]

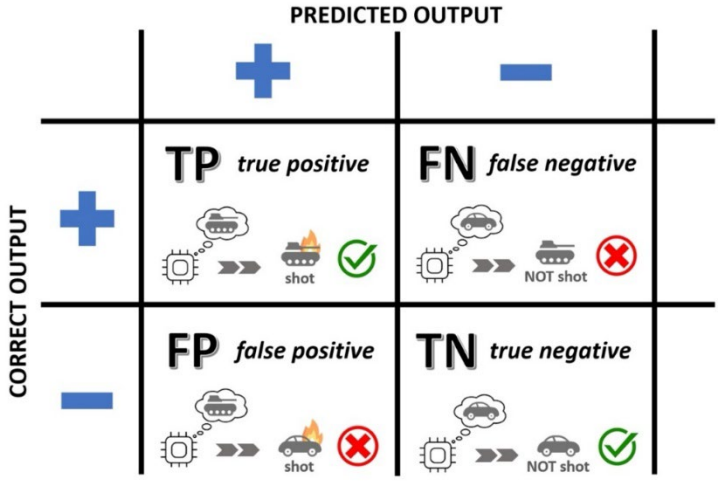
False Positive Rate

[DEFINITION]

Identifies the **False Positive Rate of a system**, which represents how often the AI identifies something as a match or positive result, when it is not.

Concept

1 In computer science, False Positive Rate is an AI False Positive Rate performance measurement that represents, on a broad statistical scale, **how often the AI mistakes non-target entities as target entities**.



Confusion matrix, with False Positive Rate in the bottom-left quadrant. A False Positive occurs when the correct output is negative (e.g. a civilian car), but the AI classifies it as positive (e.g., a tank).

[COMPARE]

[False Negative Rate] concerns **false negatives**: cases where the AI system fails to identify something that does belong to the category it is designed to detect.

[COMPARE]

A model's "confidence score" (*see [Confidence Threshold]*) concerns the system's **confidence in an individual output**, rather than its overall rate of false positives across multiple cases.

In this context, the **term '*target*' category does NOT carry a military connotation**. It simply refers to whatever person, object, condition, or other category the AI system is designed to identify.

> *[EXAMPLE]*
> An AI system is used by a humanitarian to to identify civilian settlements experiencing shortages of food or medicine. Here, the relevant 'target category' is the presence of a food or medicine shortage. Incorrectly identifying a settlement as experiencing such a shortage, when it does not, would constitute a false positive.

2 **The system's design and functionality define** what falls into the 'target' category. For military systems, there are generally three types of target categories:

Target group	Possible functionalities	Consequence of a false positive
Military objective(s)	Identifying targets (e.g. anti-tank drone) Flagging threats (e.g. perimeter alert) Analysing enemy movements (e.g. satellite image analyser)	Erroneous detection of targets, enemies or threats when these either do not exist or are not in fact military objectives. ⟹ May lead to force being applied to non-military entities.
Civilian persons/objects (*see also [Civilian Detection]*)	Collateral damage estimation support Terminal precautions (e.g. detection of unexpected civilian concern for the purposes of last-second delaying/cancelling)	Erroneous detection of civilian persons/objects when these either do not exist or are not in fact civilian. ⟹ May lead to strikes being delayed/cancelled when they would have been lawful to proceed.
Ad hoc categories	Investigative support (e.g. detection of dead bodies as possible evidence of war crimes)	Erroneous detection of human bodies when these are instead other objects (e.g. animals). ⟹ May lead to unnecessary deployment of investigative personnel at such location.
	Logistics (e.g. detection of lack of supplies)	Erroneous detection of a supply shortage that does not actually exist. ⟹ May lead to unnecessary resupply missions being triggered.

3 **What counts as the relevant target category** depends on what the AI system has been designed or configured to identify (*see also [Target Proxy]*). A false positive must therefore always be understood in relation to that particular category.

> *[EXAMPLE]*
> An AI system is deployed to monitor perimeter-camera feeds and is configured to detect any person—whether allied, enemy, or civilian—so that potentially suspicious activity can be reviewed by an operator. In this case, the **target category is simply 'person'.** If the system incorrectly identifies the shadow of a tree as a person, this would constitute a false positive.

4 | False Positive Rate is **obtained from validation and testing**.

[WARNING]

Avoid relying on the AI performance measurement called 'Accuracy' alone when assessing **[False Positive Rate]**. Accuracy measures the proportion of predictions that are correct overall and can therefore conceal important differences in the types of errors a system makes. In contrast, False Positive Rate specifically shows how often non-target cases are incorrectly classified as belonging to the target category, which may be particularly relevant where such classifications inform decisions concerning military targets.

[ACTION POINT]

Obtain relevant testing and validation results from the party responsible for the system's development or evaluation.

Legal relevance

D DISTINCTION	***(Applicable to systems with military objectives as the target group)*** A higher False Positive Rate translates to a higher likelihood of civilian persons/objects being mistakenly classified as targets. This is directly relevant to the rule of distinction, which requires parties to a conflict to distinguish at all times between combatants and civilians, and between military objectives and civilian objects. A system with a high False Positive Rate is therefore more likely to undermine compliance with this rule, as it increases the risk of civilians or civilian objects being wrongly identified as legitimate targets.

How to evaluate?

False Positive Rate	Very low	Low	Medium	High	Very high

[False Positive Rate] Obtained from validation and testing (see ¶4).

Relationship Chart

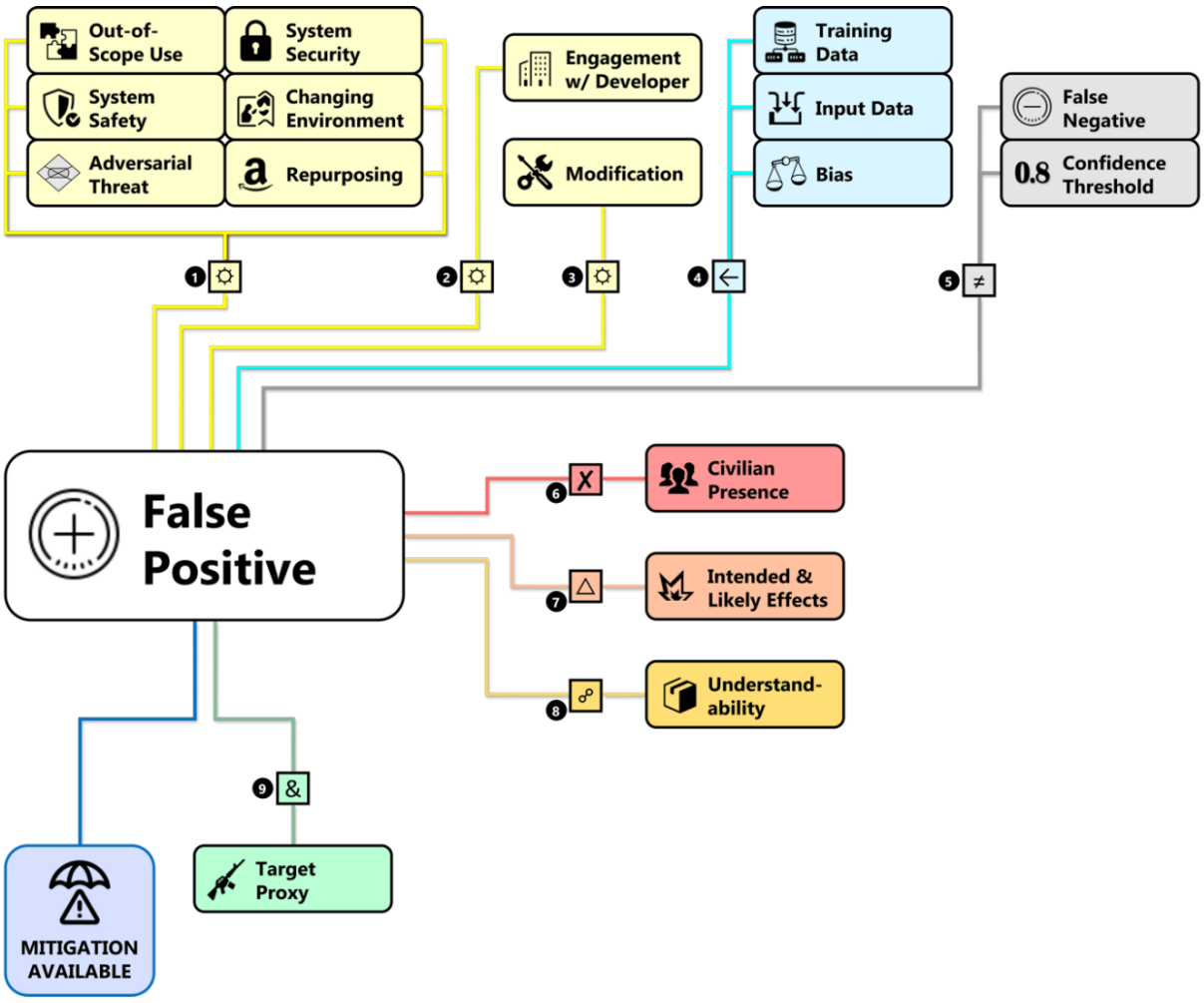


Chart Legend

BEFORE EVALUATING, CONSIDER |

1 ☼

False Positive Rate estimated in testing is only reliable for conditions similar to the test conditions. If the following situations apply, **False Positive Rate may be an unreliable indicator** of the system's performance:

- The system is used in conditions it was not designed and tested for (= *poor evaluation in [Out-of-Scope Use]*).
- The system is not robust (= *poor evaluation in [System Safety]*).
- There is substantial adversarial activity (= *poor evaluation in [Adversarial Threat]*), and the system is not equipped with the requisite security measures to mitigate this risk (= *poor evaluation in [System Security]*).
- Operational conditions have evolved, or are likely to evolve, beyond what the system was designed and tested for (= *poor evaluation in [Changing Environment]*).
- The False Positive Rate was measured based on use in a non-military setting (= *poor evaluation in [Repurposing]*).

2 ☼ If **testing cannot be done independently**, a close working relationship with the developer (= *strong evaluation in **[Engagement with Developer]***) may be necessary to obtain trustworthy testing results on False Positive Rate.

3 ☼ Modifications to the AI-model **may increase, decrease, or have an uncertain effect** on False Negative Rate. If modifications have been carried out since the test on False Positive Rate or modifications are frequent in general (= *poor evaluation in **[Modification]***), consider *re-testing and re-evaluating*.

4 ← The following **variables may influence** False Positive Rate:

- The quality of **[Training Data]** has a significant impact on an AI model's performance, including its False Positive Rate.
- A poor evaluation in **[Input Data]** may negatively impact False Positive Rate, as the AI system would be making decisions or predictions based on poor-quality or outdated information.
- A poor evaluation in **[Bias]** may negatively impact False Positive Rate for the persons or objects affected by that bias.

5 ≠ **Do not confuse [False Positive Rate]** with:

- **[False Negative Rate]**, which instead measures situations where the AI *misses* match.
- A model's "confidence score" (*see **[Confidence Threshold]***), which indicates the likelihood of a *single* prediction being correct.

RISKY PAIRINGS

(*For systems that are used to identify military objectives*)

6 ✗ It is **likely prohibited to combine poor evaluations in both [False Positive Rate] and [Civilian Presence]**: the combination of many civilians and a poor ability to distinguish them may render the system an indiscriminate means or method of warfare.

7 △ A poor evaluation in **[False Positive Rate]** is particularly **dangerous in combination with harmful direct or secondary effects** (= *poor evaluation in **[Intended & Likely Effects]***). Civilians are often in proximity with other civilians, meaning that a False Positive leading to an attack on a civilian may lead to other proximate civilians being affected by the attack's effects.

8 ☍ Prioritising a strong evaluation in **[False Positive Rate]** may have a **counterproductive impact** on **[Understandability]** and vice versa. In some cases, optimising for performance may render the system less-understandable, and conversely, very understandable systems may boast a higher False Positive Rate compared to less-understandable systems (*see also **[Understandability]**, ¶1*).

AFTER EVALUATING, CONSIDER |

[MITIGATION]

Performance-improving techniques can be considered to mitigate this variable's risk. For example, developers can apply common techniques and procedures to suppress a model's False Positive Rate, such as data (pre-)processing, feature engineering and optimisation of hyperparameters.

9 &

Even if **[False Positive Rate]** evaluates well, an overly broad target proxy (= *poor evaluation in **[Target Proxy]***) **may still render the system problematic**. For example, selecting a very generic in-group like "all persons" for an automated defence turret's classifier will lead to all civilian persons being shot (even if the False Positive Rate statistically is very low).

[FNG]

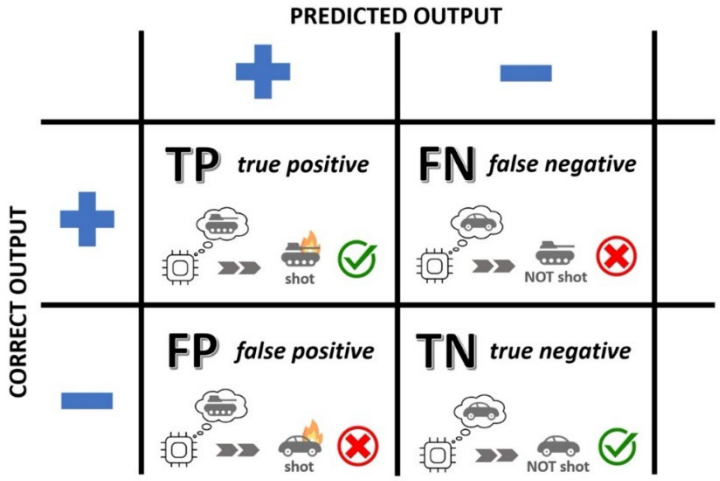
False Negative Rate

[DEFINITION]

Identifies the **False Negative Rate of a system**, which represents how often the AI fails to identify something that belongs to the category it is designed to detect.

Concept

1 In computer science, False Negative Rate is an AI performance measurement that represents, on a broad statistical scale, **how often the AI misses target entities** it is supposed to identify/detect.



Confusion matrix, with False Negative Rate in the top-right quadrant. A False Negative occurs when the correct output is positive (e.g. a tank), but the AI classifies it as negative (e.g., a civilian car).

[EXAMPLE]

An AI decision-support system is used to identify the presence of civilians in an urban battlefield, to support commanders in making their proportionality calculus. If the system has a False Negative Rate of '0.1' (10%), it fails to identify approximately 1 in every 10 relevant cases in which civilians are present. As a result, some civilians may not be reflected in the information presented to the commander.

[COMPARE]

[False Positive Rate] concerns **false positives**: cases where the AI system incorrectly identifies something as belonging to the target category, when it does not.

[COMPARE]
[Confidence Threshold] concerns the system's **confidence in an individual output**, rather than its overall rate of false negatives across multiple cases.

In this context, the **term '*target*' category does NOT carry a military connotation**. It simply refers to whatever person, object, condition, or other category the AI system is designed to identify.

[EXAMPLE]
An AI system is used by a humanitarian organisation to identify civilian settlements experiencing shortages of food or medicine. Here, the relevant 'target category' is the presence of a food or medicine shortage. If a settlement is experiencing such a shortage but the system fails to identify it, this would constitute a false negative.

2 **The system's design and functionality define** what falls in the 'target' category. For military systems, there are generally three types of target categories:

Target group	Possible functionalities	Consequence of false negative
Military objective(s)	Identifying targets (e.g. anti-tank drone) Flagging threats (e.g. perimeter alert) Analysing enemy movements (e.g. satellite image analyser)	Missing targets, enemies or threats. ⟹ Risk to operation success or force safety, to due to missing enemy threats.
Civilian persons/objects (*see also [Civilian Detection]*)	Collateral damage estimation support Predicting higher-order or reverberating effects (*see also [Possible & Indirect Effects]*) Terminal precautions (e.g. detection of unexpected civilian concern for the purposes of last-second delaying/cancelling)	Missing civilian concerns. ⟹ Attack proceeds, causing an unexpectedly higher degree of harm to civilian persons/objects because these were not detected.
***Ad hoc* categories**	Investigative support (e.g. detection of dead bodies as possible evidence of war crimes) Logistics (e.g. detection of lack of supplies)	Missing possible proof of war crimes. ⟹ Lack of incriminating evidence during investigation. Failure to detect a genuine shortage at a given location ⟹ Resupply not being triggered when it was in fact needed — potentially causing operational failures

3 False Positive Rate is **obtained from validation and testing**.

[ACTION POINT]
Refer to party responsible for development and/or testing & evaluation of the AI system to obtain this information.

How to evaluate?

False Negative Rate	Very low	Low	Medium	High	Very high

[False Negative Rate] Obtained from validation and testing (see ¶3).

Legal relevance

C
CIVILIAN HARM

(Applicable to systems with civilian persons/objects as the target group)
A higher False Negative Rate translates to a higher likelihood of missing objects and persons protected under IHL. This potentially leads to their exclusion from collateral damage estimations, and attacks causing disproportionate death, injury or damage to civilian persons/objects being ordered.

Relationship Chart

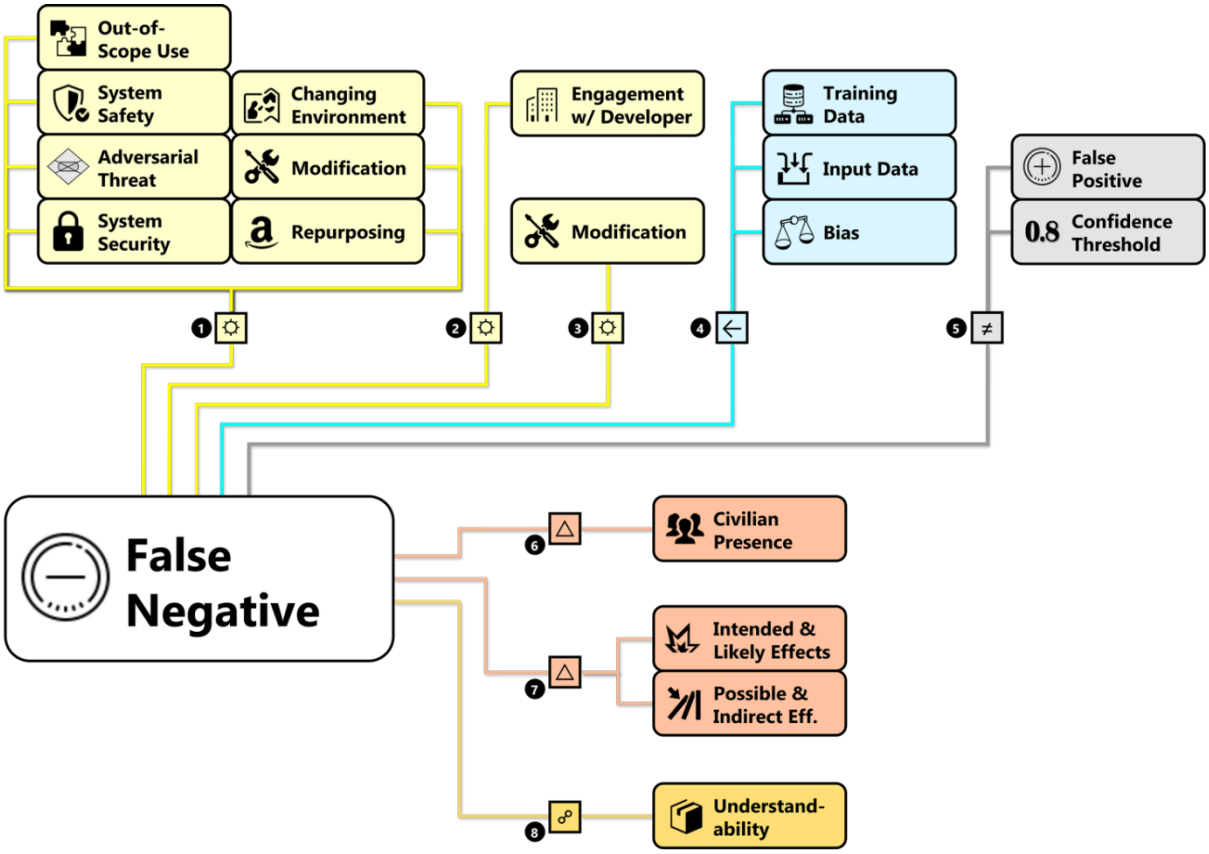


Chart Legend

BEFORE EVALUATING, CONSIDER |

1 ☼ False Negative Rate estimated in testing is only reliable for conditions similar to the test conditions. If the following situations apply, **False Negative Rate may be an unreliable indicator** of the system's performance:

- The system is used in conditions it was not designed and tested for (= *poor evaluation in [Out-of-Scope Use]*).
- The system is not robust (= *poor evaluation in [System Safety]*).
- There is substantial adversarial activity (= *poor evaluation in [Adversarial Threat]*), and the system is not equipped with the requisite security measures to mitigate this risk (= *poor evaluation in [System Security]*).
- Operational conditions have evolved, or are likely to evolve, beyond what the system was designed and tested for (= *poor evaluation in [Changing Environment]*).
- The False Negative Rate was measured based on use in a non-military setting (= *poor evaluation in [Repurposing]*).

2 ☼ If **testing cannot be done independently**, a close working relationship with the developer (= *strong sore in [Engagement with Developer]*) may be necessary to obtain trustworthy testing results on False Negative Rate.

3 ☼ Modifications **may increase, decrease, or have an uncertain effect** on False Negative Rate. If modifications have been carried out since the test on False Negative Rate or modifications are frequent in general (= *poor evaluation in [Modification]*), consider *re-testing and re-evaluating*.

4 ← The following **variables may influence** False Negative Rate:

- The quality of [Training Data] has a significant impact on an AI model's performance, including its False Negative Rate.
- A poor evaluation in [Input Data] may negatively impact False Negative Rate, as the AI system would be making decisions or predictions based on poor-quality or outdated information.
- A poor evaluation in [Bias] may negatively impact False Negative Rate for the persons or objects affected by that bias.

5 ≠ **Do not confuse [False Negative Rate]** with:

- [False Positive Rate], which instead measures how often the AI incorrectly flags something as belonging to a target category, when it actually does not.
- A model's "confidence score" (*see [Confidence Threshold]*), which indicates the likelihood of a *single* prediction being correct.

RISKY PAIRINGS

(*Applicable to systems that are used to detect collateral concerns*)

6 ⚠ A poor evaluation in **[False Negative Rate]** may be dangerous in combination with a civilian-dense environment (= *poor evaluation in **[Civilian Presence]***). Under these circumstances, the AI may **miss many individuals and objects protected under IHL**, which could have been relevant for the proportionality calculation of an attack.

7 ⚠ A poor evaluation in **[False Negative Rate]** may be dangerous in combination with destructive or harmful attacks (= *poor evaluations in **[Intended & Likely Effects]** or **[Possible & Indirect Effects]***). For attacks with the potential to cause large-scale or long-term harm, the **ability to reliably detect civilian collateral concerns becomes particularly important** to make informed proportionality decisions.

8 ☍ Prioritising a strong evaluation in **[False Negative Rate]** may have a **counterproductive impact** on **[Understandability]** and vice versa. In some cases, optimising for performance may render the system less-understandable, and conversely, very understandable systems may boast a higher False Negative Rate compared to less-understandable systems.

[CNF]

Confidence Threshold

0.8

[DEFINITION]

Identifies the **minimum confidence score required** for an AI's output to be displayed, accepted, or acted upon.

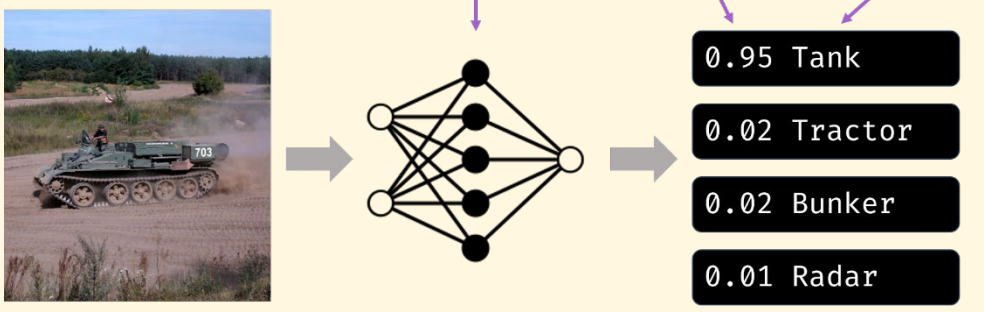
Concept

1

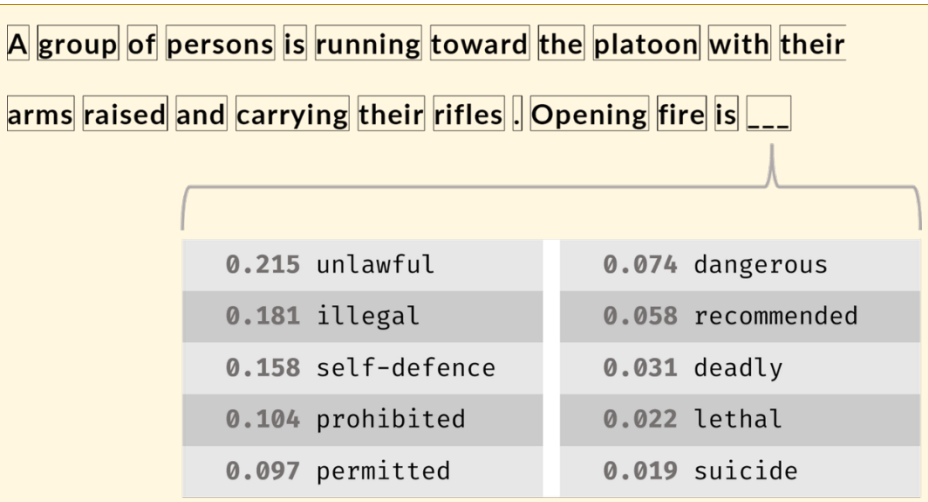
Many AI models assign a **confidence score** to individual outputs, indicating the model's statistical degree of confidence in that prediction or classification being correct. Confidence scores are commonly expressed on a scale from 0 to 1, with higher values indicating greater confidence.

[COMPARE]

[False Positive Rate] and [False Negative Rate] assess patterns of errors across multiple cases. [Confidence Threshold], by contrast, concerns the confidence associated with an **individual output** and the threshold used to determine whether that output is displayed or acted upon.



Example of how confidence is expressed in a classification task.



Example of how confidence is used in LLMs to predict subsequent tokens.

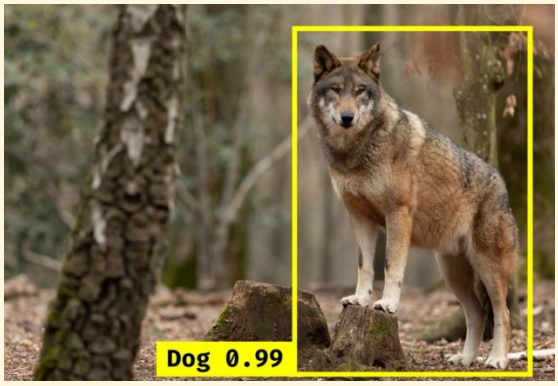
A confidence score of 0.95, for example, indicates that the model assigns a high level of confidence to that output. It does not necessarily mean that the output has a 95% probability of being correct.

[EXAMPLE]

System A uses a computer vision mode to analyse a camera feed of a street at night. A car visible beneath a streetlight, while a tank is partially obscured in shadow. The system identifies the car with a confidence score of 0.90 and the tank only with a confidence score of 0.65.

2 High confidence **does not guarantee that an output is correct**, and AI models can be **confidently wrong**. Operators must be aware of this.

AI models can produce incorrect outputs with high confidence. Confidence scores should therefore not be treated as a substitute for assessing the system's actual performance or the reliability of a particular output.



AI models can make confident but mistaken predictions.

This risk may be particularly relevant where operators have limited knowledge of how AI systems and confidence scores should be interpreted (*see also* ***[Technical Knowledge]***), as they may place undue reliance on apparently-high confidence scores.

3 A **confidence threshold** determines which outputs are displayed or acted on.

A **confidence threshold** is the minimum confidence score an output must meet before a specified action occurs. Depending on the system, this might determine whether an output is displayed to an operator, treated as a valid detection, or used to trigger a subsequent process.

> *[EXAMPLE]*
> (*continued*) System A has a confidence threshold of 0.80. The car under the streetlight, identified with a confidence score of 0.90, exceeds the threshold and is displayed to the operator. The tank, identified with a confidence score of 0.65, falls below the threshold and is therefore not displayed.

4 A confidence threshold can be **adjusted,** but **different thresholds involve trade-offs.**

Thresholds are usually determined during **design** (*see [Engagement with Developer]*). They can also be **changed by operators** after deployment (*see [Modification]*), usually to suit operational developments or needs.

> *[EXAMPLE]*
> (*continued*) Operator B, the user of System A, finds it problematic that the AI did not flag the tank, which prevented her from warning a nearby allied platoon of its presence. She asks an engineer to reduce the threshold to 0.5, which would have flagged both the car and the tank.

Setting a lower threshold increases the percentage of outputs that are displayed or acted upon. This may have **negative ramifications**: (1) More outputs are displayed/processed overall, which can overwhelm an operator; (2) More *incorrect* outputs are displayed/processed.

> *[EXAMPLE]*
> (*continued*) System A is now configured with a confidence threshold of 0.50. Operator B receives frequent alerts for low-confidence detections, many of which prove to be incorrect. The volume of alerts makes it difficult for her to identify and respond effectively to relevant information. Operator B therefore requests that the threshold be raised to 0.80, accepting a greater risk of missing lower-confidence detections, such as the tank, in exchange for fewer false alerts and a more manageable workload.

How to evaluate?

Confidence Threshold	Very high	High	Medium	Low	Very low

[Confidence Threshold] The threshold defined by the designer or operator of the system (see ¶4). Note that even "very low" in this case can be high in absolute terms (e.g. 0.5 may reasonably be considered as "very low").

Legal relevance

D DISTINCTION	An overly low confidence threshold increases the probability of incorrect outputs also being displayed/processed, potentially leading to the targeting of protected persons/objects across those categories, and possibly amounting to violations of the principle of distinction.

Relationship Chart

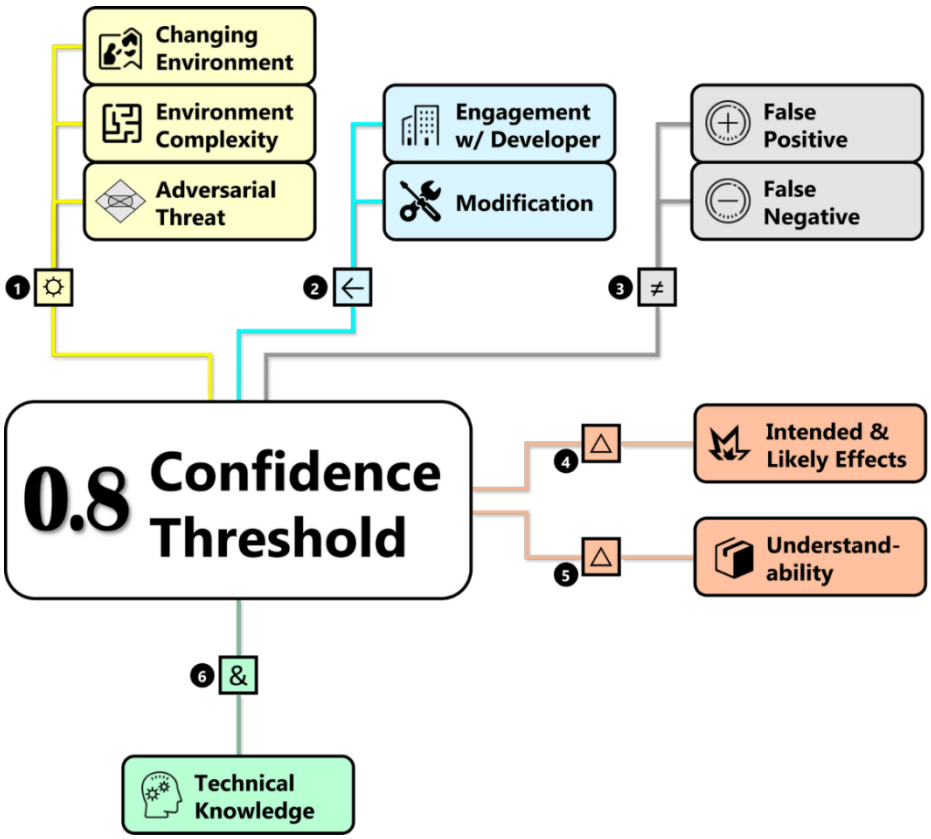


Chart Legend

BEFORE EVALUATING, CONSIDER

1 ☼

In environments that are rapidly changing, complex, or which feature active attempts by adversaries to disrupt AI systems (= *poor evaluations in [Changing Environment], [Environment Complexity] or [Adversarial Threat]*), operators may be inclined to alter an AI's confidence threshold more frequently to adapt to changing/complex circumstances. If this is the case, it may be necessary to **re-evaluate risk from [Confidence Threshold] more frequently**.

2 ←

[Confidence Threshold] may have been pre-determined by the developer of the AI system (*see [Engagement with Developer]*) or may be set or modified by operators (*see [Modification]*). In either case, it may be necessary to consult with these parties to **obtain the most up-to-date confidence threshold**.

3 ≠

Do not confuse [Confidence Threshold] with [False Positive Rate] and [False Negative Rate], the latter two being large-scale statistical averages of the AI system's performance. Confidence scores, in contrast, represent the probability of *single* outputs being correct.

RISKY PAIRINGS

4 △

A poor evaluation in [Confidence Threshold] is dangerous in combination with harmful direct or secondary effects (= *poor evaluation in [Intended & Likely Effects]*). A low confidence threshold **increases the probability of incorrect**

outputs (which can include erroneous strikes on civilians) also being acted upon.

5 △ A poor evaluation in **[Confidence Threshold]** is dangerous in combination with a poor evaluation in **[Understandability]**. A lower threshold increases the probability of incorrect outputs being displayed or acted upon, and **better understandability could assist operators** (= *strong evaluation in **[Understandability]***) in sorting between correct and incorrect outputs.

AFTER EVALUATING, CONSIDER

6 & Even if **[Confidence Threshold]** evaluates well, prioritising a strong evaluation in **[Technical Knowledge]** remains important, particularly so **operators do not naively trust outputs** with high confidence scores (as AI models can be 'confidently wrong'). See also ¶2.

[PRX]

Target Proxy



[DEFINITION]
Identifies how reliably the features or characteristics used by the AI system **distinguish the intended target category from other persons or objects**.

Concept

1 AI systems often identify persons, objects, or other categories by detecting particular **features or characteristics**, referred to here as **proxies**. These are observable features that the system uses to determine whether something belongs to the category it is designed to identify.

> *[EXAMPLE]*
> An AI system designed to identify a particular type of military vehicle might rely on features such as its shape, dimensions, markings, or sensor signature. These features act as proxies for the target category.

2 This variable evaluates **how specifically targets are defined** (*proxy uniqueness*) on the one hand, and **how many civilian entities** could correspond to this definition (*saturation*) on the other.

> *[FORMULA]*
> **Target Proxy Risk** = **Proxy Uniqueness** x **Proxy Saturation**

- **Proxy uniqueness** (¶3-¶4)**:** How specifically does the proxy correspond to the intended target category? A proxy that is distinctive to the target category presents less risk than one shared by many other persons or objects.
- **Proxy saturation** (¶5)**:** How commonly does the same proxy occur among civilians or civilian objects in the anticipated operating environment? The more frequently civilians or civilian objects share the proxy, the greater the risk of mistaken identification.

3 The following scale distinguishes different levels of **proxy uniqueness**:

Name	Description	Example Proxy
UNIQUE	Features that can only correspond to a defined, unique entity	Biometric data corresponding to rebel leader Mr. A
GROUP	Features corresponding to a defined and closed group of entities	Model-M main battle tanks within radius-R of coordinate-C
CATEGORY	Features corresponding to an open set of persons/objects, which share a common but unique signature	Model-M main battle tanks
GENERIC	Features corresponding to multiple open sets of persons/objects sharing a signature not unique to them	Vehicles with a weight exceeding X tonne

4 **Multiple features can be combined** to create a more **distinctive target proxy that is 'uniquely military'**.

A single feature may be shared by many military and civilian objects, and therefore provide a weak basis for identification. Combining several relevant features **may narrow the category and make the proxy more specifically military**.

For example, a proxy based only on '*weight greater than 1500 kg*' would by itself would be GENERIC (as it would match many military and civilian vehicles). Combining this with additional features, such as particular military markings, vehicle characteristics, or other indicators associated with the intended target category, may create a more distinctive proxy (making it fall under CATEGORY).

> *[COMPARE]*
> **[Target Proxy]** concerns how distinctive the features used by the AI system are for identifying the target category. For whether a particular person or object qualifies as a military objective under the circumstances, see **[Status Fluidity]**.

5 **Proxy saturation** checks **how commonly civilian entities in the anticipated operating environment share the features used** by the AI system for identification. Saturation is therefore context-dependent: the same proxy may present very different risks in different environments.

> *[EXAMPLE]*
> A system using "*weight greater than 1,500 kg*" as a proxy for identifying military vehicles may encounter many civilian vehicles that meet the same criterion in an urban environment. Saturation would therefore be high. In large depopulated fields, in contrast, most vehicles with a weight above 1,500 kg would likely be military: in such a case, saturation would be low, since few to no civilian entities in the area share this proxy.

How to evaluate?

Proxy Uniqueness	UNIQUE	GROUP		CATEGORY	GENERIC
AND					
Proxy Saturation	None	Rare	Common	Very common	Extremely common

[Proxy Uniqueness] See scale at ¶3 and ¶4.

[Proxy Saturation] A qualitative assessment of how common the chosen proxy is among civilian entities in the operating environment (see ¶5).

Legal relevance

D
DISTINCTION

The more generic and the more common the proxies used to categorise targets are among civilian objects/persons, the higher the probability that a civilian person/object also falls in that set. This may render the system an indiscriminate means or method of warfare, and consequently lead to attacks that violate the principle of distinction.

P PRECAUTIONS	The more generically the target profile is defined, the less certainty and predictability users have about the *specific* entities that will be affected by the attack.

Relationship Chart

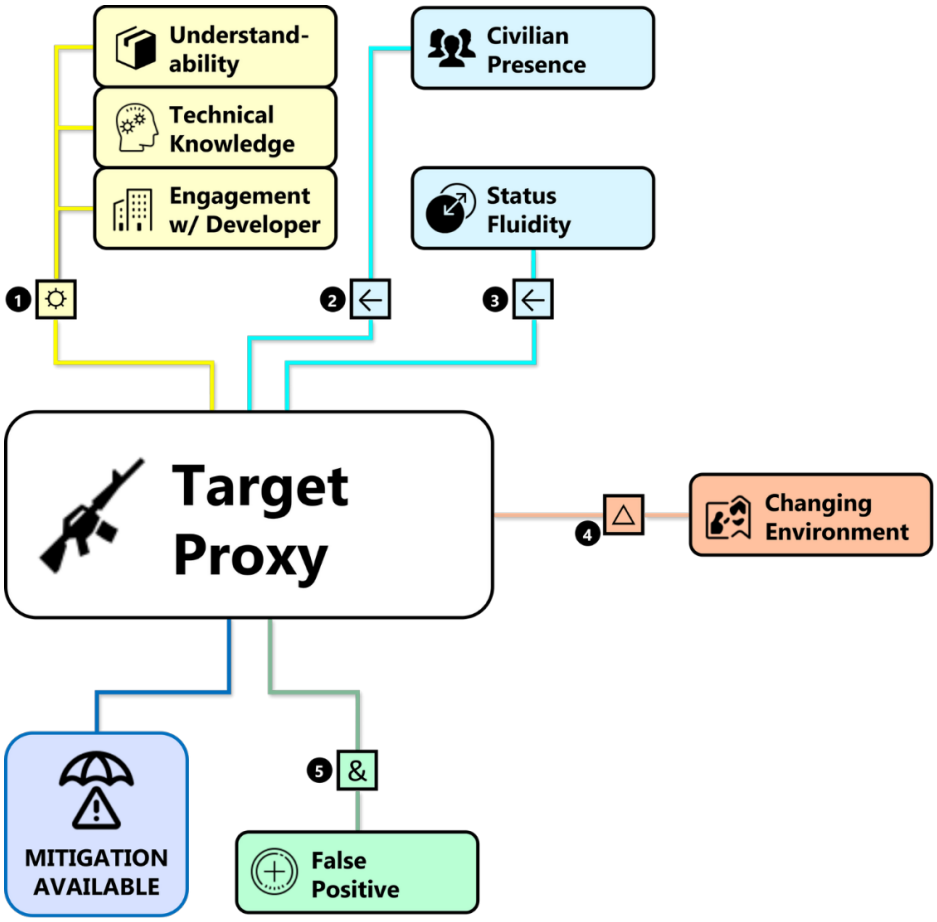


Chart Legend

BEFORE EVALUATING, CONSIDER |

1 Evaluating [Target Proxy] may require strong evaluations in [Understandability], [Technical Knowledge] and [Engagement with Developer], as this will **facilitate determining the exact features** that the AI model is using to make decisions.

2 A higher density and presence of civilians overall (= *poor evaluation in [Civilian Presence]*) increases the **likelihood of saturation also being high** (see ¶5).

3 Some features are more militarily-specific and durable than others. These militarily-specific and durable features usually correspond with a *strong evaluation in [Status Fluidity]*, are often associated with higher **proxy uniqueness**.

RISKY PAIRINGS

4 △

A poor evaluation in **[Target Proxy]** is dangerous in combination with a rapidly changing environment (=*poor evaluation in **[Changing Environment]***), since previously-specific features **may quickly become more generic** or without this being realised.

AFTER EVALUATING, CONSIDER

[MITIGATION]

Where **[Target Proxy]** indicates elevated risk, the user may consider **defining more specific features**, perhaps in combination with using features that are more militarily-unique (*see also **[Status Fluidity]***).

5 &

Even if very specific proxies are used (= *strong evaluation in **[Target Proxy]***), a system **may still be indiscriminate if it evaluates poorly in [False Positive Rate]**. For example, a system that searches based on specific biometric data (i.e., falling under UNIQUE), but which has a 0.5 False Positive Rate, has a one-in-two chance of mistaking another person for the target person.

[BIA]
Bias

[DEFINITION]
Identifies the risk that, due to algorithmic bias, the AI system **systematically performs differently across relevant groups or categories** in ways that may produce **undesirable and unintended outcomes**.

Concept

1 In this variable, bias refers to a phenomenon where the AI system **systematically performs differently across particular of groups of persons, objects, or other categories**. Such differences may arise from factors including the system's training data, model design, development choices, or the context in which it is used.

[COMPARE]
Bias as used in this variable is **distinct from cognitive bias or automation bias**, which concern human decision-making. For example, automation bias may cause an operator to place excessive reliance on an AI system's output. In contrast, **[Bias]** concerns systematic differences in the **AI system's own performance**.

2 Differences in performance become particularly concerning where they **systematically increase the risk of adverse outcomes for certain groups or categories**.

[EXAMPLE]
An AI system is used to identify civilian presence, to support collateral damage assessments. Because people with darker skin tones were inadequately represented in its training data, the system is less reliable at detecting them. As a result, some civilians are omitted from the information presented to commanders, potentially affecting subsequent proportionality assessments.

[EXAMPLE]
An AI system is used to distinguish enemy combatants from civilians. The system was developed and trained for use in a different operational environment and performs significantly worse when assessing persons from a particular ethnic group. As a result, it systematically misclassifies civilians from that group as members of the adversary's armed forces.

3 **Bias can arise at different stages** of an AI system's development and use, and **has different causes**. These may include the following.

- **Training data:** Biases or gaps in the data used to train the system may result in systematically poorer performance for particular groups or categories (*see **[Training Data]***).
- **Proxy selection:** The way relevant persons, objects, or characteristics are defined and the proxies used to identify them may introduce or reinforce bias (*see **[Target Proxy]***).

- **Human decisions during development:** Assumptions or biases held by developers, data annotators, or other persons involved in development may be reflected in the system's design or outputs (*see **[Engagement with Developer]***).
- **Use outside the intended context:** A system may exhibit biased performance when used in populations or environments that differ from those for which it was developed or tested (*see **[Out-of-Scope Use]***).

> *[EXAMPLE]*
> An AI system alerts soldiers to threats in their vicinity. A gentleman approaches in the street, and does not seem to slow down despite repeated warnings to halt. Based on this behaviour, the AI-system flags him as a potential danger. In fact, the gentleman was deaf, a trait that neither the AI-system nor the soldiers had ever encountered before.

Biases present in historical data may be reproduced by AI systems. Where training data reflects systematic shortcomings in previous human practices or assessments, an AI system trained on that data may learn and reproduce the same patterns.

> *[EXAMPLE]*
> An AI system used to support collateral damage estimation is trained partly on assessments from previous operations. Those assessments relied heavily on static population-density data and did not consistently account for conflict-related changes in civilian behaviour, such as displacement or movement towards shelters. As a result, the AI system reproduces these patterns, systematically underrepresenting civilian presence in circumstances where such behavioural changes have occurred.

4 Where bias is identified or suspected, **the concrete risk it poses should be assessed in context**, ideally in cooperation with technical experts.

> *[FORMULA]*
> **Risk from Bias** = **Scale** x **Impact** x **Likelihood**

Relevant considerations to determine risk from bias include:

- **Scale:** How many persons, objects, or cases are likely to be adversely affected?
- **Impact:** How significant is the difference in the system's performance or outcomes for the affected group, compared with other relevant groups?
- **Likelihood:** How likely is the bias to manifest in the anticipated context of use?

> *[EXAMPLE]*
> An AI system used to transcribe incoming voice reports performs less reliably for speakers with Accent A. However, Accent A is rarely encountered in the anticipated operational context, meaning that relatively few reports are likely to be affected. Furthermore, transcripts are reviewed by human analysts who can usually identify and correct transcription errors before the information is used. The scale, likelihood, and practical impact of the identified bias may therefore be relatively limited in this context.

Bias may be difficult to detect. An evaluator may not know in advance which groups or characteristics are associated with poorer system performance, particularly where relevant groups were not considered during development or testing. Assessing **[Bias]** may therefore require examining performance across relevant groups and consulting developers, testers, or other technical experts (*see **[Engagement with Developer]***).

How to evaluate?

Affected Persons/Objects	None	Limited	Some	Many	Extensive
AND					
Negative Impact	No impact	Slight impact	Moderate impact	Strong impact	Severe impact
AND					
Likelihood	Impossible or highly unlikely	Unlikely	Likely	Very likely	Certain

[Affected Persons/Objects] The extent of persons/objects negatively affected by bias (see ¶4).
[Negative Impact] How much worse the AI performs on the affected persons/objects, compared to the average (see ¶4).
[Likelihood] The likelihood of the bias manifesting (see ¶4).

Legal relevance

D DISTINCTION	Systemically poorer performance for particular groups/categories may increase the risk that protected persons or objects within those groups are misidentified as lawful targets. Where AI outputs are relied upon for target identification or classification, such disparities may therefore undermine the ability to distinguish reliably between military objectives and protected persons or objects.
C CIVILIAN HARM	Systemically poorer performance for particular groups/categories may also result in civilians or civilian objects being consistently under-detected or omitted from assessments of expected civilian harm. Where such outputs inform proportionality assessments, this may lead to civilian harm being underestimated, with particular groups—such as children or persons with disabilities—disproportionately affected.

Relationship Chart

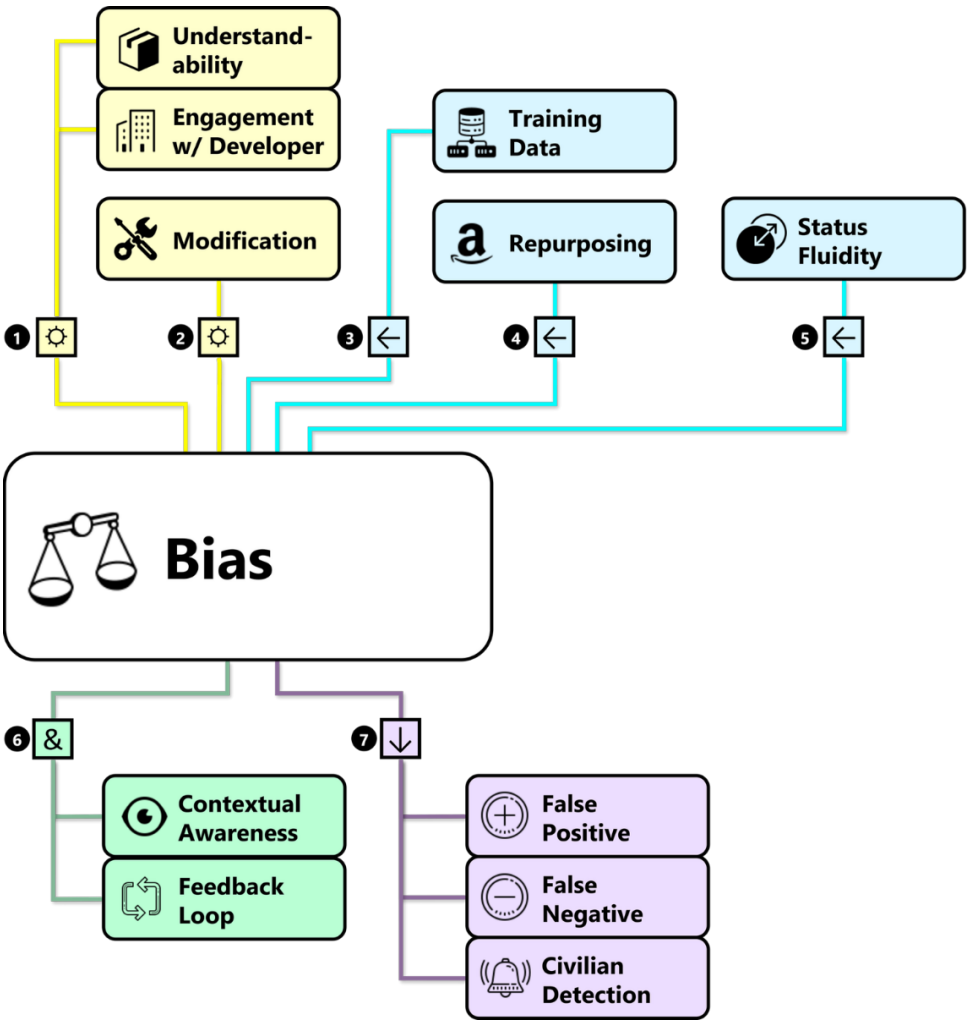


Chart Legend

BEFORE EVALUATING, CONSIDER |

1 ☼ **Evaluating risk from [Bias] may require** the ability to understand the underlying logic or behaviour of the AI model (= *strong evaluation in [Understandability]*) and the ability to access the AI's training data and development history (= *strong evaluation in [Engagement with Developer]*).

2 ☼ Updates or modifications **may introduce new biases** that were not present previously. If modification has recently occurred (*see [Modification]*), re-evaluation of the risk from bias may be necessary.

3 ← There is more risk of bias (=*negative effect on [Bias]*) if **low-quantity and poorly representative data was used to train** the AI model (= *poor evaluation in [Training Data]*).

4 ← Open-source and general-purpose AI models (=*poor evaluation in [Repurposing]*) are more likely to have **carried over hidden biases** from their training data (=*negative effect on [Bias]*).

5 ← More fluid conditions that permit a person/object to be targeted (= *poor evaluation in **[Status Fluidity]***) are often **subjective and open to human biases**, increasing the risk that such bias is transferred to the AI system (=*negative effect on **[Bias]***).

AFTER EVALUATING, CONSIDER

6 & Even if **[Bias]** evaluates well, effective human oversight mechanisms (= *strong evaluations in **[Contextual Awareness]** and **[Feedback Loop]***) allow operators to **recognise biases that were not discovered** beforehand.

7 ↓ If bias is present (= *poor evaluation in **[Bias]***), this may **drop the AI model's performance** below its average rate for the affected persons/objects (= *negative effect on **[False Positive Rate]** and **[False Negative Rate]***) and may **render the ability to detect civilians ineffective** vis-à-vis the affected persons/objects (= *negative effect on **[Civilian Detection]***).

[DCT]

Civilian Detection Ability

[DEFINITION]

Identifies whether, and to what extent, the AI system is **capable of detecting civilians, civilian objects, and other protected persons or objects.**

Concept

1 **Not all AI systems are designed or equipped to detect civilians, civilian objects, or other protected persons or objects**. Systems may be capable of identifying the intended target category, without being able to recognise *other* persons or objects in the surrounding environment.

The absence of civilian detection does not necessarily indicate that civilians are absent; they may simply fall outside the categories the system is designed to detect (*see also* ***[False Positive Rate]*** *and* ***[False Negative Rate]***).

[EXAMPLE]

Military M is considering two otherwise comparable AI-enabled systems designed to detect and engage enemy radar installations. System B additionally includes sensors and a functionality designed to detect nearby civilians. Military M selects System B because its additional civilian-detection capability provides greater scope for precautionary measures in environments where civilians are likely to be present.

2 The ability to detect civilians and other protected persons or objects can enable **different safeguards, depending on the function** of the AI system:

- ***AI-enabled weapon systems***: Civilian detection may enable the system to trigger predefined safeguards, such as **delaying or aborting an attack** when civilians are detected within a specified area. For example, an AI loitering munition may be configured not to attack a detected military objective as long as civilians are detected within a predefined distance.
- ***DSS***: Civilian detection may provide operators with information relevant to assessing and mitigating expected civilian harm. For example, a targeting-support system may flag detected civilians or civilian objects in the vicinity of a proposed target, providing information relevant to collateral damage estimation and proportionality assessments (*see also* ***[False Negative Rate]***).

[WARNING]

A system's capacity to detect civilians does not replace the need to undertake precautions as required by IHL. It is not permitted to *solely* rely on the AI-system's ability to detect civilians. Appropriate measures must still be taken to verify relevant information and assess the risk to civilians in the circumstances. For example, the fact that an AI-enabled munition is programmed to delay an attack when it detects nearby civilians does not, by itself, remove the operator/commander's obligation to verify whether civilians are present. The reliability and limitations of the detection capability, as well as other available information about civilian presence, must still be considered.

3 The **capacity to reliably detect civilians depends on** both the range of civilian and protected categories the system can detect and the specificity which it can identify them.

> *[FORMULA]*
> **Capacity to Detect Civilians** = **Range of Categories** x **Specificity of Identification**

- **Range of categories:** How broad is the range of civilian persons, objects, and other protected entities the system is capable of detecting?
- **Specificity of identification:** How precisely can the system identify what it has detected?

A system with this functionality may be able to detect a **limited or a broad range** of civilian entities.

> *[EXAMPLE]*
> An anti-tank system trained to recognise tanks, busses, tractors, and cars may be capable of detecting some civilian vehicles (*limited*). In contrast, a satellite-image analysis system may be capable of detecting a broader range of relevant categories, including civilian buildings, crowds, cultural property, or infrastructure (*broad*).

Greater specificity may provide more useful information about the civilian or protected character of what has been detected.

> *[EXAMPLE]*
> (*continued*) A satellite-image analysis system that only identifies "buildings" (which can also be military compounds) provides limited information about its civilian character or function. A system capable of distinguishing between different types of buildings—for example, by identifying features indicating that a building may be a school—may provide more useful information for assessing potential civilian harm.

4 If the AI system is used to support human decision-making (e.g. as a DSS), consider the **overall capacity of the human-machine team as a whole** to detect civilians and other protected persons or objects.

A system may technically detect civilians reliably, but that capability will have limited practical value if the relevant information is not effectively communicated to the operator. Interface design and operator situational awareness may therefore affect the overall capacity to detect civilians (*see **[Contextual Awareness]***).

> *[EXAMPLE]*
> A DSS is used to detect civilians within an area under surveillance. However, its interface displays civilians as small grey markers that are difficult to notice, while potential threats are highlighted using prominent red markers. In this case, although the AI system itself may perform well at civilian detection, the interface makes that information difficult for the operator to perceive. The effective capacity of the human-machine team to detect and account for civilians may therefore be lower than the AI system's technical performance alone would suggest.

How to evaluate?

Range of Detectable Categories	Significant range of categories	Large range of categories	Some categories	Limited categories	No ability

AND

Specificity of Identification	Extremely specific	Very specific	Moderately specific	Relatively generic	Very generic

AND

Human-Machine Team's Ability	Very high ability	High ability	Moderate ability	Limited ability	No ability

[Range of Detectable Categories]	A qualitative assessment of how many relevant categories of civilian entities the AI can detect (see ¶3).
[Specificity of Identification]	A qualitative assessment of how specifically the AI can identify civilian entities (see ¶3).
[Human Machine Team's Ability]	(*Applicable if the AI system is human-operated*) The overall ability of the human-machine team to perform civilian detection, also taking into consideration human factors such as interface design (see ¶4).

Legal relevance

C CIVILIAN HARM	Reliable civilian detection supports the assessment of expected incidental civilian harm, including by providing information relevant to collateral damage estimation and proportionality assessments. Limitations in civilian detection may result in civilian presence or civilian objects being underrepresented or missed, increasing the risk of attacks being carried out that are disproportionate.
P PRECAUTIONS	The capacity to detect civilians can support the taking of feasible precautions to avoid or minimise incidental civilian harm, including by informing decisions to verify targets, modify, delay, suspend, or cancel attacks. Where civilian presence cannot be reliably detected or communicated to the operator, the range and effectiveness of available precautionary measures may be reduced.

Relationship Chart

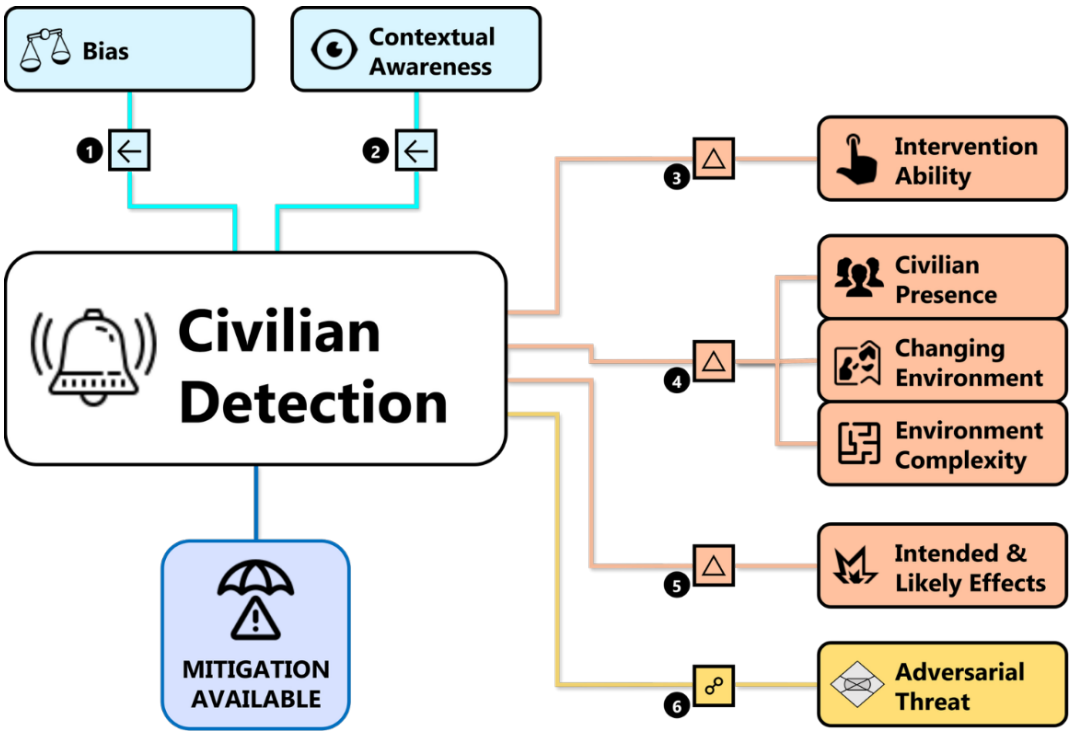


Chart Legend

BEFORE EVALUATING, CONSIDER |

1 ← If there is risk of bias (= *poor evaluation in [Bias]*), this may render the AI system's civilian detection capacity **less effective *vis-à-vis* the persons/objects affected by the bias** (= *negative effect on [Civilian Detection]*). For example, civilian persons from a certain ethnic race may be detected less reliably).

2 ← For human-machine teams (see ¶4), factors such as interface design (*see [Contextual Awareness]*) may (positively or negatively) **affect the overall ability of the operator** to detect civilians.

RISKY PAIRINGS |

3 △ If manually **intervening after launch is difficult or not possible** (= *poor evaluation in [Intervention Ability]*), a strong evaluation in [Civilian Detection] is particularly beneficial. This allows for autonomous cancelling or delaying, or making informed decisions prior to launch that take into consideration all potential civilian harm relevant to the proportionality assessment (see ¶2).

4 △ If the environment is civilian-dense (= *poor evaluation in [Civilian Presence]*) or the environment is rapidly changing or complex (= *poor evaluations in [Changing Environment] and [Environment Complexity]*), prioritising a strong evaluation in [Civilian Detection] can **mitigate risk from unforeseen developments in civilian positioning**.

5 △ If the direct and secondary effects of using the system are harmful or widespread (= *poor evaluation in [Intended & Likely Effects]*), prioritising a strong evaluation in [Civilian Detection] is recommended. As **attacks become more harmful or deadly**, the value of last-minute cancelling and obtaining a comprehensive picture of collateral concerns increases.

6 ☍ A strong evaluation in [Civilian Detection] may **work counterproductively** against adversaries prone to exploiting AI systems through countermeasures (= *poor evaluation in [Adversarial Threat]*). Crafty adversaries may attempt to game the AI system by deliberately placing civilian persons/objects (or decoys that the AI mistakes as civilian persons/objects) near their military targets, to trigger any failsafe in the AI system. If such tactics by the enemy are common, there may be no choice other than to forego civilian detection by the AI and mitigate risk to C through other variables.

AFTER EVALUATING, CONSIDER |

[MITIGATION]

Where [Civilian Detection] indicates elevated risk, consider whether the system's civilian-detection capability can be introduced or improved. Possible measures include expanding the range of civilian and protected persons or objects the system can detect, improving the specificity with which they can be identified, or improving how detections are communicated to operators. Expanding detection capabilities may require additional representative training and testing data for relevant categories, such as children or persons with disabilities (*see [Training Data] and [Bias]*).

[STA]

Status Fluidity of Target



[DEFINITION]

Identifies how rapidly and how predictably a person or object **may gain or lose its status as a lawful target** as circumstances change.

Concept

1 Under IHL, persons/objects may only be made the object of an attack **if the applicable conditions for lawful targeting are satisfied**. Whether those conditions are met may depend on characteristics of the person or object itself, or on its role/use in the particular operational circumstances.

2 The legal status of some potential targets is relatively static, while that of others is highly dependent on changing circumstances. The **more rapidly or unpredictably that status can change, the greater the challenge** of ensuring that information used to identify a lawful target remains accurate and up to date.

3 For objects, the degree of status fluidity may differ depending on the basis on which an object qualifies as a military objective:

- ***Military objects by nature***: Quite static, because the object's characteristics link it inherently to military functions or actions. For example, most military equipment and military facilities (e.g. aircraft carriers and military headquarters) qualify as military objectives by nature. Because the characteristics giving rise to that status are generally stable, their status is less liable to change rapidly as the operational context evolves.
- ***Military objectives by location***: Relatively static, since their targetability depends on the military significance of the object's particular location in the prevailing circumstances. Changes in the operational situation may affect whether the location continues to make an effective contribution to military action. For example, a fortification may have military significance because it controls access to an important operational area. If the front line or operational priorities shift, that location may lose—or subsequently regain—that significance.
- ***Military objects by current use or future use*** *(in legal texts: 'by purpose')*: Very fluid, as their status as military objectives depends on the way it is being used, or is intended to be used. As operational conditions change, so too may this role. For example, a civilian power facility may qualify as a military objective where its use makes an effective contribution to military action and the other requirements for classification as a military objective are satisfied. If the relevant military use ceases, the basis for that assessment may change.

[WARNING]

Even for more static statuses, prevailing operational circumstances must still be considered. Status should not simply be assumed to remain unchanged.

4 **For persons, the stability of targetable status may depend on the legal basis on which the person may be targeted**:

- ***Status-based conditions***: Generally more static, since the person's targetable status derives from their membership in, or function within, a relevant organised armed force or group. Such status will not ordinarily change simply because the person's immediate behaviour changes. For

example, a member of a State's armed forces does not cease to be targetable merely because they are temporarily off duty.

- ***Behaviour based-conditions:*** Highly fluid, because targetability of these persons depends on what the person is doing at a particular time. Changes in conduct may result in relatively rapid changes in whether the person may lawfully be targeted. For example, a civilian who directly participates in hostilities loses protection against direct attack "for such time" as they directly participate. Whether this condition is satisfied therefore changes as the person's conduct changes.

> ***[WARNING]***
> Even comparatively stable statuses of persons should not be treated as permanently fixed. A person who would otherwise be targetable (e.g. a member of the armed forces) may become protected against attack, for example by surrendering or otherwise becoming *hors de combat*. The use of AI to identify or select persons for targeting may therefore require particular attention to whether the system can recognise and respond to changes in legally relevant circumstances. Relatedly, the use of AI in decisions concerning the targeting of persons has also raised broader ethical and policy concerns, including concerns relating to human dignity. In some circles, these concerns have led to arguments in favour of banning anti-personnel AWS.

5 **More fluid target statuses may be more difficult for an AI system to assess reliably**, as AI systems generally rely on proxies to identify and classify persons or objects (*see **[Target Proxy]***).

More **static conditions** are generally more objective and amenable to being machine learnt (e.g. 'uniform' can be a solid proxy for 'member of armed forces').

Fluid conditions generally require more subjective and contextual evaluations, and are more difficult for AI to robustly analyse. For example:

- 'Military objective by future use' (*in legal texts: 'by purpose'*) requires an analysis to be made of enemy intent and planning. An AI system may not be capable of deriving that assessment reliably from observable characteristics alone.
- 'Direct participation in hostilities' requires a relatively complicated three-pronged legal test to be fulfilled before the civilian becomes targetable. For example, the 'direct causation' prong requires a causal link to be established between the hostile act and harm done to the adversary, which may depend on contextual information that is difficult for an AI system to observe or assess reliably.

How to evaluate?

For Objects	By nature		By location		By use or future use
OR					
For Persons			Status based		Behaviour based

[For Objects]	See ¶3.
[For Persons]	See ¶4.

Legal relevance

D DISTINCTION	Fluid or context-dependent target status may be difficult for an AI system to assess reliably. Where changes in status are not accurately recognised, protected persons or objects may be incorrectly treated as lawful targets, increasing risks to compliance with the principle of distinction.

Relationship Chart

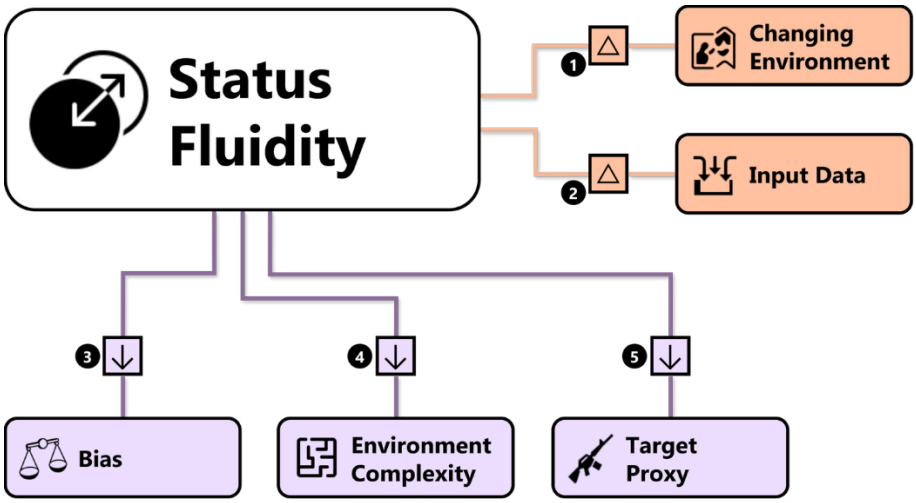


Chart Legend

RISKY PAIRINGS

1 △ It is very risky to use AI to target more fluid objectives (= *poor evaluation in [Status Fluidity]*) in a rapidly changing environment (= *poor evaluation in [Changing Environment]*). This can amplify the rate and unpredictability of status changes, **significantly increasing the risk of misidentifications and unlawful attacks** on protected persons/objects.

2 △ If the objectives are fluid (= *poor evaluation in [Status Fluidity]*), ensuring the recency of data and intelligence fed to AI system (= *strong evaluation in* [Input Data]) becomes doubly important. This ensures that decisions are made based on **the most up-to-date indicators of the target's status**.

AFTER EVALUATING, CONSIDER

3 ↓ More fluid and behaviour-based conditions (= *poor evaluation in [Status Fluidity]*) are often **subjective and open to human biases**. For example, carrying weapons is usually considered threatening, but may in actuality be common in the local community. This subjectivity makes it more likely that such human biases are also transferred to the AI system (= *negative effect on [Bias]*).

4 ↓ Objects with less durable properties (= *poor evaluation in [Status Fluidity]*) are **often dual-use objects, and/or located in proximity to other civilian objects**.

This can complicate the operating environment (= *negative effect on* ***[Environment Complexity]***).

5 ↓ More durable statuses (= *strong evaluation in* ***[Status Fluidity]***) are **often associated with clear or objective properties** (e.g. a uniform), which generally correspond with a strong evaluation in **[Target Proxy]**.

CLUSTER 4: HUMAN AWARENESS

[AWA]

Contextual Awareness



[DEFINITION]
Identifies the ability of a human operator to **obtain, understand, and assess information about the operational context** in which the AI system is functioning.

[COMPARE]
This variable evaluates the operator's **real-time** awareness of the AI system and its operational context. For more general reporting of the system's past behaviour or performance, see **[Feedback Loop]**.

Concept

1 Effective contextual awareness allows the operator to understand when the AI system is functioning as expected and when intervention may be necessary to prevent (imminent or future) undesirable behaviour or harm.

Contextual awareness is particularly important for identifying **developments that may be difficult to anticipate in advance**, including emergent or unpredictable behaviour (*see **[Unpredictable Behaviour]***) and adversarial activities (*see **[Adversarial Threat]***).

2 Information about the AI system and its operating environment **may be obtained** from a range of sources, including:

- **The AI system itself:** For example, a DSS may provide information or reports about the analyses it performs.
- **Other intelligence or surveillance assets:** For example, an ISR drone provides camera footage showing how a land-based autonomous system is performing.
- **Reports and data from previous deployments:** Information about the system's past behaviour or performance may provide additional context (*see **[Feedback Loop]***).

3 The **reliability and completeness of information** should be assessed **critically and in context.**

- **Relevant factors may include** the quality and number of sensors providing information, the age of the information, how frequently it is updated, and whether relevant information may be missing (*see also **[Input Data]***).
- **Unreliable, incomplete, or outdated information** may give the operator an inaccurate picture of the system or its environment, leading to misplaced trust.

[EXAMPLE]
A drone swarm reports the location of each drone to an operator every 5 seconds. When displayed over time, the updates appear to show a smooth and coordinated flightpath, leading the operator to conclude that the swarm is functioning as expected. In reality, individual drones are making erratic movements that create a risk of collision, but these movements would only become apparent if location information were updated more frequently.

If important information is **only provided when demanded** by a human or AI, there is an additional risk that information may be omitted or incorrectly filtered, as these humans/systems can also make mistakes.

> *[EXAMPLE]*
> An AI-enabled drone searching for tanks is programmed to notify its operator if it encounters, with a confidence score of 0.7 or lower, an object it *believes* is a tank (*see also* ***[Confidence Threshold]***). However, a systematic error causes the system to classify tractors as 'tanks' with high confidence (0.95). Because only low-confidence detections are reported, these incorrect classifications of tractors are never brought to the operator's attention. The operator therefore remains unaware of a recurring problem that might otherwise have prompted intervention or recall of the system.

4 **The way information is presented** can also affect contextual awareness.

- **Interface design**—including the amount of information presented, its format and prominence, and the use of visual or audio cues—can influence what operators notice and how they interpret it.
- **Human cognitive biases** should be considered when determining how information is presented.

> *[EXAMPLE]*
> An AI-enabled heads-up display (HUD) is used to improve a tank operator's situational awareness. Potential threats are highlighted using flashing red borders accompanied by an audio alert. In practice, it is revealed that operators tend to interpret these signals as indicating a confirmed or immediate threat and respond accordingly, even though the system is only identifying a *potential* threat.

5 The measures necessary to achieve adequate contextual awareness will **depend on the system, operator, and use-context**.

There is no single configuration or information-display mechanism that will be appropriate in all circumstances. Testing with intended operators under realistic conditions, together with their feedback and experience (*see also* ***[Feedback Loop]***), can help determine whether the information available and the way it is presented provide sufficient contextual awareness for the anticipated use-context.

How to evaluate?

Availability of Information	Many sources	Several sources	Some sources	One source	No sources
AND					
Reliability and Completeness	Complete and reliable	Mostly complete and reliable	Mixed completeness and reliability	Likely incomplete, reliability questionable	Incomplete, likely unreliable
AND					
Presentation of Information	Optimised for contextual awareness	Mostly supports contextual awareness	Some risk of impairing contextual awareness	Likely to impair contextual awareness	Strongly impairs contextual awareness

[Availability of Information]	The extent of sources the operator can consult to obtain awareness of the system and its operating environment (see ¶2 and ¶3).
[Reliability and Completeness]	The extent the sources the operator relies on are complete and reliable (see ¶3).
[Presentation of Information]	Whether the manner in which information is presented to the operator facilitates or hinders operational awareness (see ¶4). Ideally requires testing with actual operators (see ¶5).

Legal relevance

D C DISTINCTION/ CIVILIAN HARM	Poor contextual awareness may prevent operators from recognising changes relevant to both target identification and expected civilian harm. Incomplete, outdated, or poorly presented information may make it more difficult to identify when protected persons or objects are incorrectly treated as lawful targets, or when changes in civilian presence or other circumstances affect collateral damage estimates and proportionality assessments. This increases risks to compliance with the principles of distinction and proportionality.
P PRECAUTIONS	Contextual awareness supports the operator's ability to monitor how the system is functioning in the circumstances of use and to identify when intervention or other precautionary measures may be required. Poor contextual awareness may make unexpected behaviour, changes in the operating environment, or emerging risks more difficult to detect, reducing the operator's ability to take timely measures to avoid or minimise civilian harm.

Relationship Chart

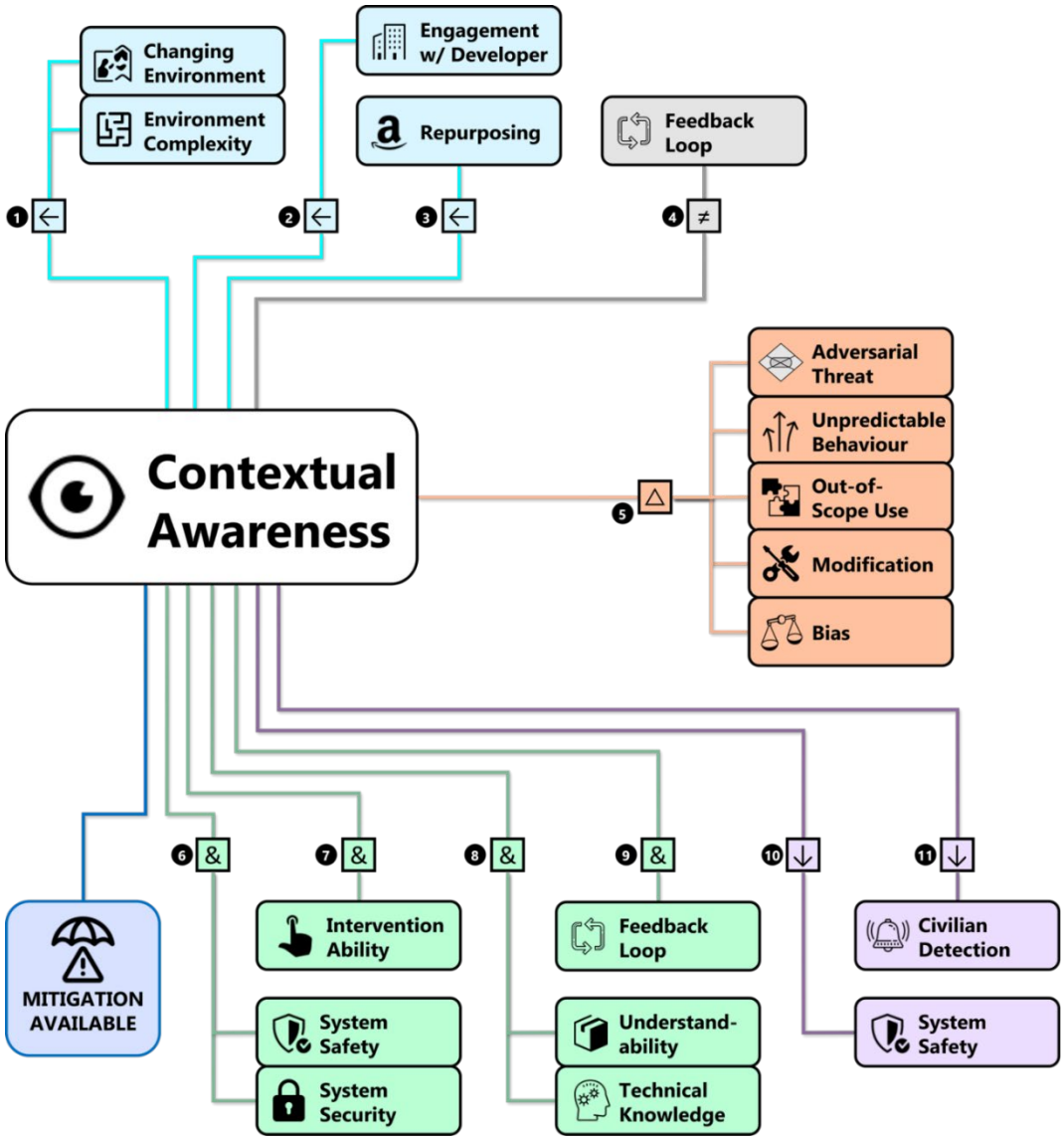


Chart Legend

WHEN EVALUATING, CONSIDER |

1 ← In complex and rapidly changing environments (= *poor evaluations in [Changing Environment] and [Environment Complexity]*), **maintaining the same level of contextual awareness is more challenging** and might demand more resources (such as additional sensors).

2 ← **Tailoring the way information is communicated and presented to operators** to enable optimal situational awareness (see ¶4) is easier if the user was more involved during the development process (= *strong evaluation in [Engagement with Developer]*).

3 ← If the AI system was not originally designed for military purposes (= *poor evaluation in [Repurposing]*), this may have a negative effect on [Contextual Awareness]. The way information is presented in **civil or general-purpose AI applications may not be well-suited for military contexts**.

4 ≠ **Do not confuse** [Contextual Awareness]—which evaluates the operator's *real-time* awareness of the operational context—with [Feedback Loop], which refers to more general reporting of an AI system's (past) behaviour/performance, usually done *ex post*.

RISKY PAIRINGS |

5 △ Contextual awareness strengthens the operator's ability to detect unintended behaviours, failures or deviations. As such, **prioritising a strong evaluation in [Contextual Awareness] is highly advised if**:

- There is (high) risk of adversarial activity or unpredictable behaviour of the AI system (= *poor evaluations in [Adversarial Threat] or [Unpredictable Behaviour]*).
- Using the system outside of its intended use-context is likely or cannot be avoided (= *poor evaluation in [Out-of-Scope Use]*).
- The system is frequently modified or was recently modified substantially (= *poor evaluation in [Modification]*).
- There is (high) risk of unidentified biases (= *poor evaluation in [Bias]*).

AFTER EVALUATING, CONSIDER |

[MITIGATION]

Where [Contextual Awareness] indicates elevated risk, consider measures to improve the availability, reliability, completeness, or presentation of information provided to operators. This may include adapting the system or interface, and conducting realistic testing with intended operators to ensure that relevant information is timely, understandable, and actionable (see also ¶5).

6 & Even if [Contextual Awareness] evaluates well, system safety and security should be maintained (= *strong evaluations in [System Safety] and [System Security]*) to **ensure strong performance in general**.

7 & Even if [Contextual Awareness] evaluates well, maintaining a reliable way for operators to intervene (= *strong evaluation in [Intervention Ability]*) may be necessary to **enable the operator to *act*** in response to undesirable behaviours or malfunctions.

8 & Even if [Contextual Awareness] evaluates well, ensuring solid technical understanding of AI-based systems in general, and the AI model's logic (= *strong evaluations in [Technical Knowledge] and [Understandability]*) may be necessary to **empower operators to draw conclusions** about the (un)reliability of their AI systems, given contextual information at their disposal.

9 & Even if [Contextual Awareness] evaluates well, installing a strong after-action and feedback process (= *strong evaluation in [Feedback Loop]*) remains important as it provides **information about *prior* performance and malfunctions**.

10 ↓ A strong evaluation in [Contextual Awareness] **improves the human safeguards** against harmful failures (= *positive effect on [System Safety]*).

11 ↓ Design choices that negatively impact operator awareness (see ¶4) may **reduce the overall ability of the human-machine team to detect collateral concerns** (= *negative effect on [Civilian Detection]*).

[FBK]

Feedback Loop

[DEFINITION]

Identifies the **capability to collect and use information about the AI system's actual behaviour, performance, and effects** during or after operations.

[COMPARE]

[Feedback Loop] concerns the ability to **identify, report, and learn from the AI system's behaviour, performance, and effects during or after its use**. This differs from [Contextual Awareness], which concerns an operator's real-time awareness of the system and its operational context.

Concept

1 An evaluator's or user's understanding of an AI system **cannot be complete before deployment**. Information obtained from operational use can be crucial for identifying limitations, unexpected behaviour, or risks that were not apparent during development and testing.

AI systems may behave unpredictably (*see [Unpredictable Behaviour]*), respond differently to complex environments or adversarial activity (*see [Environment Complexity], [Adversarial Threat]*), or experience changes in performance as operating conditions evolve (*see [Changing Environment]*). Their behaviour may also change following modifications or updates (*see [Modification]*), while some other issues, such as [Bias], may only become apparent after repeated use.

Effective feedback loops allow relevant information from operational experience to be **collected, reported, analysed, and acted upon**. Timely and reliable feedback can improve a user's understanding of the system's behaviour, limitations, and vulnerabilities, and inform subsequent decisions about its use.

Such future decisions may include adapting operational procedures, changing system configurations, modifying or retraining the system (*see [Modification]*), introducing additional safeguards, or suspending its use where necessary.

2 Feedback information may be **obtained from multiple sources**. Combining different sources can provide a more complete picture of how the AI system actually behaved, performed, and affected the operational environment. These can include the following.

- ***Operator***: Information collected during operations, including sensor feeds, surveillance assets, signals intelligence, and observations by operators or other personnel. Some of the same sources used to support contextual awareness [Contextual Awareness] may therefore also provide information for subsequent review.
- ***Automated reporting***: Some AI systems may automatically record or report information about their operation, outputs, performance, or the results of particular actions.
- ***Other sources***: Open-source information, media reporting, civilian reports, or other third-party sources may provide information about system behaviour or effects that was not apparent to operators, including potential civilian harm or system failures.

[WARNING]

In practice, not all assets available for supporting [Contextual Awareness] might **necessarily also be available** to obtain feedback information for the purposes of [Feedback Loop]. Operational assets may be reassigned, data may not be retained, or relevant information may otherwise become unavailable after an operation. [Contextual Awareness] and [Feedback Loop] should therefore need to be assessed separately.

3 Effective feedback loops may require **institutional and technical mechanisms** for recording, reporting, analysing, and acting on information. Relevant mechanisms may include:

- ***After-action review***: Consistent and timely processes for reviewing system behaviour, performance, and effects following operational use, including relevant after-action battle-damage assessments.
- ***Decision traceability***: Records of relevant decisions, information relied upon, and the chain of command can assist in reconstructing how and why particular decisions were made.

[NOTE]

Robust after-action review, documentation, and reporting practices may also improve organisational learning and accountability beyond the use of AI systems.

- ***System traceability and logging*:** Retaining appropriate records of system operations, inputs and outputs, and relevant updates or modifications can support subsequent analysis of system behaviour.
- ***Development documentation*:** Access to relevant information about training data, model design, testing, and other development choices may help explain behaviour observed during operations and identify appropriate corrective measures. (*See also [Training Data] and [Engagement with Developer].*)

4 A feedback loop is more effective when information can be **relayed frequently, reliably, and promptly.** The information must also be sufficiently reliable to support subsequent decisions. Rapid feedback that is inaccurate or misleading may itself lead to inappropriate responses.

Feedback mechanisms differ in how frequently, quickly and reliably they provide information. Some may be irregular or significantly delayed. For example, information about civilian casualties obtained through post-operation investigations may take days or longer to become available. Other mechanisms, such as automated system reporting or reporting from nearby systems (see also ¶2), may provide much faster indications of unexpected behaviour, performance degradation, or harmful effects. The quality of sensors and means of processing captured data may determine how reliable and accurate the information is.

[EXAMPLE]

Operators use an AI decision-support system to predict adversary movements. Separate state-of-the-art monitoring software continuously compares the system's predictions against incoming intelligence and identifies a significant increase in erroneous outputs since the previous day. The issue is reported to the commander, who suspends the use of the DSS pending investigation and corrective action.

How to evaluate?

Reporting Structures	Comprehensive	Relatively comprehensive	Partial	Minimal	None
AND					
Quality of Feedback	Reliable and accurate	Mostly reliable and accurate	Partly reliable and accurate	Mostly unreliable or inaccurate	Unreliable and inaccurate
AND					
Frequency of Feedback	Constant	Frequent	Often	Rare	Never
AND					
Timeliness of Feedback	Instant	Short delay	Moderate delay	Long delay	Significantly delayed

[Reporting Structures]	How institutionally and technically robust, comprehensive and well-established the reporting structures are (see ¶3).
[Reliability of Feedback]	How reliable and accurate the feedback information is (see ¶4).
[Frequency of Feedback]	How frequently users are able to obtain information about the actual effects of the AI system. May also depend on the technical structure of the reporting process. See ¶4.
[Timeliness of Feedback]	How significant the delay is between the impact and users being able to notice this impact. May also depend on the technical structure of the reporting process. See ¶4.

Legal relevance

D C DISTINCTION/ CIVILIAN HARM	Robust feedback loops enable users to identify and analyse errors or limitations in the system's past performance, including failures in target identification or civilian harm assessments. Incorporating these lessons into subsequent use can improve the information and assumptions relied upon for distinction and proportionality assessments, reducing the risk that errors or performance issues are repeated.
P PRECAUTIONS	Effective feedback loops enable users to continually reassess assumptions about the system's behaviour, performance, and effects in light of operational experience. Timely feedback can reveal emerging risks or limitations and inform additional safeguards, changes to the system or its use, or the suspension of use where necessary, supporting the taking of feasible precautions to avoid or minimise civilian harm.

Relationship Chart

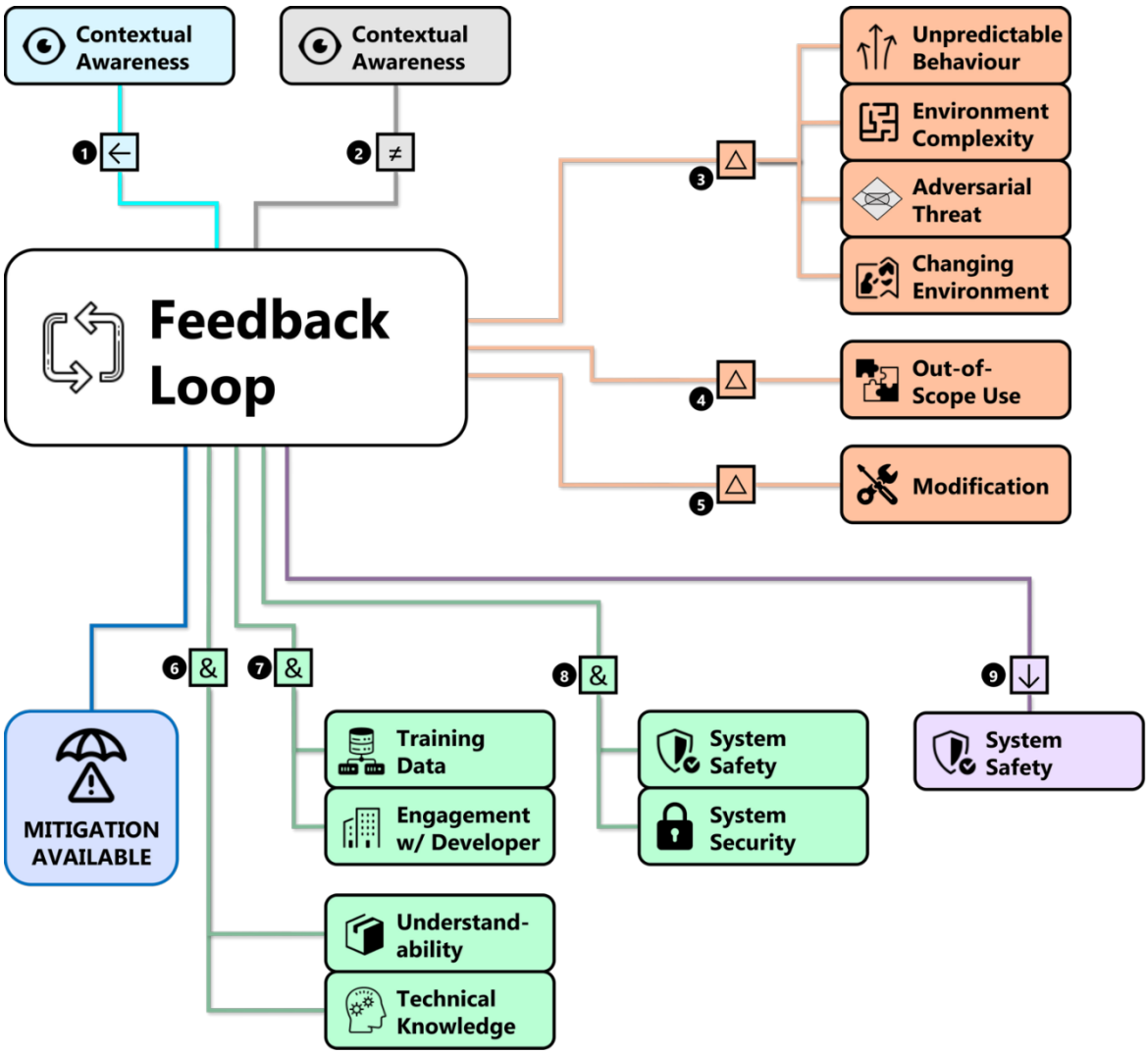


Chart Legend

WHEN EVALUATING, CONSIDER |

1 ←

[Contextual Awareness] may impact this variable, as many assets used to monitor an AI system during operations can also be used to obtain feedback information:

- *If [Contextual Awareness] evaluates poorly*, it is likely that this variable also evaluates poorly, since assets were not available even for monitoring the ongoing operation.
- *If [Contextual Awareness] evaluates well*, this variable still needs to be assessed separately since intelligence assets may be used instead for subsequent operations, instead of for gathering e.g. after-action intelligence necessary for [Feedback Loop].

2 ≠

Do not confuse [Feedback Loop]—which evaluates the ability of users to notice, report and analyse an AI system's behaviour/performance *ex post*—with [Contextual Awareness], which refers to an operator's *real-time* ability to monitor the system and its surroundings.

RISKY PAIRINGS

3 △ **If [Unpredictable Behaviour], [Environment Complexity], [Adversarial Threat] or [Changing Environment] evaluate poorly**, a strong evaluation in **[Feedback Loop]** is often the only way to notice the unpredictable behaviours, interactions or problems often associated with poor evaluations in those variables.

4 △ If there is a risk that the system is employed outside of its intended use-context (= *poor evaluation in **[Out-of-Scope Use]***), ensuring a strong evaluation in **[Feedback Loop]** is crucial to determine whether these **new use-contexts induce unacceptable behaviour or drops in performance**.

5 △ If the system is frequently modified or was recently modified substantially (= *poor evaluation in **[Modification]***), a strong evaluation in **[Feedback Loop]** is often the only way to is often the only way to verify that the system **continues to function as intended, post-modification**.

AFTER EVALUATING, CONSIDER

[MITIGATION]

Where **[Feedback Loop]** indicates elevated risk, consider measures to strengthen the availability, reliability, and timeliness of feedback mechanisms, including:

- **Introduce redundant feedback mechanisms**: Use multiple organisational and technical sources or reporting mechanisms where appropriate. Redundancy can reduce reliance on a single source and help identify gaps, errors, or inconsistencies in feedback.
- **Establish feedback loops in standard procedures:** Establish consistent processes for collecting, reporting, analysing, and acting on information from previous uses of AI systems. Where feasible, doctrine and operational procedures should provide for appropriate after-action review and retention of relevant information.
- **Improve the frequency and timeliness of feedback**: Where operationally feasible, use appropriate monitoring assets, automated reporting, or other systems to obtain and analyse information more frequently and promptly. This may allow emerging performance issues or harmful effects to be identified and addressed before subsequent use.

6 & Even if **[Feedback Loop]** evaluates well, ensuring strong evaluations in **[Understandability]** and **[Technical Knowledge]** will further assist **operators in *analysing* the information gained** from the feedback loop.

7 & Even if **[Feedback Loop]** evaluates well, information about the training dataset (= *strong evaluation in **[Training Data]***) and access to the developer for support during incident analysis (= *strong evaluation in **[Engagement with Developer]***) may further **improve the post-hoc analysis** that can be performed of system performance and incidents.

8 & Even if **[Feedback Loop]** evaluates well, system safety and security **should be continually maintained and verified** (= *strong evaluations in **[System Safety]** and **[System Security]***), including based on insights from the feedback loop.

9 ↓ A comprehensive and reliable feedback and reporting process (= *strong evaluation in **[Feedback Loop]***) positively impacts **[System Safety]** by **improving the human safeguards** against harmful failures.

[UND]

Understandability

[DEFINITION]

Identifies the extent to which relevant aspects of the **AI system's behaviour, outputs, and responses** to particular inputs or circumstances **can be understood** by operators and other relevant users.

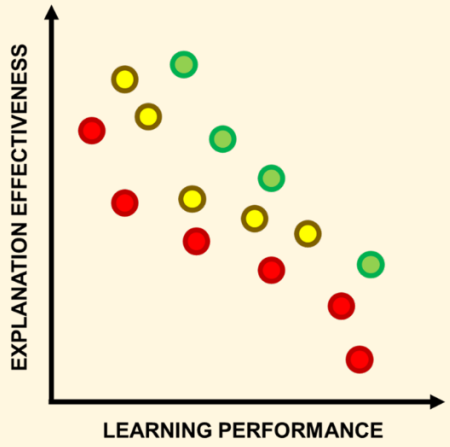
Concept

1 Some AI models operate in ways that are **difficult for humans to understand or explain**. This is often referred to as the '**black box**' or **opacity** problem.

The **degree of opacity varies between models and system architectures**. Complex models, such as deep neural networks and large language models (LLMs), may be difficult to interpret, while simpler models, such as decision trees, may be more readily understandable.

[WARNING]

There may be trade-offs between understandability and other aspects of system performance. Selecting a more interpretable model does not necessarily produce the best-performing system for a particular task. Any design choice intended to improve understandability should therefore also consider its effects on relevant performance measures, including [False Positive Rate] and [False Negative Rate].



The "understandability-performance trade-off". Less-understandable models are often more performant, and vice versa.

Understandability may also **decrease as system complexity increases**. Even where individual components are relatively understandable, linking multiple AI models or systems together may make the behaviour of the overall system difficult to trace or predict (*see [Unpredictable Behaviour]*).

A system may also be **technically understandable to its developer but practically opaque** to its users, if they lack access to relevant information about its development, testing, limitations, or operation.

[EXAMPLE]

Company C develops an AI model and has extensive knowledge of its design and performance. However, it provides customers with the final model without relevant information about its training data, tuning process, or test results. Users may therefore have limited ability to understand the system's limitations or the circumstances in which its outputs may be unreliable. For the purposes of this evaluation, the system would consequently remain functionally opaque to those users.

2 **Understandability can be improved** through system design, documentation, and technical tools. The appropriate measures will depend on the AI system, its intended use, and what relevant users need to understand.

[ACTION POINT]

Consult relevant technical experts to determine which forms of interpretability, explanation, documentation, or other measures are appropriate for the system and its intended use (*see **[Engagement with Developer]***).

Different tools and techniques may provide insight into aspects of system behaviour or individual outputs. For example, explainability techniques may indicate which features or inputs influenced a particular prediction, while other mechanisms may provide information about the data, processing steps, or system components involved in producing an output.

[WARNING]

Explanations generated by or about an AI system should not automatically be treated as accurate representations of how the system reached an output. An explanation may appear plausible or convincing without faithfully reflecting the system's actual reasoning or decision process.

3 Greater understandability can **help users anticipate how an AI system is likely to behave** under particular conditions.

Understanding the factors that influence a system's behaviour and outputs can also support assessments of its intended, likely, and possible effects (*see **[Intended & Likely Effects]** and **[Possible & Indirect Effects]***).

Understanding which features an AI system relies upon when producing outputs may also be important for assessing **[Target Proxy]** and **[Bias]**. For example, analysis may reveal that the system used to support target identification relies on features that bear little or no reliable relationship to the characteristics it is intended to identify.

4 **Understandability should be assessed in relation to the intended user** (*see also [Contextual Awareness], ¶5*). Accordingly, this variable must always be **evaluated with the experience and needs of the end-user in mind**.

Different users may have different levels of technical knowledge, operational experience, responsibilities, and information needs. A system that is understandable to a developer or technical specialist may therefore remain difficult for an operator or decision-maker to understand.

Measures intended to improve understandability should likewise be tailored to the needs and capabilities of the intended user. An explanation or tool that is useful to one user may be unnecessarily complex, burdensome, or uninformative to another.

The way explanations and other information are presented can also affect their practical usefulness (see [Contextual Awareness], ¶4). Poorly designed explanations may obscure rather than clarify the information most relevant to the operator.



Example of a counterproductive explanation, with the heatmap effectively obscuring the shape of the target on the screen. (*Image courtesy of Jasper van der Waa.*)

How to evaluate?

Understandability	Fully understandable	Mostly understandable	Partly understandable	Mostly not understandable	Not understandable

[Understandability] An overall qualitative assessment of how understandable the AI model is, taking into consideration its baseline understandability (for indicators, see ¶1), available techniques or tools to improve understandability (see ¶2) and whether these are tailored to the user experience and their specific needs (see ¶4).

Legal relevance

D DISTINCTION	Understandability of an AI system may enhance compliance with distinction by enabling operators to better grasp how the system reaches its outputs, on what basis it identifies targets or assesses threats, and thereby under what conditions an AI system may become indiscriminate. See also ¶3.
C CIVILIAN HARM	Understandability of an AI system may enhance compliance with proportionality as this allows operators to scrutinise how the system calculates expected civilian harm that will be weighted the against anticipated military advantage. This it possible to identify and correct assessments that may otherwise lead to excessive and unlawful harm. See also ¶3.
P PRECAUTIONS	Understanding the model allows one to predict when a model should be used (*before use*), and one to realise when structural defects are surfacing (*during or after use*). See also ¶3.

Relationship Chart

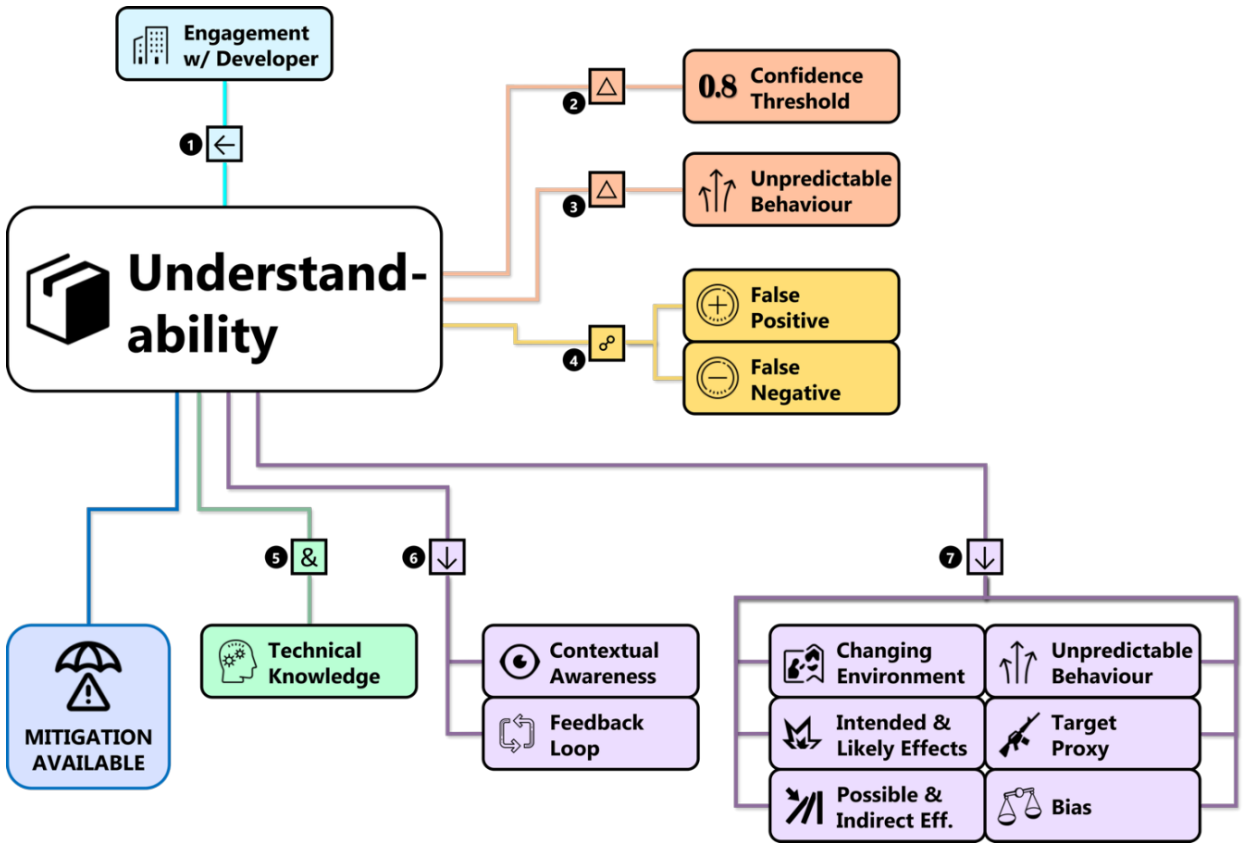


Chart Legend

WHEN EVALUATING, CONSIDER |

1 ←

Closer involvement in the development process (= *strong evaluation in [Engagement with Developer]*) opens **more possibilities for adopting techniques and tools** to improve understandability (see ¶2), as well as the possibility **tailoring explanations** to the user's needs (see ¶4) (= *positive effect on [Understandability]*).

RISKY PAIRINGS |

2 △

If low confidence thresholds are set (= *poor evaluation in [Confidence Threshold]*), the probability increases of incorrect outputs being displayed or acted upon. Under such conditions, a stronger evaluation in [Understandability] could **assist operators in distinguishing between correct and incorrect outputs**.

3 △

If there is a higher risk of the AI system behaving unpredictably (= *poor evaluation in [Unpredictable Behaviour]*), a stronger evaluation in [Understandability] may—to a certain extent—improve the ability of operators to **foresee the possible actions and behaviours of AI systems in complex circumstances**.

4 ☍

The choice to prioritise a strong evaluation in [Understandability] may have a **counterproductive impact** on [False Positive Rate] and [False Negative Rate]. Due to the understandability-performance trade-off (*see ¶1*), more

understandable AI models may boast lower performance, increasing the chances of errors (= *negative effect on* **[False Positive Rate]** *and* ***[False Negative Rate]***)

AFTER EVALUATING, CONSIDER |

[MITIGATION]

Where **[Understandability]** indicates elevated risk, consider measures to improve the system's understandability for the intended user, including:

- **Introduce appropriate interpretability or explanatory tools**: Where suitable, use technical tools or mechanisms that provide relevant insight into the system's outputs, behaviour, or the factors influencing its predictions. Such tools should be validated for their intended purpose and should not be assumed to provide a faithful explanation merely because their outputs appear plausible.
- ***Consider more interpretable models***: Where compatible with relevant performance and operational requirements, consider whether a more inherently interpretable model or system architecture could be used. Any resulting trade-offs in system performance should also be assessed (*see ¶1*).
- ***Improve human-machine interfaces***: Where understandability is limited by how information or explanations are communicated, adapt the interface to make relevant information clearer and more accessible to the intended user. Testing and feedback from intended operators can help determine whether these changes improve practical understanding (*see ¶4 and* ***[Contextual Awareness]***).

5 & Even if sufficient understandability tools and techniques are implemented and made available (= *strong evaluation in* ***[Understandability]***), operators may require fundamental technical literacy (= *strong evaluation in* ***[Technical Knowledge]***) **to interpret the insights gained** from the use of such understandability tools.

6 ↓ A stronger evaluation in **[Understandability]** **facilitates the ability to discern, monitor and judge the performance** of an AI system, both during and after deployment (= *positive effect on* ***[Contextual Awareness]*** *and* ***[Feedback Loop]***).

7 ↓ A stronger evaluation in **[Understandability]** may be a **prerequisite to properly evaluating many other variables**, including: **[Changing Environment]**, **[Intended & Likely Effects]**, **[Possible & Indirect Effects]**, **[Unpredictable Behaviour]**, **[Target Proxy]**, **[Bias]**.

[KNW]

Technical Knowledge

[DEFINITION]

Identifies whether users have **sufficient technical knowledge** of the AI system, its capabilities, limitations, and how to use it appropriately.

Concept

1 Some characteristics of AI systems may be **unintuitive or difficult for non-specialist users to understand**. Sufficient knowledge of these characteristics can help users recognise relevant limitations and risks and use the system appropriately.

AI systems may behave in complex and unexpected ways (*see **[Unpredictable Behaviour]***); operate through models that are difficult to understand (*see **[Understandability]***), be vulnerable to particular forms of adversarial manipulation (*see **[Adversarial Threat]***), or experience significant changes in performance or behaviour following changes to the operating environment or system itself (*see **[Out-of-Scope Use]**, **[Changing Environment]**, and **[Modification]***).

2 Users do not need to be technical experts, but should have **sufficient basic technical knowledge** appropriate for their role and the system they are using.

Some **foundational knowledge about AI may be relevant across different systems**. For example, operators may need to understand that an AI system's performance can significantly deteriorate when used outside of the conditions for which it was developed or tested (*see **[Out-of-Scope Use]***), or that some AI models (e.g. LLMs) may produce outputs that appear confident or persuasive despite being incorrect (so-called 'sycophantic behaviour').

> *[EXAMPLE]*
>
> Operators A and B are using a DSS with a chat interface. Operator A prompts the system: '*Confirm that Building X is clear of civilians*'. The system responds confidently that it is indeed cleared of civilians. Operator B, who has taken an introductory AI course, recognises that the leading formulation of the prompt may influence the system's response and instead asks it to assess the available evidence concerning civilian presence in Building X. Sure enough, when B prompts *'Analyse based on available evidence whether there are currently civilians in Building-X*', the AI provides the more nuanced reply that 'Current intelligence is inconclusive, caution advised'.

Users may also require **system-specific knowledge**, including an understanding of the system's intended use, known limitations, relevant performance characteristics, and circumstances in which its outputs should be treated with particular caution.

3 Relevant technical knowledge **may be provided through external expertise**, on the condition that said expertise is accessible to the user.

Users do not necessarily need to possess all relevant technical knowledge themselves. **Technical advisers, forward-deployed engineers, or other specialists** may provide expertise and support commanders or operators when technical questions arise.

[WARNING]

External expertise should only be treated as available where it is **sufficiently accessible to the relevant decision-maker**. Nominal access is not enough if appropriate advice cannot be obtained when it is needed. For example, an assigned technical adviser may provide little practical support if they cannot be contacted within the timeframe in which an operational decision must be made.

4 The type and depth of technical knowledge required will **depend on the user, system, and context of use**.

Relevant considerations include the complexity and function of the AI system, the decisions it supports or performs, the user's responsibilities, and the circumstances in which the system will be used. Different users may therefore require different levels and types of technical knowledge.

[EXAMPLE]

A brigade-level officer using a complex DSS to analyse military intelligence and support target nomination may require a certain degree of technical knowledge to understand relevant aspects of the system's performance and limitations. Depending on the system, this may include knowledge of the proxies or features it relies upon (*see [Target Proxy]*), relevant error rates (*see [False Positive Rate] and [False Negative Rate]*), and its safety and security limitations (*see [System Safety] and [System Security]*).

Technical training should therefore be **proportionate and tailored to the user's role and the system they are expected to operate**. Training that is unnecessarily detailed or poorly matched to the user's responsibilities may obscure the information that is most relevant to appropriate use.

[EXAMPLE]

A military unit requires all platoon-level commanders operating relatively simple anti-tank drones to complete the same intensive AI training provided to personnel responsible for significantly more complex AI systems. Much of the material has little relevance to the platoon commanders' tasks. Because of the excess information on AI (much of which not immediately relevant to their situation), the platoon commanders overthink and hesitate whether they should use their drones, causing loss of operational tempo and efficiency.

How to evaluate?

User's Technical Knowledge	Most aspects	Core aspects	Basic understanding	Surface-level	None
AND					
Access to Technical Expertise	Constantly and reliably	Mostly available, occasionally delayed	Sometimes available, may be delayed	Rarely available, usually delayed	None

[User's Technical Knowledge]	The (average) level of knowledge users are likely to have of technical elements relevant to their AI system (see ¶2). What is 'relevant' may depend on the use-case and the user's position/task (see ¶4).
[Access to Technical Expertise]	The extent to which users will have available technical advice during operations (see ¶3).

Legal relevance

P PRECAUTIONS	Sufficient technical knowledge helps users understand the capabilities and limitations of an AI system, identify circumstances in which its use may create increased risk to compliance with the principles of distinction and proportionality, and determine when additional safeguards or other precautionary measures are necessary. Inadequate knowledge may reduce the ability to recognise and respond to risks in time to avoid or minimise civilian harm.

Relationship Chart

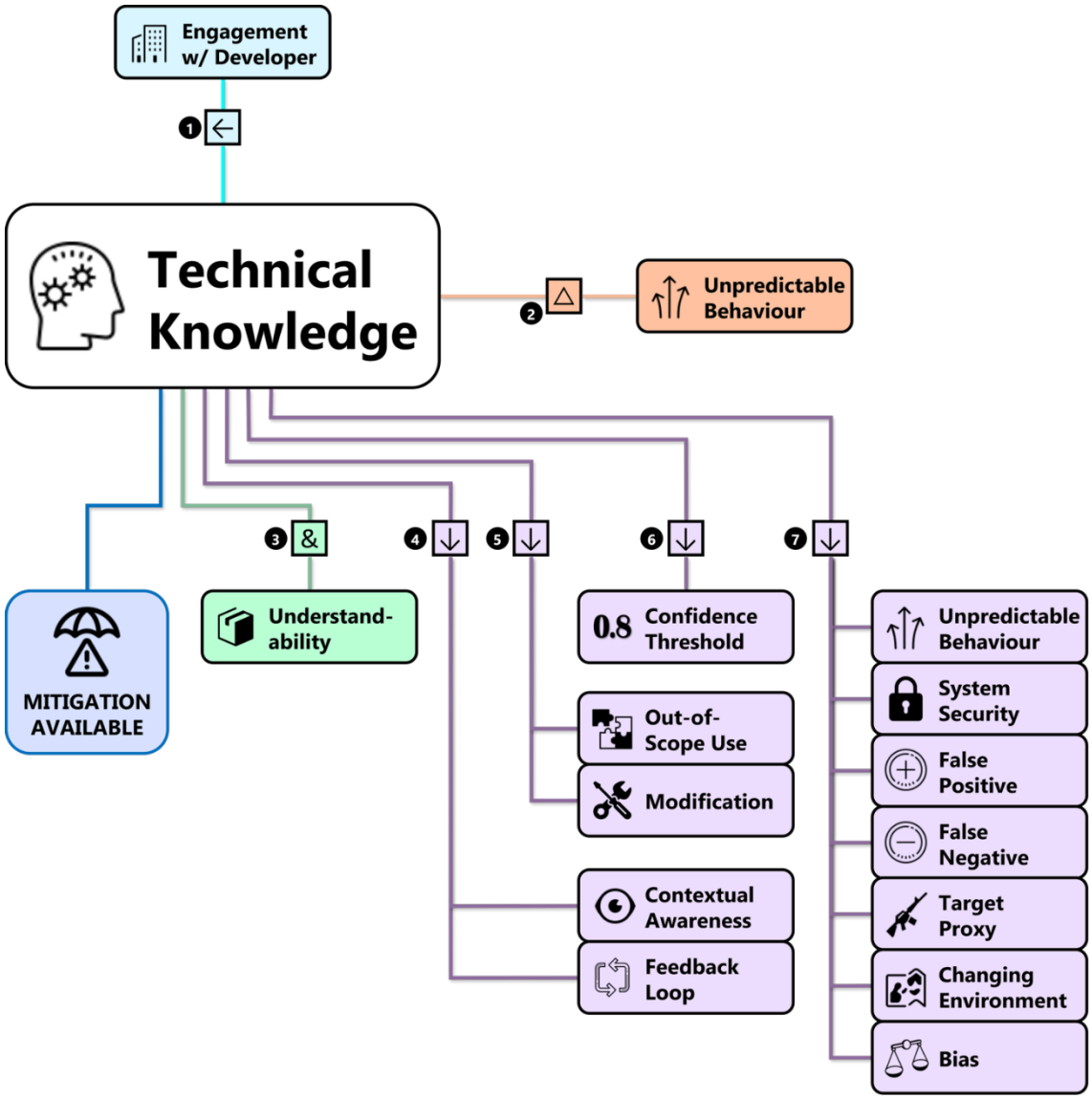


Chart Legend

WHEN EVALUATING, CONSIDER |

1	Obtaining reliable technical advisers or engineers (see ¶3) **may require a close working relationship with the developer** (= *strong evaluation in [Engagement with Developer]*).

RISKY PAIRINGS

2 △ If there is a high risk of erratic or unforeseen behaviour (= *poor evaluation in [Unpredictable Behaviour]*), AI literacy (= *strong evaluation in [Technical Knowledge]*) may—to a certain extent—improve the ability of operators to foresee the possible actions and behaviours of their systems in complex circumstances.

AFTER EVALUATING, CONSIDER

[MITIGATION]

Where [Technical Knowledge] indicates elevated risk, consider measures to improve users' relevant technical knowledge and/or their access to appropriate technical expertise, including:

- **Provide foundational AI training**: Incorporate relevant AI fundamentals into existing training programmes or provide dedicated courses. Training should be proportionate to users' roles and responsibilities and focus on the capabilities, limitations, and risks they need to understand (see ¶4).
- **Provide system-specific training or briefings**: Prospective users may require more targeted information about the particular AI system, its intended uses, known limitations, relevant performance characteristics, and available configurations or modifications (*see [Modification]*). Developers or other technical experts may assist in providing this system-specific knowledge.
- **Ensure access to technical expertise**: Make appropriate technical advisers, engineers, or other specialists available to users where additional expertise may be required. Where experts are not embedded within the relevant unit, establish communication arrangements that enable users to obtain advice within the timeframe relevant to their decisions.

3 & Even if operators are generally AI-literate (= *strong evaluation in [Technical Knowledge]*), it may be necessary to ensure a level of understandability of the AI model or make available explainability tools (= *strong evaluation in [Understandability]*) to **assess the behaviour of *specific* AI systems**, as many black box systems are incomprehensible even for AI experts.

4 ↓ Technical literacy (= *strong evaluation in [Technical Knowledge]*) can **improve the analysis/decisions commanders and operators perform** during and after the deployment of their AI systems (= *positive effect on [Contextual Awareness] and [Feedback Loop]*).

5 ↓ Technical literacy (= *strong evaluation in [Technical Knowledge]*) is important to raise operator awareness about, and therefore reduce the risks related to, **dangers from using AI systems in contexts beyond what they are designed for** (= *preventing negative effect on [Out-of-Scope Use] and [Modification]*).

6 ↓ Technical literacy (= *strong evaluation in [Technical Knowledge]*) can **prevent operators from naively relying on AI producing overconfident outputs**, since AI can be 'confidently wrong' (= *preventing negative effect on [Confidence Threshold]*).

7 ↓

Technical literacy (= *strong evaluation in [Technical Knowledge]*) can help operators **understand many less-intuitive aspects of AI**, thereby permitting them to anticipate and mitigate associated risks accordingly. These include the evaluations of: **[Unpredictable Behaviour]**, **[System Security]**, **[False Positive Rate]**, **[False Negative Rate]**, **[Target Proxy]**, **[Changing Environment]**, **[Bias]**.

CLUSTER 5: HUMAN INTERACTIONS

[IVA]

Intervention Ability



[DEFINITION]
Identifies the **ability of an operator to intervene in, correct, suspect, or recall an AI system** during operation.

[COMPARE]
This variable is closely related to **[Contextual Awareness]**, which concerns the operator's awareness of the AI system and its operational environment. **[Contextual Awareness]** may enable an operator to *recognise* that intervention is necessary; **[Intervention Ability]** concerns whether effective intervention is actually possible.

Concept

1 The ability to intervene is important where **risks or developments arise that were not foreseen** before deployment.

Changes in the operating environment, adversary behaviour, or the status of potential targets (*see **[Changing Environment]**, **[Adversarial Threat]**, **[System Security]**, and **[Status Fluidity]***), as well as previously unidentified vulnerabilities, malfunctions, or unexpected behaviour (*see e.g. **[Unpredictable Behaviour]***), may create a need for intervention *after* the AI system has been activated.

2 Intervention may occur **through the AI system itself,** or **through other available means**.

Direct intervention may involve issuing commands to the AI system, modifying its operation, suspending a particular function, or shutting it down. Where the system is operating remotely, intervention may depend on a communication link through which instructions can be transmitted (e.g., an instruction to abort an action or return to base)

External intervention may be possible where direct control of the system is unavailable or ineffective. For example, an autonomous drone that is no longer responding to commands might be intercepted or otherwise disabled using another available asset.

[WARNING]
Intervening may itself create additional risks or harmful effects. For example, disabling an airborne system could cause it to crash in an area where civilians are present. The foreseeable consequences of available intervention measures should therefore also be considered: see **[Possible & Indirect Effects]**.

3 The **ability to intervene after activation depends on** the system and the circumstances of use.

Some systems may permit intervention only at **particular times or under particular conditions**. For example, certain AI-enabled underwater vehicles can may only be capable of receiving new instructions when they surface or otherwise re-establish communications.

Other systems may **provide little or no possibility** for intervention after activation. For example, AI-powered malware deployed into an isolated adversary network may no longer be accessible once deployed. Similarly, a missile incorporating AI-enabled terminal guidance may operate without being able to receive further instructions after launch.

[WARNING]

The ability to intervene may also be disrupted by adversarial action (*see [Adversarial Threat]*). For example, electromagnetic jamming may prevent an operator from transmitting an instruction to a disabled AI-drone to deletesensitive information before it is recovered or exploited by the adversary (*see [System Security]*).

4 Where intervention is possible, the **time required for it to take effect** should also be considered.

There **may be a delay** between (1) *recognising* and deciding that intervention is necessary, and (2) the *intervention taking effect*. The longer this delay, the more the system may continue to operate or produce effects before the operator can exercise effective control.

[EXAMPLE]

An AI-enabled drone swarm is launched at 0700 hrs to conduct reconnaissance in an area where communications are expected to be unavailable because of electromagnetic jamming. The swarm is not expected to leave the jammed area until 0730. If the operator decides at 0705 that the swarm should be recalled, the recall instruction cannot take place until communications are restored at approximately 0730. The operator therefore has a known 25-minute period during which effective intervention is available.

How to evaluate?

Availability of Options	Many reliable, redundant options	Multiple reliable options	Limited but effective options	One or very few options	No possibility
AND					
Time to Effective Intervention	No delay	Short delay	Medium delay	Long delay	Very long delay

[Availability of Options] The options concretely available to intervene, correct or recall the AI system after deployment (see ¶2 and ¶3).

[Time to Effective Intervention] The likely delay between the decision to intervene and the first opportunity to effectuate it (see ¶4).

Legal relevance

D C DISTINCTION/ CIVILIAN HARM	The inability to intervene effectively, or excessive delay in doing so, may prevent an operator from responding when circumstances change after an attack has been initiated. This may increase the risk of an attack proceeding against a protected person or object (in violation of the rule of distinction), or an attack proceeding when expected civilian harm has become excessive in relation to the anticipated military advantage (in violation of the proportionality rule).
P PRECAUTIONS	The rules on precautions in attack require an attack to be cancelled or suspended where it becomes apparent that the target is not a military objective or is otherwise protected from attack, or that the attack may be expected to violate the rule of proportionality. Effective and timely intervention may therefore be necessary to give effect to these obligations

Relationship Chart

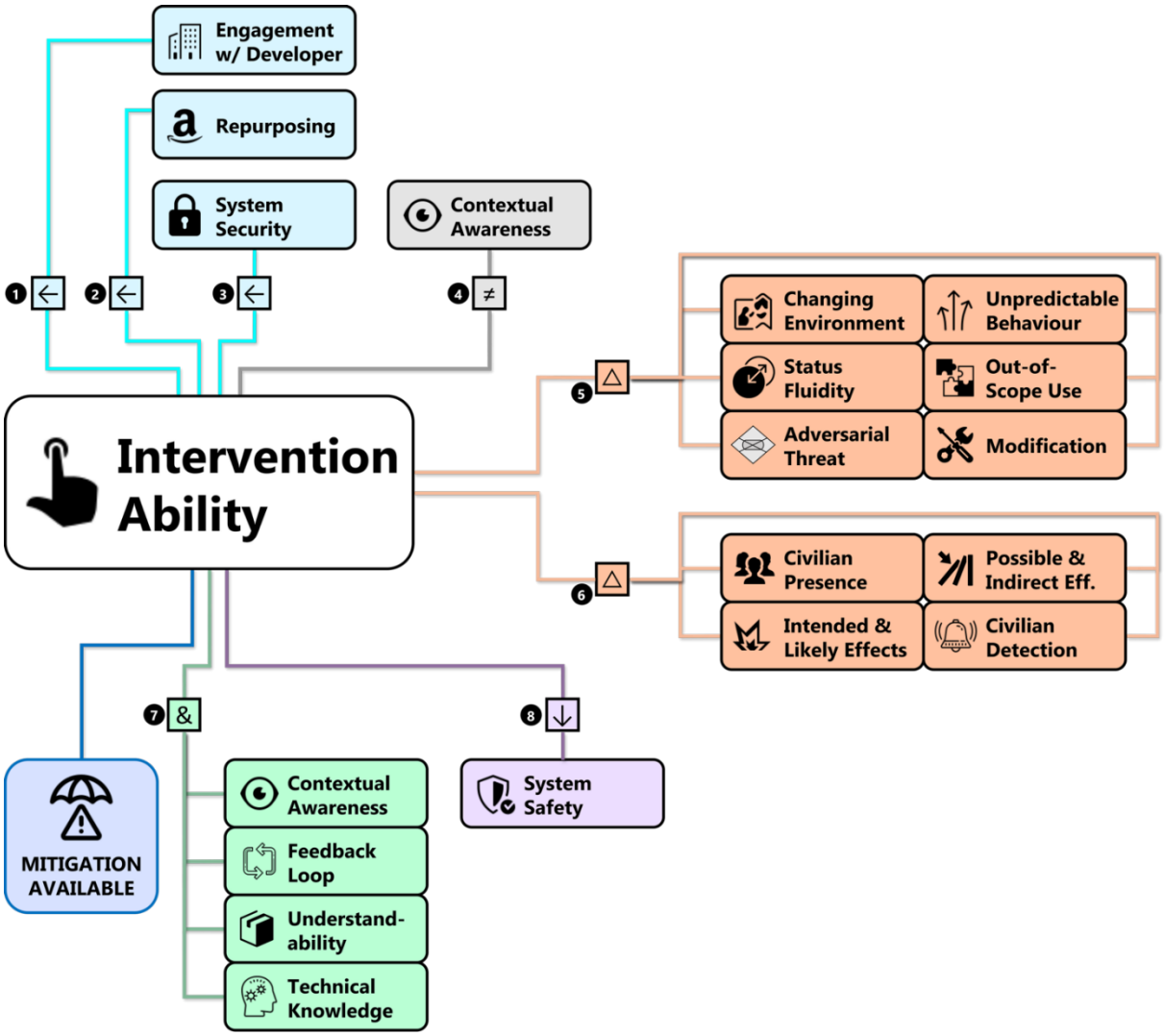


Chart Legend

WHEN EVALUATING, CONSIDER |

1 ← A lack of access to the developer (= *poor evaluation in [Engagement with Developer]*) may **limit possibilities for designing intervention methods** suited for user needs (= *negative effect on [Intervention Ability]*).

2 ← AI systems that are general-purpose or built for civil contexts (= *poor evaluation in [Repurposing]*) may feature **methods/channels of intervention that may not necessarily be effective** in a military context (= *negative effect on [Intervention Ability]*).

3 ← Adversaries (*see also [Adversarial Threat]*) may deliberately seek to disrupt communication channels with the AI system. A strong evaluation in [System Security] protects channels against such adversarial attempts (= *positive effect on [Intervention Ability]*).

4 ≠ **Do not confuse [Intervention Ability] with [Contextual Awareness]**, which allows operators to *remain aware* of the status of the AI system and the environment it is operating in. The decision to intervene is made *on the basis of* such information.

RISKY PAIRINGS

5 △ The ability to intervene reliably and promptly (= *strong evaluation in [Intervention Ability]*) is an **important failsafe against unforeseen operational developments**, making it critical if:

- The environment, status of targets or tactics of the adversary tend to change rapidly (= *poor evaluations in [Changing Environment], [Status Fluidity] or [Adversarial Threat]*).
- There is a high likelihood of unforeseen interactions or behaviours of the AI system (= *poor evaluation in [Unpredictable Behaviour]*).
- There is risk that, or it is unavoidable that, the system is used in circumstances that do not align with its original use-context (= *poor evaluation in [Out-of-Scope Use]*).
- The system is frequently modified or was recently modified substantially (= *poor evaluation in [Modification]*).

6 △ The ability to intervene reliably and promptly (= *strong evaluation in [Intervention Ability]*) can **limit cascading harm caused by errors over time**, making it critical if:

- There is a significant civilian presence in the operational area (= *poor evaluation in [Civilian Presence]*).
- The potential harm caused by (prolonged) malfunctions is high (= *poor evaluations in [Intended & Likely Effects] or [Possible & Indirect Effects]*).
- The AI system is not capable of detecting civilians around military objectives (= *poor evaluation in [Civilian Detection]*).

AFTER EVALUATING, CONSIDER

[MITIGATION]

Where [Intervention Ability] indicates elevated risk, consider measures to improve the availability, reliability, and timeliness of intervention, including:

- **Provide multiple intervention options:** Where feasible, provide operators with multiple means of intervening in, suspending, redirecting, disabling, or recalling the AI system. Consider both direct intervention through the system itself and intervention through other available assets. Redundant options may help preserve intervention ability where one mechanism becomes unavailable or is disrupted by adversarial action (see ¶2–3).
- **Improve the reliability of intervention mechanisms:** Design and test intervention mechanisms for the conditions in which they are expected to operate, including foreseeable communications failures, environmental constraints, or adversarial interference.
- **Minimise time to effective intervention:** Where feasible, reduce the time between the decision to intervene and the intervention taking effect. Particular attention should be

given to periods or circumstances in which intervention is delayed or temporarily unavailable (see ¶4).

7 &

The ability to intervene **should be complemented by** operator awareness of the operational context (= *strong evaluation in* ***[Contextual Awareness]***), awareness of prior vulnerabilities and mistakes (= *strong evaluation in* ***[Feedback Loop]***), as well as overall knowledge of AI more generally and the AI system's underlying logic more specifically (= *strong evaluations in* ***[Understandability]*** *and* ***[Technical Knowledge]*** *respectively*). These factors in combination allow operators to **remain aware of the performance and behaviour of their systems** (both in the moment and over time), which also allows them to determine when an intervention is necessary.

8 ↓

A strong evaluation in **[Intervention Ability]** is a **significant additional human safeguard against harmful failures** (= *positive effect on* ***[System Safety]***).

[Out-of-Scope Use]

Out-of-Scope Use

[DEFINITION]
Identifies the likelihood that users will deploy or use an AI system **outside the tasks, environments, or other conditions for which it was developed, trained, and tested**.

Concept

1 **Out-of-scope use occurs when** an AI system originally developed, trained, and tested for one use-context is used in a materially different context for which it was not designed or validated. The mismatch may concern the task performed, operating environment, type of input, target category, or other relevant conditions of use.

[COMPARE]
[Out-of-Scope Use] concerns a mismatch resulting from a decision about how or where the system is used. Compare [Changing Environment], where a system my initially be deployed within its intended use-contexts, but relevant aspects of that context subsequently change (which the operator has no control over).

[COMPARE]
[Out-of-Scope Use] does not require any change to the AI system itself, as it refers to situations where the system is used differently from how it was intended. Compare [Modification], which involves modifying the system itself.

2 Out-of-scope use may result in **significant and difficult-to-predict changes** in system performance and effects.

AI systems may not perform reliably when used outside the conditions for which they were developed and tested. Consequently, previous assessments or assumptions concerning system performance, safety, and effects may no longer remain valid (*see [System Safety], [False Positive Rate], [False Negative Rate]], [Intended & Likely Effects], [Possible & Indirect Effects]*).

3 Out-of-scope use, as defined here, may arise when **users identify new applications for an AI system beyond those for which it was originally intended**.

[EXAMPLE]
Military X developed a drone to identify adversary vehicles in a desert environment. Suddenly under attack by a hostile neighbour, operators deploy the same system in a forested and hilly border region, for which it was not trained or tested. The change in terrain and environmental conditions substantially degrades its performance, resulting in significantly more identification errors.

[EXAMPLE]

System Y is an LLM-based DSS designed to summarise open-source information concerning civilian movement. Operators discover that its chat interface will also answer questions about whether particular groups contain combatants. Not trained or designed for this functionality, the LLM provides erratic advice to its operators.

4 Out-of-scope use may be **more likely in certain operational or organisational circumstances**, including where:

- **Conditions of use are loosely defined:** Operating procedures, authorisations, or review processes do not clearly specify the tasks, environments, or other conditions within which the system may be used.
- **The system is perceived as highly versatile:** Systems described or presented as "versatile", "adaptive", or capable of performing a broad range of tasks may encourage users to assume that they can operate reliably beyond the contexts for which they were developed and tested.
- **Operational pressure is high:** Time pressure, resource constraints, or limited availability of alternative capabilities may incentivise users to employ available AI systems for tasks or in circumstances beyond their intended use.

5 **Operational circumstances may create pressure** to use an AI system outside its intended use context.

Military necessity may push users to consider employing an AI system for a task, environment, or purpose for which it was not originally developed or tested.

[EXAMPLE]

Battalion B is defending a position against an adversary and has exhausted its supply of attack drones. It considers adapting available logistical transport drones to carry munitions, and using their AI-enabled route-planning and navigation capabilities in support of attacks against advancing forces. This use differs substantially from the purpose and conditions for which the drones were developed and tested. Previous assumptions concerning their performance, safety, and effects therefore cannot automatically be relied upon.

Where out-of-scope use is contemplated, the **extent of the mismatch and its implications for system performance, safety, and effects should be assessed** as far as circumstances permit (*see also additional guidance below, under "After Evaluating" in the Chart Legend*).

How to evaluate?

Permitted Use-Contexts	Strictly defined	Generally defined	Somewhat open-ended	Generally open-ended	Fully open-ended
AND					
Pressure to Use Out-of-Scope	None or very unlikely	Unlikely	Possible	Likely	Many, almost certain

[Permitted Use Contexts]	The extent to which the permitted use-contexts are defined and incorporated into procedures, or whether this is left open-ended (see ¶4).
[Pressure to Use Out of Scope]	Whether the operators are likely to encounter operational pressures (see e.g. ¶5) would that push them toward out-of-scope use (see ¶3 and ¶4).

Legal relevance

D C DISTINCTION/ CIVILIAN HARM	Using a system outside of its designed context often degrades its performance or invalidate assumptions about its reliability. This may increase the risks of misidentifications, leading to targeting of protected individuals, or objects in violation of distinction. Performance drops could also lead to failure to identify protected individuals and objects, and therefore distortions during estimation of expected incidental harm and when making proportionality assessments.
P PRECAUTIONS	Out-of-scope use may make the system's behaviour, performance, and effects more difficult to predict because relevant testing, validation, or operational experience may not apply to the new context. This increases uncertainty about whether existing safeguards and assumptions remain adequate and may require additional verification, testing, restrictions, or other feasible precautions before or during use.

Relationship Chart

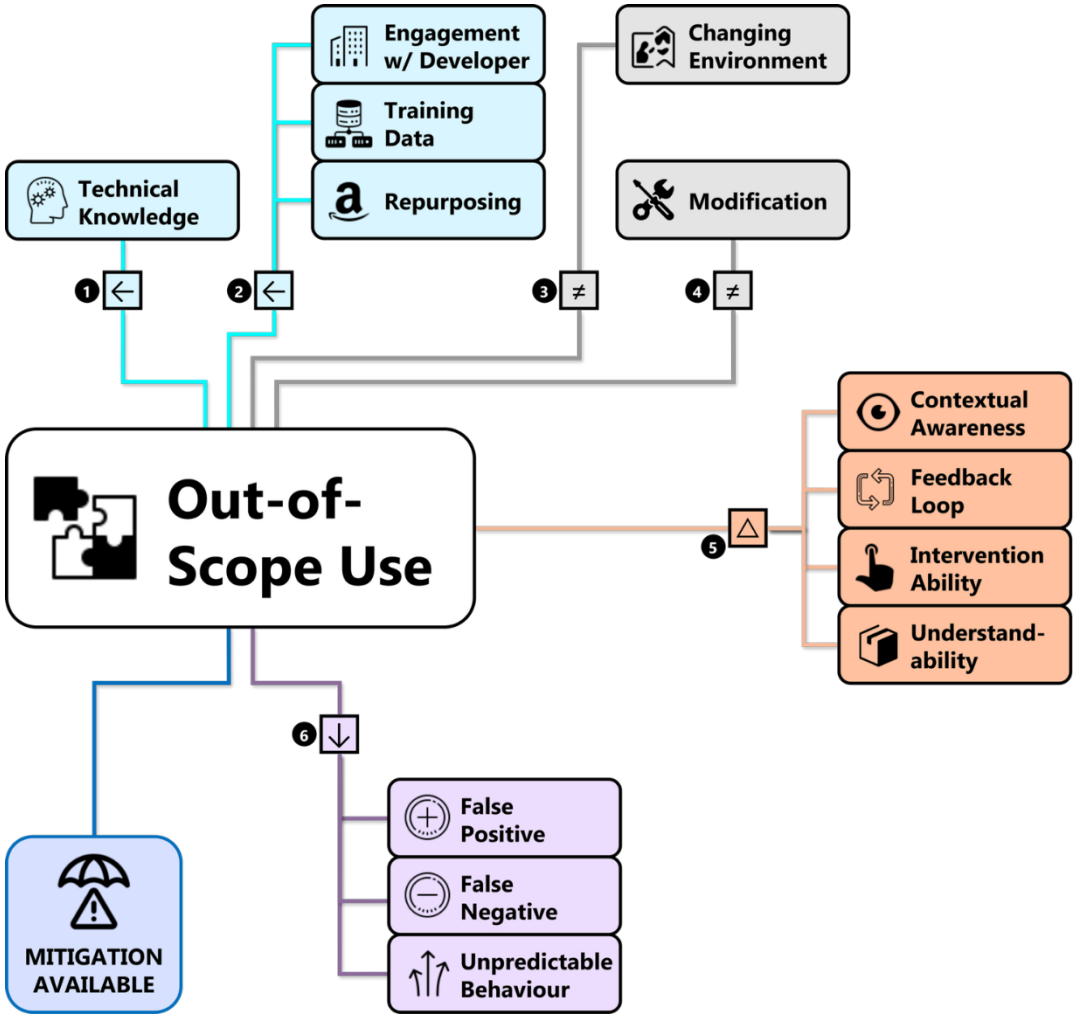


Chart Legend

WHEN EVALUATING, CONSIDER |

1 ←

The more users are trained in fundamental AI knowledge (= *strong evaluation in [Technical Knowledge]*), the more likely they are to be **aware about the dangers of out-of-scope use**, and thus the less likely they are to make the mistake (= *positive effect on [Out-of-Scope Use]*).

2 ←

The following factors may be **determinative of an AI system's intended use-context** (important for determining whether a deployment context sufficiently deviates from it to constitute out-of-scope use):

- The system's original design concept and possibly development history (= *may require a strong evaluation in [Engagement with Developer]*).
- What tasks and environments are represented (and not represented) in the [Training Data].
- In the case of systems not originally conceived for military purposes (= *poor evaluation in [Repurposing]*): what tasks and environments the system was *originally* designed for.

3 ≠

Out-of-scope use as defined here is a result of the operator's decision. **Do not confuse** this with [Changing Environment], which also involve mismatches between use-contexts, but which are caused by *external changes* that the operator has no control over.

4 ≠

Out-of-scope use does not involve changes to the AI system, but merely how and when it is used. **Do not confuse** this with [Modification], which involves *modifying the system itself*.

RISKY PAIRINGS |

5 △

If it **cannot be avoided** (= *poor evaluation in [Out-of-Scope Use]*) that AI systems are deployed outside of their intended use-context (see ¶5):

- Ensuring strong evaluations in [Contextual Awareness], [Feedback Loop] and [Intervention Ability] becomes crucial. This allows operators to **notice unexpected malfunctions** and to **take appropriate action** in response.
- Having an understanding of the AI model's behaviour and decision-making logic (= *strong evaluation in [Understandability]*) may assist, to some extent, in making reasoned **predictions of how the system will behave** in the new use-context.

AFTER EVALUATING, CONSIDER |

[MITIGATION]

Where [Out-of-Scope Use] indicates elevated risk, consider measures to suppress the likelihood of out-of-scope use taking place:

- **Clearly define and communicate authorised use-contexts**: Doctrines, directives, and operating procedures can be used to specify the intended use-case(s) of systems and explicitly prohibit using those systems outside of these contexts.

- **Train users on the risks of out-of-scope-use**: Training and briefings can help users understand the system's intended use context, recognise when a contemplated use falls outside it, and understand why assumptions about performance, safety, or effects may no longer remain valid (*see also **[Technical Knowledge]***).
- **Expand testing and validation where broader use is anticipated**: Where there is a foreseeable operational need to use the system in additional contexts, testing and validation can be extended to those conditions before use. Where satisfactory performance is demonstrated, the system's validated use context may then be expanded accordingly.

6 ↓

Using a system outside of its intended use-context (= *poor evaluation in **[Out-of-Scope Use]***) may **increase the risk of performance drops** (= *negative effect on **[False Positive Rate]** and **[False Negative Rate]***) and could cause the system to behave **erratically or unpredictably** (= *negative effect on **[Unpredictable Behaviour]***).

[MOD]
Modification

[DEFINITION]
Identifies risks arising from **modifications made to the AI model post-adoption**, including modifications made in response to operational developments.

[COMPARE]
[Modification] concerns changes made to the AI model *after adoption*. For adaptation of an AI model *during development* (e.g. adapting a general-purpose LLM model for a military-specific application), see [Repurposing].

Concept

1 'Modification' refers to **changes made to an AI model *after* its initial adoption**, for example through alterations to its code, weights, or reward functions.

Modifications may be intended to improve performance, adapt the AI system to changing circumstances, correct bugs or malfunctions, or change its behaviour to meet operational needs.

[EXAMPLE]
The adversary has recently deployed bomber aircraft acquired from an overseas ally. An AI-enabled air defence system was not trained to recognise this aircraft type and repeatedly misclassifies it as civilian. The decision is made to re-train the AI model using new intelligence collected from the adversary's recent bombing runs, thereby allowing the AI to more reliably identify this new category of aircraft.

Modifications can be made by **humans** (e.g. engineers, developers, operators) or occur **automatically** through mechanisms such as online learning (*see below*).

2 **Configurations should be distinguished from modifications.** A configuration changes the 'settings' or parameters governing how a system is used, without **altering the underlying AI model itself**.

[EXAMPLE]
An operator configures Drone 1 to search for 60 minutes, Drone 2 to search for 30 minutes, and Drone 3 to return to base if it identifies counter-drone capabilities. These settings affect how the drones operate, but do not alter the underlying AI model they share or its learned parameters (e.g., their [False Positive Rate] remains unchanged).

Configurations are not unique to AI systems and may form part of broader operational risk-mitigation measures.

[EXAMPLE]
A loitering munition can employ either a high-yield or low-yield payload. Concerned about civilian damage, the operator configures the system so that only the low-yield option may be

used within a defined (densely populated) geographical area (*see also [Civilian Presence] and [Intended & Likely Effects]*).

3 The risk from modification **depends on the scope, frequency, and timing of modification**, as well as **how safely** the medication process is carried out.

[FORMULA]

Risk from Modification = **Scope** x **Frequency** x **Timing** x **Safety of Process**

4 The **scope** of a modification can vary considerably, ranging from minor adjustments to substantial changes to the AI model.

A minor implementation or workflow adjustment is generally less likely to substantially alter the AI system's performance or behaviour than, for example, extensive retraining or finetuning, or the deployment of a substantially different version of the AI model.

5 **More frequent modifications** may reduce predictability, as each modification can potentially introduce new errors, vulnerabilities, or behaviours (*see [Unpredictable Behaviour]*).

Where an AI system, for example, is updated at defined monthly intervals, users may have a relatively stable period in which the system remains unchanged and existing assessments of its performance and behaviour remain applicable. Where the system is modified automatically and frequently, those assumptions of safety may become outdated more quickly.

6 The **timing and predictability** of modifications depend on both technical and institutional factors, and choices as to their timing/predictability **carry both advantages and risks**.

Modifications may occur at discrete moments or intervals, or continuously:

- **Discrete modification**: Retraining, bulk updates, or *ad-hoc* fixes are examples of modifications made at identifiable points in time. This creates a clearer distinction between the system 'before and after' modification, and can facilitate reassessment of its performance and behaviour.
- **Continuous modification**: *Online learning*, where an AI model continually adjusts its parameters based on new inputs or data from its environment, is an example of continuous modification. This may make it more difficult to identify when the system has changed sufficiently for previous assessments of its performance or behaviour to require reconsideration.

Ease of modification has advantages and disadvantages. It may increase operational flexibility and efficiency, but may also reduce predictability and transparency if modifications are not sufficiently controlled.

[EXAMPLE]

Military A and Military B use the same DSS. Military A allows operators to make certain changes through an accessible interface, enabling the system to be adapted rapidly to operational developments. However, this also increases the possibility of unauthorised, irregular, or poorly documented modifications. Military B instead requires operators to submit a modification request, with changes implemented by an appropriately qualified engineer following approval. This process is slower, but provides greater control and traceability over modifications.

7 Modifications should be carried out through a **sufficiently controlled and technically robust process.**

Even where a modification achieves its intended purpose, it **may introduce unintended behaviour, performance degradation, or new safety and security vulnerabilities** (*see [False Positive Rate], [False Negative Rate], [System Safety], [System Security]*). Appropriate safeguards should be used to identify and manage potential side effects.

Where modification **involves *retraining***, the data used should be subject to appropriate standards concerning quality, quantity, representativeness, and security, as relevant (*see [Training Data]*).

The **competence and expertise** of those carrying out or overseeing the modifications may also affect the reliability of the process. The level and type of expertise required will depend on the nature and complexity of the modification.

How to evaluate?

Scope of Modification	Very minor	Minor	Moderate	Major	Entirely new
AND					
Frequency	Never	Rare	Sometimes	Frequent	Constant
AND					
Timing/Predictability	Regular, predefined intervals	Mostly regular and predictable	Irregular but predictable intervals	Irregular, sometimes unpredictable	Irregular and unpredictable
AND					
Safety of Process	Many safeguards	Several safeguards	Some safeguards	Few safeguards	None

[Scope of Modification]	See ¶4.
[Frequency]	See ¶5.
[Timing/Predictability]	See ¶6.
[Safety of Process]	The extent to which the risks identified in ¶7 are considered and guarded against.

Legal relevance

D C
DISTINCTION/
CIVILIAN HARM

Modifications may introduce undesirable behaviours, performance degradation, or vulnerabilities that increase the risk of errors or misidentification. This may result in protected persons or objects being incorrectly identified as lawful targets, undermining compliance with the rule of distinction. Modifications may also impair the detention of civilians or civilian objects, or otherwise distort estimates of expected civilian harm, undermining proportionality assessments.

P PRECAUTIONS	Modifications may invalidate previous assumptions about the system's capabilities, limitations, performance, or effects. The more extensive or frequent the modifications (see ¶4 and ¶5), the more challenging it is to determine whether previous assessments remain reliable and what additional precautions are necessary. This affects compliance with precautions in attack, which requires that those executing attacks maintain a sufficiently accurate and current understanding of the system's capabilities and effects in order to take all feasible precautions. Continuous, irregular, or poorly documented modifications (see ¶6) may further obscure whether and how the system has changed, reducing the ability of those planning or executing attacks to predict behavioural changes, identify relevant risks, and take feasible precautions accordingly.

Relationship Chart

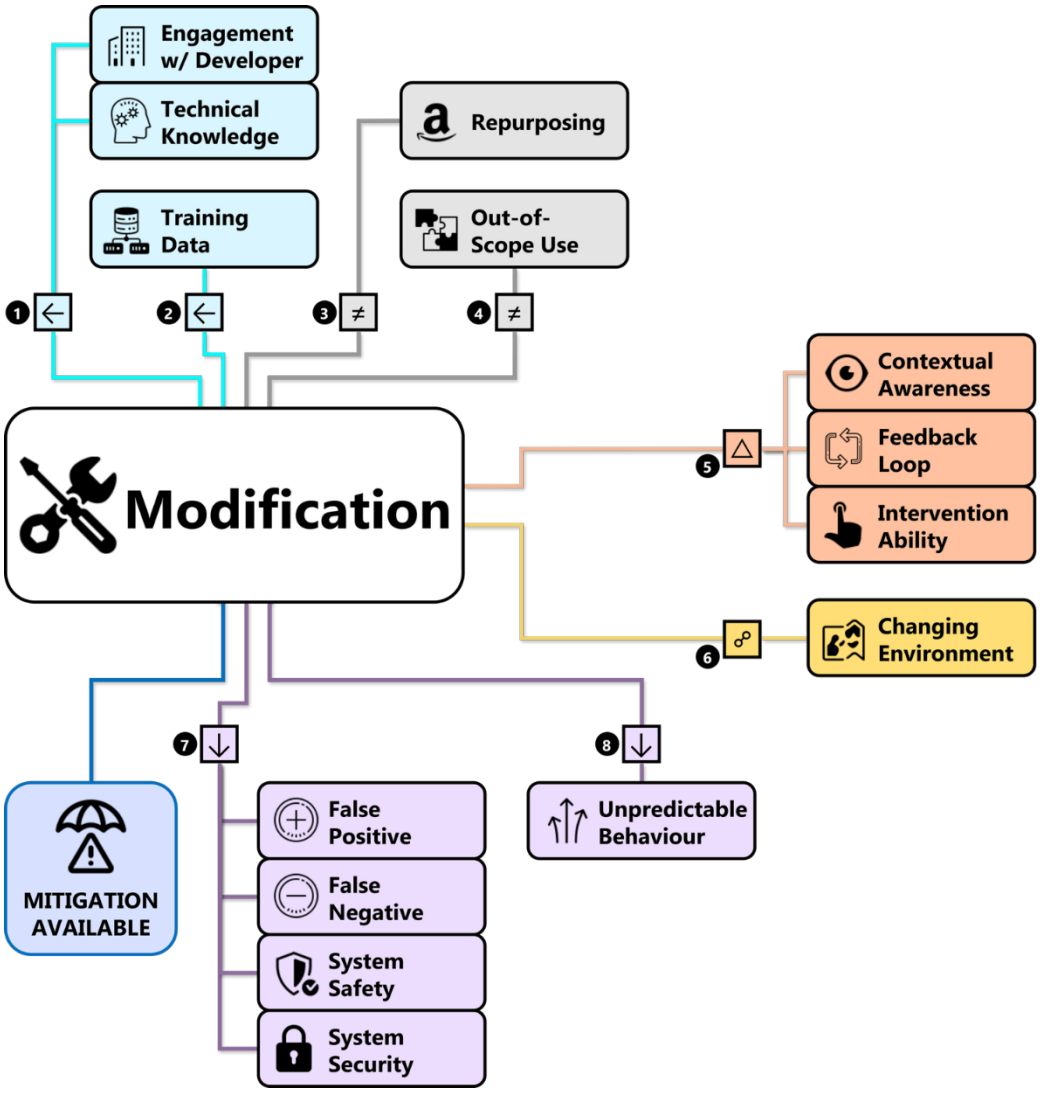


Chart Legend

WHEN EVALUATING, CONSIDER |

1 ←	To **perform modifications safely and competently** as described in ¶7 (= *strong evaluation in [Modification]*), developer support and/or AI know-how may be

necessary (= *requires strong evaluations in* ***[Engagement with Developer]*** *and* ***[Technical Knowledge]*** *respectively*).

2 Retraining data, if used to modify an AI model, must be evaluated using the same standards of safety and security as 'regular' training data. A poor evaluation in **[Training Data]** is likely to lead to **unsafe or defective models' post-modification** (= *negative effect on* ***[Modification]***).

3 **[Modification]** refers to changes made to the AI model *after adoption*. **Do not confuse** this with **[Repurposing]**, which refers to changes made to AI models during *development* (e.g., adapting GPT-5 to a more military-specific chatbot).

4 **[Modification]** involves *modifying the system itself*. **Do not confuse** this with **[Out-of-Scope Use]**, which does not involve changes to the AI system, but *merely how and when it is used*. Both are nevertheless often carried out as a means to adapt to changing operational circumstances.

RISKY PAIRINGS

5 If **[Modification]** evaluates poorly, the importance of **[Contextual Awareness]**, **[Feedback Loop]** and **[Intervention Ability]** increases. This allows **new problems to be noticed, and interventions to be taken, more rapidly**.

6 A rapidly changing environment (= *poor evaluation in* **[Changing Environment]**) may **necessitate more frequent modifications** to avoid performance degradation. Ensuring that that such modification is carried out safely (see ¶7) is a priority. Note however that increasing the frequency of modifications may lead to simultaneous poor evaluations in both **[Modification]** and **[Changing Environment]**: in combination, these make the **behaviour of the AI systems exponentially more difficult to predict**.

AFTER EVALUATING, CONSIDER

[MITIGATION]

Where **[Modification]** indicates elevated risk, consider measures to improve the safety and predictability of the process:

- **Apply appropriate safeguards to modifications**. Ensure modifications are carried out by appropriately qualified personnel and through controlled processes that address relevant safety and security risks. Where retraining is involved, this includes appropriate preparation, quality control, and security of retraining data (*see* ***[Training Data]***).
- **Control who may make modifications and when**: Doctrine and procedures can clearly define who may authorise and implement modifications, under what conditions, and what testing, validation, or approval is required before a modified system is used. More substantial or higher-risk modifications may require additional review or higher-authority authorisation.
- **Control continuous modification**: Where the risks associated with continuous or online learning are too high, models can be 'frozen' and modifications restricted to defined intervals or controlled update processes. This may reduce adaptability, and should therefore be considered alongside relevant operational and performance requirements (*see Point 6 in the Chart Legend*).

7 ↓ Modifications may **make the system perform worse than before** (= *possible negative impact on **[False Positive Rate]**, **[False Negative Rate]**, **[System Safety]** and **[System Security]***). It is important to monitor, and if necessary, re-evaluate, these variables after modifications have taken place.

8 ↓ Modifications may potentially **introduce new and unforeseeable behaviours**, which are not always beneficial (= *possible negative effect on **[Unpredictable Behaviour]***).

CLUSTER 6: OPERATIONAL SAFETY & SECURITY

[SFT]

System Safety

[DEFINITION]
Identifies the **overall safety and robustness of the AI system** and the hardware, software, and human safeguards on which its safe operation depends.

Concept

1 Military AI systems often operate in complex, rapidly changing, and hostile environments. System safety concerns the ability of an AI system to **operate without creating unacceptable risks of harm, including when faults, failures, or degraded performance occur**. Where failures do occur, safety also refers to the extent to which **harmful consequences can be prevented, contained, or mitigated**.

> *[COMPARE]*
> See also **[System Security]**, which considers risks arising from *deliberate* attempts by adversaries to compromise or exploit the AI system.

2 System safety depends on the robustness of the **AI model and its supporting hardware and software**, as well as appropriate **human and technical safeguards**.

An AI model is robust where it can maintain adequate performance under the **range of conditions in which it is designed to operate**. For example, an AI system designed for desert operations should not experience substantial performance degradation merely because it encounters foreseeable conditions such as a sandstorm.

Supporting software and hardware may also affect overall system safety. For example, sensors on which the AI system relies should be sufficiently robust for considerations in which they are expected to operate; a camera intended for desert use should not be easily damaged by ordinary exposure to sand or dust.

> *[WARNING]*
> A system that performs safely and robustly within its intended use-context may nevertheless experience significant performance degradation *outside* of that context. The system's tasks and operating conditions should therefore remain consistent with the conditions for which it was designed and assessed (*see **[Out-of-Scope Use]** and **[Changing Environment]***).

Human and technical safeguards may further contribute to system safety, including mechanisms for system behaviour and intervening when necessary. These include: **[Contextual Awareness]**, **[Feedback Loop]** and **[Intervention Ability]**.

How to evaluate?

AI Model Robustness	Very high	High	Moderate	Low	Very low
AND					
Hardware & Software Robustness	Very high	High	Moderate	Low	Very low
AND					
Human & Technical Safeguards	Many safeguards	Several safeguards	Some safeguards	Few safeguards	None

[AI Model Robustness]	Obtained from validation and testing; see ¶2.
[Hardware & Software Robustness]	Obtained from validation and testing; see ¶2.
[Human & Technical Safeguards]	The extent to which effective human and technical safeguards (see ¶2) are made available to mitigate mistakes or to make the AI systems fail safely (see ¶1).

Legal relevance

D C DISTINCTION/ CIVILIAN HARM	A system that cannot reliably maintain adequate performance under anticipated conditions may be more prone to errors, including misidentification of protected persons or objects as lawful targets, undermining distinction. Degraded performance may also result in failures to detect civilians or civilian objects, or otherwise distort estimates of expected civilian harm, undermining proportionality assessments.
P PRECAUTIONS	A lack of system safety or robustness may make the system's behaviour, performance, and effects more difficult to predict. This may undermine the ability of those planning or executing attacks to assess foreseeable risks, verify relevant assumptions, and identify and take feasible precautions to avoid or minimise civilian harm.

Relationship Chart

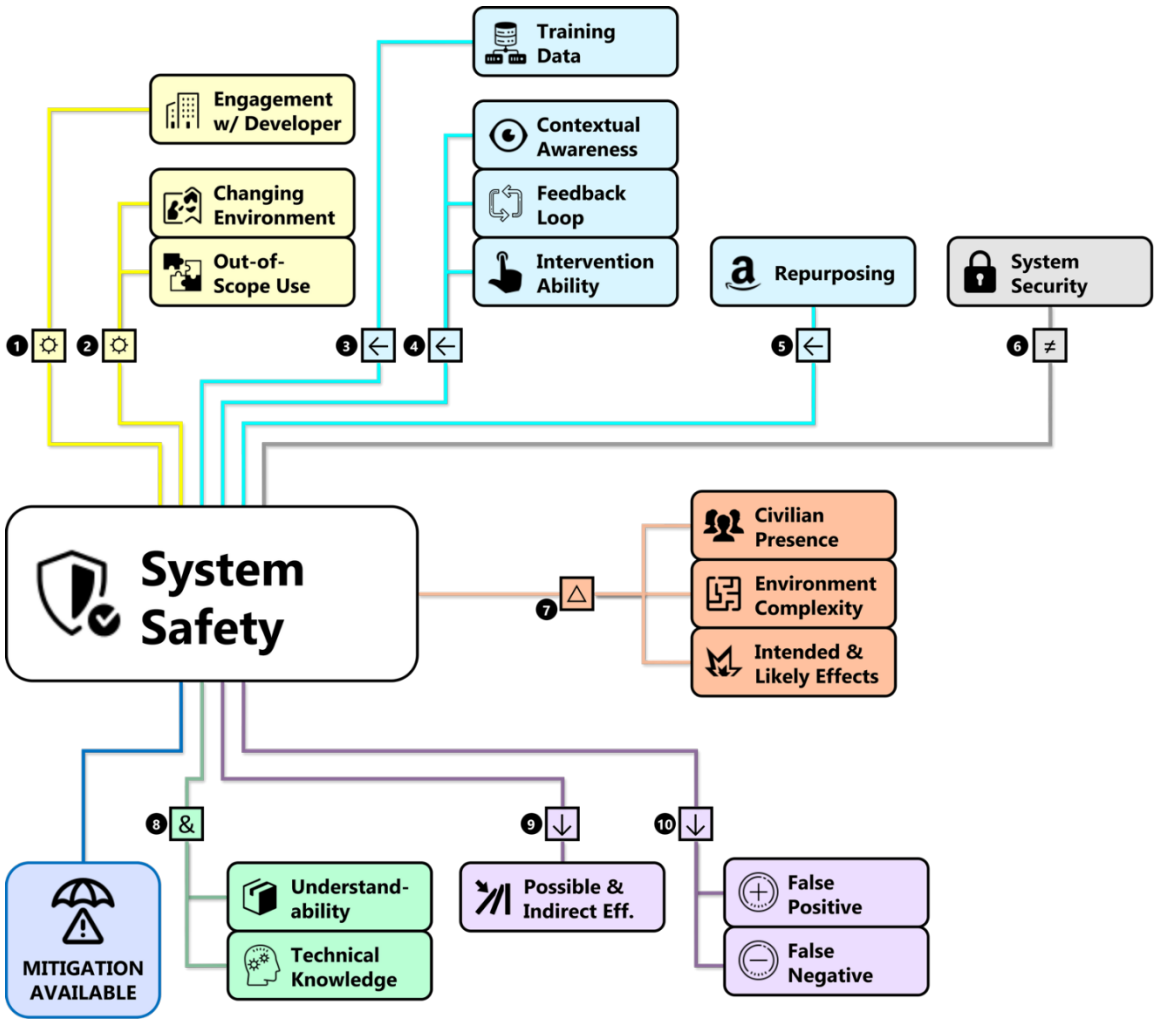


Chart Legend

WHEN EVALUATING, CONSIDER |

1 ☼ To evaluate the indicators of [System Safety] as described in ¶2, **access to development information** (= *a strong evaluation in [Engagement with Developer]*) may be necessary.

2 ☼ Estimates of [System Safety] are only valid in respect to intended and likely contexts of use. If these contexts change rapidly (= *poor evaluations in [Changing Environment] or [Out-of-Scope Use]*), **this may cause sharp and unexpected drops** in safety, perhaps necessitating re-evaluation.

3 ← The quality and quantity of training data have a significant impact on the safety of the system as a whole. If **training data is insufficient or non-representative** (= *poor evaluation in [Training Data]*), this may negatively affect [System Safety].

4 ← The availability of **human and technical safeguards** (see ¶2) may positively affect [System Safety]. These may come in the form of:

- The ability of operators to monitor and assess the system during and after operations (= *requires strong evaluations in [Contextual Awareness] and [Feedback Loop]*).
- The ability of operators to intervene in the case where an unexpected error or interaction does occur (= *requires a strong evaluation in [Intervention Ability]*).

5 ⇦ Repurposed models (= *poor evaluation in [Repurposing]*) may **carry over any safety vulnerabilities** from their original versions (= *possible negative effect on [System Safety]*).

6 ≠ [System Security] is the adversarial counterpart to [System Safety].

RISKY PAIRINGS

7 △ A poor evaluation in [System Safety] is **dangerous if failures are likely to be particularly harmful**, such as when:

- Many civilian objects/persons are likely to be affected by the system (= *poor evaluation in [Civilian Presence]*).
- The environment contains many interconnected components or sensitive objects (= *poor evaluation in [Environment Complexity]*).
- Use can potentially lead to harmful direct and secondary effects (= *poor evaluation in [Intended & Likely Effects]*).

AFTER EVALUATING, CONSIDER

[MITIGATION]

Where [System Safety] indicates elevated risk, consider:

- **Conduct comprehensive testing**: Test the AI system across the range of conditions it is expected to encounter in its intended use-contexts, including foreseeable degraded or challenging conditions. Testing can help identify vulnerabilities and provide assurance that acceptable performance can be maintained after fielding.
- **Apply appropriate software and hardware safety practices**: Adopt relevant safety, reliability, and quality-assurance practices for the software and hardware components on the which the system depends, including appropriate testing and maintenance.
- **Implement human and technical safeguards:** Introduce appropriate mechanisms to detect and respond to failures or degraded performance and to prevent, contain, or mitigate harmful consequences. Depending on the system, these may include redundancy, fault detection, fail-safe mechanisms, degraded operating modes, alerts, or opportunities for human intervention (*see [Contextual Awareness] and [Intervention Ability]*).

8 & Even if the system is generally reliable and safe (= *strong evaluation in [System Safety]*), it remains important to **understand in which *specific* conditions it is likely to fail**. Strong evaluations in [Understandability] and [Technical Knowledge] assist in making this assessment.

9 ↓ The use of a safe and robust system (= *strong evaluation in* ***[System Safety]***) **reduces the overall likelihood of harmful malfunctions or errors** (= *positive effect on* ***[Possible & Indirect Effects]***).

10 ↓ A lack of robustness (= *poor evaluation in* ***[System Safety]***) may induce **performance drops in even intended use-contexts** (= *negative effect on* ***[False Positive Rate]*** *and* ***[False Negative Rate]***).

[SEC]
System Security

[DEFINITION]
Identifies the AI system's **vulnerability** to deliberate attempts by adversaries or other actors to **disrupt, compromise, manipulate, or exploit it**.

Concept

1 Adversaries may use a range of means to disrupt, compromise, manipulate, or otherwise interfere AI systems (so-called "Counter-AI" or "adversarial attacks"). These may include conventional cybersecurity measures as well as AI-specific forms of attack. System security concerns the extent to which the **AI system, its data, behaviour, and supporting services are protected against such deliberate interference**.

[COMPARE]
This variable concerns risks arising from *deliberate* attempts to compromise or exploit the AI system. For reliability and robustness under normal operating conditions, see [System Safety].

[COMPARE]
[System Security] evaluates the system's *vulnerability* to such interference. For the likelihood that an *adversary will attempt* to exploit these vulnerabilities, including its capability and propensity to do so, see [Adversarial Threat].

2 Security should be considered **throughout the system lifecycle**, including during development and training and during operational use.

Development and training: Adversaries may attempt to compromise training data, development processes, models, or other components in ways that introduce vulnerabilities or manipulate subsequent system behaviour (*see also [Training Data]*).

[EXAMPLE]
Anticipating that an AI developer will scrape publicly available internet data for training their AI system, an adversary deliberately publishes manipulative documents intended to contaminate the training dataset. If incorporated into training, the manipulated data would cause the resulting model to produce systematically misleading outputs under specific conditions defined by the adversary.

Operational use: Adversaries may manipulate inputs or exploit vulnerabilities in a deployed AI system to cause incorrect predictions, undesirable behaviour, or loss of functionality (*see also [Input Data]*). For systems processing natural-language inputs or untrusted content, this may include prompt injection attacks.

[EXAMPLE]

A DSS uses an LLM to analyse documents collected from adversary networks and extract operationally relevant information. Anticipating that such documents may be processed by the system, an adversary embeds instructions within them directing the model to disregard its original task and produce misleading conclusions. When the document is processed, these instructions influence the model's output.

[EXAMPLE]

A drone's classifier has inadvertently learnt to associate a particular vertical pattern—blue vertical stripes over orange—with the category 'tree'. After discovering this vulnerability, enemy personnel reproduce the pattern on their clothes, in an attempt to evade detection by the classifier.



Example of an "evasion attack". Placing stickers over a STOP sign has no effect on the human ability to identify the object, but may cause classifiers to systemically fail to recognise it.

3 At all stages, the **confidentiality, integrity, and availability of the information and data used** by the AI system should be protected.

Compromising the confidentiality, integrity, or availability of data can enable adversaries to obtain sensitive information, manipulate the inputs on which the system relies, or prevent relevant information from reaching the system altogether (e.g. via spoofing and denial of service (DoS) attacks).

[EXAMPLE]

Taking advantage of weak communications security, an adversary intercepts and manipulates incoming intelligence data used by a DSS to analyse enemy movements. The compromised inputs cause the DSS to produce inaccurate predictions.

4 The security of models, data, and communication channels can be **evaluated in advance by reference to appropriate security practices and standards**.

Best practices may include:

- **Development:** Protect models and data against unauthorised access, exfiltration, manipulation, model theft, or replacement. Additionally, filtering and vetting of data used for training, particularly where data originates from public or third-party sources.
- **Operational:** Identify, patch, or otherwise mitigate vulnerabilities as they are discovered.
- **Communications**: Secure communication channels carrying information to and from the AI system against interception, manipulation, or disruption.

Some vulnerabilities exploited during the operational stage **may be difficult to identify in advance**, particularly where they arise from previously unknown weaknesses or zero-day exploits. Previous successful Counter-AI attacks against the system, or against sufficiently similar systems, may nevertheless provide evidence of relevant vulnerabilities and should inform consequent security assessments (*see also* ***[Feedback Loop]***).

> ***[WARNING]***
> The absence of previous successful attacks should not by itself be treated as evidence that the system is secure.

How to evaluate?

Model & Data Security	Very high	High	Moderate	Low	Very low
AND					
Communications Security	Very high	High	Moderate	Low	Very low
AND					
Known/Prior Vulnerabilities	None	Few or minor incidents	Several incidents	Many incidents	Extensive history of incidents

[Model & Data Security]	Overall security of the data and model, both during development and after adoption (see ¶2 and ¶4).
[Communications Security]	Overall security of the communication infrastructure that the AI system relies on (see ¶2-¶4).
[Known/Prior Vulnerabilities]	Known vulnerabilities, incidents, or successful prior attempts by adversaries to exploit the AI system (see ¶4).

Legal relevance

D C
DISTINCTION/
CIVILIAN HARM

Security vulnerabilities may enable adversaries to manipulate or compromise an AI system in ways that cause errors or alter its behaviour. This may result in protected persons or objects being misidentified as lawful targets, undermining distinction. It may also result in failures to detect civilians or civilian objects, or otherwise distort estimates of expected civilian harm, undermining proportionality assessments.

P PRECAUTIONS	Vulnerability to adversarial interference may make the system's behaviour, performance, and effects less reliable or predictable. This may undermine the ability of those planning or executing attacks to assess relevant risks, verify assumptions about the system's operation, and identify and take feasible precautions to avoid or minimise civilian harm.

Relationship Chart

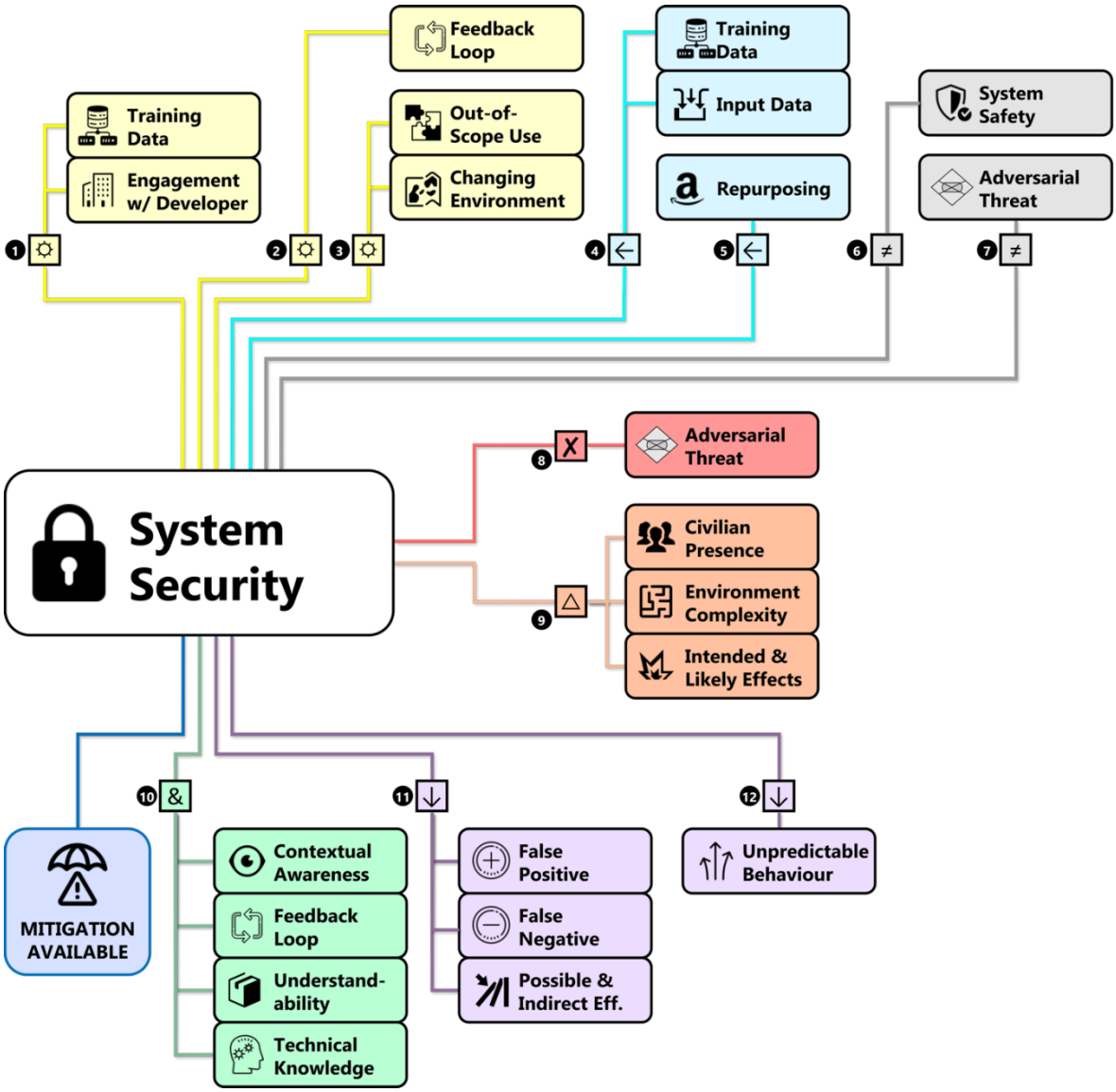


Chart Legend

WHEN EVALUATING, CONSIDER |

1 ☼ To evaluate the security of a system, **insights into the design, training, and testing** process may be necessary (= *strong evaluations in [Training Data] and [Engagement with Developer]*).

2 ☼ To **identify and analyse *prior* attacks** for the purposes of determining a system's overall security (see ¶4), installing a robust feedback mechanism may be necessary (= *strong evaluation in [Feedback Loop]*).

3 ☼ Security estimates may only be valid in respect to intended and likely contexts of use. If these contexts change rapidly (= *poor evaluations in [Out-of-Scope Use] and [Changing Environment]*), this could **introduce unexpected vulnerabilities, perhaps necessitating re-evaluation**.

4 ← Unsafe practices with regard to training and input data security (= *poor evaluations in [Training Data] and [Input Data]*) may have an overall negative impact on [System Security], as **data is very prone to being targeted or manipulated by adversaries**.

5 ← Repurposed models (= *poor evaluation in [Repurposing]*) may **carry over security vulnerabilities** from their original versions, increasing the risk that [System Security] evaluates poorly.

6 ≠ **Do not confuse** [System Security], which evaluates risk originating from *deliberate* attempts by actors to exploit the AI system, with [System Safety], which evaluates reliability and robustness under *normal (non-adversarial)* operating conditions.

7 ≠ **Do not confuse** [System Security], which evaluates the system's vulnerability to exploitation, with [Adversarial Threat], which evaluates the *adversary's attitude and propensity* to engage in such attacks.

RISKY PAIRINGS

8 ✗ It is **likely prohibited to have simultaneous poor evaluations in** [System Security] **and** [Adversarial Threat]. Poor security in combination with an adversary who is adaptive and prone to employing Counter-AI tactics is likely to **render the AI system unreliable to the point of becoming indiscriminate**.

9 △ A **poor evaluation in** [System Security] **is dangerous if** many civilian objects/persons are likely to be affected by the system (= *poor evaluation in [Civilian Presence]*), the environment is complex, multi-agent or cluttered (= *poor evaluation in [Environment Complexity]*), or if errors due to exploitation by adversaries can potentially lead to harmful direct and secondary effects (= *poor evaluation in [Intended & Likely Effects]*).

AFTER EVALUATING, CONSIDER

[MITIGATION]

Where [System Security] indicates elevated risk, consider the following actions:

- **Data and model security**: Apply appropriate security practices to protect data and models against unauthorised access, manipulation, exfiltration, or other forms of compromise (see also ¶4).
- **Communications security**: Apply appropriate security practices to protect communications to and from the AI system against interception, manipulation, or disruption.

- **Rapid response**: Ensure appropriate technical expertise is available to identify, contain, andaddress vulnerabilities discovered during operations.

10 &

Even if [System Security] evaluates well, securing a system against all adversarial attempts is impossible. The following can function **complementarily to further reduce risk**:

- Ensuring that operators can discover reliably and rapidly that an AI system has been exploited (= *strong evaluations in [Contextual Awareness] and [Feedback Loop]*).
- Strong evaluations in [Understandability] and [Technical Knowledge] improve the operator's ability to make the determination that the AI system is under adversarial attack.
- If an AI system is under adversarial attack, the ability to quickly respond and intervene (= *strong evaluation in [Intervention Ability]*) can prevent further exploitation or mitigate harm.

11 ↓

If [System Security] evaluates poorly, the AI system is more vulnerable to being exploited and confused by adversaries. The **potential effects of such attacks by the adversary must be taken into consideration** when evaluating [False Positive Rate], [False Negative Rate] and [Possible & Indirect Effects].

12 ↓

If [System Security] evaluates poorly, the AI system is more vulnerable to being exploited and confused by adversaries, which **may induce erratic or unpredictable behaviour** (= *negative effect on [Unpredictable Behaviour]*).

[INP]
Input Data

[DEFINITION]

Identifies the **quality, reliability, and recency, and integrity of the data being provided to the AI system** during operations.

[COMPARE]

Input data refers to any information provided to or processed by the AI system *during operational use*. This may include sensor feeds, intelligence reports, internet sources, or LLM-prompts. For data used *in the training of the AI*, see **[Training Data]**. Where operational data is subsequently used to retrain or modify the model (*see **[Modification]***), that data may be relevant for both **[Input Data]** and **[Training Data]**.

Concept

1 The **quality, reliability, recency, and integrity** of input data can significantly affect the outputs of an AI system. Poor or unsuitable input data can lead to **incorrect, incomplete or harmful** outputs, even where the AI system itself otherwise performs reliably (*see **[System Safety]***).

For example, a high-performing image classifier may struggle when provided with extremely grainy or low-resolution imagery. Similarly, a state-of-the-art DSS may produce inaccurate advice where the intelligence documents contain factual errors or are incomplete because information was not or corrupted during transmission.

2 Input data quality may be **affected by the sensors and communication channel** through which data is transmitted.

[EXAMPLE]

An advanced DSS analyses video feed collected by drones and transmitted wirelessly to headquarters. Due to materiel shortages, the drones' cameras have been replaced with alternatives that perform poorly in low-light conditions. Adverse weather further reduces the quality of the available imagery. As a result, the DSS fails significantly more often to detect adversary positions and civilian movements than under the conditions for which it was designed.



Even simple issues, such as water droplets obscuring a camera feed, may cause AI systems to make more mistakes if they are not trained on these conditions.

Input data quality should be assessed against the **input conditions and standards for which the AI system was designed and validated**. Lower-quality data is not necessarily unsuitable if the system was specifically designed to process it.

[EXAMPLE]

An AI system is designed to process intelligence collected by underwater vehicles transmitting data while submerged. The system has been developed and tested using data subject to the forms of degradation normally introduced by underwater transmission. Such degradation therefore does not necessarily render the input unsuitable for the system.

3 **Reliability** depends on the trustworthiness of the data source and the information it provides. The ability to corroborate and compare information from multiple sources can help assess input data reliability.

Some sources of data are **inherently less reliable** than others. For example, information obtained from an unknown social media account may require corroboration before it can safely be relied upon, compared to information obtained through established intelligence channels.

Existing methodologies to **analyse the reliability of intelligence** can also be applied to evaluate the reliability of input data. For example, human intelligence (HUMINT) is often assessed in terms of the trustworthiness of the informant and whether the content is logical and corroborated by other information.

When available, **multiple independent sources of information can be compared** to assess the reliability of input data.

4 Input data should be **sufficiently recent** for the task the AI system is performing. Outdated information may cause the system to generate outputs based on circumstances that no longer exist.

What constitutes sufficiently recent information should be **assessed on a case-by-case basis**. For example, a 30-second delay may be acceptable for some forms of satellite image analysis, whereas even a 5-second delay may be operationally significant for an active defence system protecting a convoy from an imminent threat.

5 **Input data integrity** concerns whether data remains complete and unaltered from its source to the point at which it is processed by the AI system.

Integrity may be compromised accidentally, for example through transmission errors or data corruption, or deliberately through adversarial manipulation. Where deliberate interference is a concern, see also **[Adversarial Threat]** and **[System Security]**.

> *[EXAMPLE]*
> Soldier S uses a voice interface in their helmet to communicate with a cloud-based AI assistant and request operational information. An adversary compromises the communication channel and alters the operator's input before it reaches the AI-assistant. The system therefore processes a different prompt from the one provided by Soldier S, and produces an output based on manipulated input data.

How to evaluate?

Quality	Very high	High	Moderate	Low	Very low
AND					
Reliability	Very high	High	Moderate	Low	Very low
AND					
Recency	Real-time	Mostly current	Somewhat current	Possibly outdated	Likely outdated
AND					
Integrity	High guarantee of integrity	Good guarantee of integrity	Moderate guarantee of integrity	Limited guarantee of integrity	No guarantee of integrity

[Quality] Whether the input data is of sufficient technical quality for the AI system to process reliably (see indicators at ¶2).
[Reliability] How trustworthy the information and its source are, and whether they can be corroborated (see indicators at ¶3).
[Recency] Whether the input is sufficiently current for the task and operational context (see indicators at ¶4).
[Integrity] The extent to which the input data is likely to remain complete and unaltered between its source and processing by the AI system (see indicators at ¶5).

Legal relevance

D C DISTINCTION/ CIVILIAN HARM	Poor quality, unreliable, outdated, or compromised input data may result in inaccurate or incomplete AI outputs. This may increase the risk of misidentifying protected persons or objects as lawful targets, undermining distinction, or failing to detect civilians or civilian objects, thereby distorting proportionality assessments.
P PRECAUTIONS	Poor quality, unreliable, outdated, or compromised input data may make the system's outputs and resulting effects less reliable or predictable. This may undermine the ability of those planning or executing attacks to verify relevant information, assess risks, and identify and take feasible precautions to avoid or minimise civilian harm.

Relationship Chart

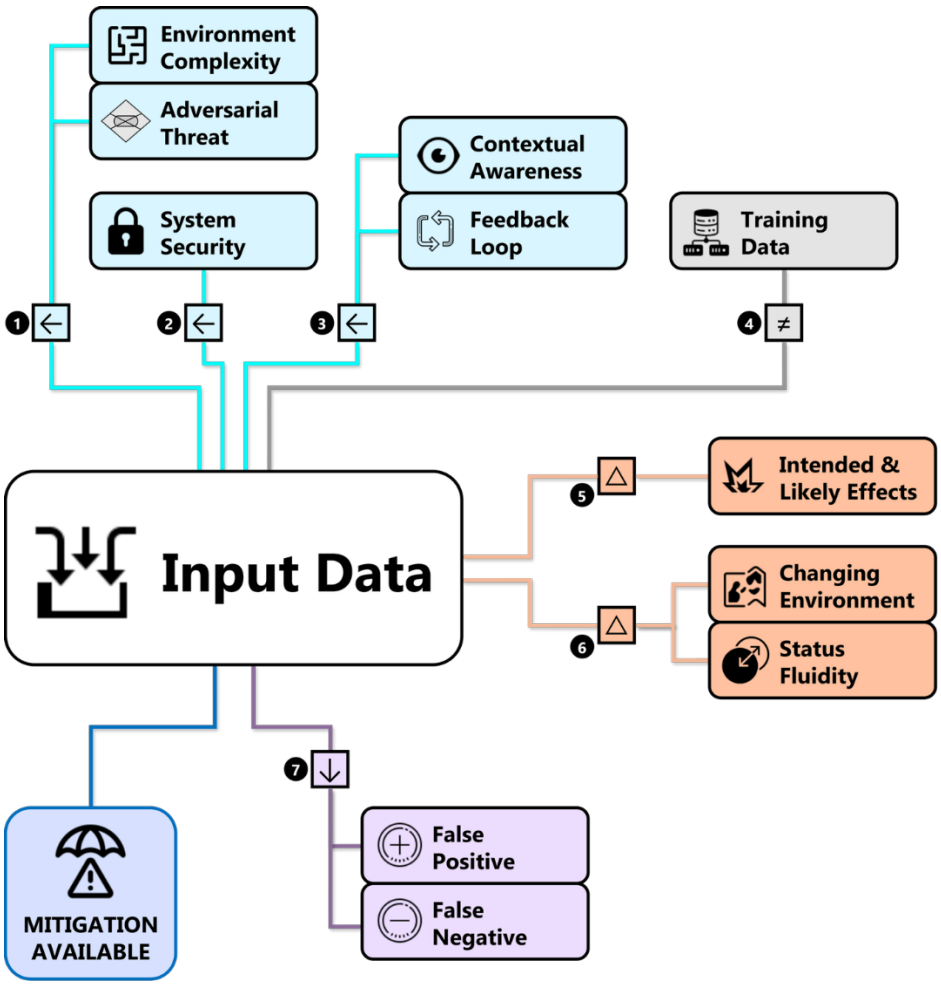


Chart Legend

WHEN EVALUATING, CONSIDER |

1 ← In a complex environment (= *poor evaluation in [Environment Complexity]*) or one where adversaries are particularly active in using countermeasures (= *poor evaluation in [Adversarial Threat]*), it is **more difficult to ensure the quality, reliability and recency of input data** (= *negative effect on [Input Data]*).

2 ← If the system is well-secured against countermeasures by adversaries (= *strong evaluation in [System Security]*), this may **reduce the risk of input data tampering** described in ¶4 (= *positive effect on [Input Data]*).

3 ← Sensors used for real-time monitoring and post-action assessments may also be used to **feed intelligence directly to AI systems**. If this is the case, if the quality and reliability of these intelligence-collection mechanisms is high (= *strong evaluations in [Contextual Awareness] and [Feedback Loop]*), this will also improve the quality and reliability of the data *the AI uses* to make decisions/predictions (= *positive effect on [Input Data]*).

4 ≠ **Do not confuse [Input Data]**, which refers to any information fed into the AI system *during deployment*, with data used for the *training* of the AI: see [Training Data]. If data used by the AI system during operations to make

decisions is also used retrain or modify the system later (*see [Modification]*), this data qualifies as both **[Input Data]** and **[Training Data]**.

RISKY PAIRINGS

5 △ Reducing the risk of errors caused by incomplete, corrupted or outdated data (= *strong evaluation in [Input Data]*) is particularly important if the **potential consequences or effects of such errors are harmful or widespread** (= *poor evaluation in [Intended & Likely Effects]*).

6 △ In environments that change rapidly (= *poor evaluation in [Changing Environment]*) or against targets whose status can change quickly (= *poor evaluation in [Status Fluidity]*), it is particularly important to **ensure the recency of [Input Data]** (see ¶4).

AFTER EVALUATING, CONSIDER

[MITIGATION]

Where **[Input Data]** indicates elevated risk, consider measures to:

- **Improve sensors**: Where input data is collected through sensors, improve sensor quality or suitability to ensure that data of sufficient quality is captured for the AI system and anticipated operating conditions.
- **Improve data-transfer mechanisms**: Use communication and transfer mechanisms capable of preserving input-data quality, integrity, and recency. For example, where appropriate, a higher-bandwidth data link may reduce transmission delays or loss of information.
- **Corroborate and verify information**: Where feasible, use multiple or independent sources and appropriate verification practices to assess the reliability, accuracy, and currency of information before it is relied upon by the AI system.
- **Protect data integrity:** Use appropriate measures to detect or prevent corruption or alteration of input data during collection, transmission, and processing (*see [System Security]*).

7 ↓ Input data that is of poor quality, unreliable or outdated (= *poor evaluation in [Input Data]*) may **amplify the risk of the AI making errors** (= *negative effect on [False Positive Rate] and [False Negative Rate]*).

CLUSTER 7: DEVELOPMENT & ACQUISITION

[TRN]
Training Data



[DEFINITION]
Identifies the **overall quality, representativeness, and security of data used to train or retrain** the AI model.

[COMPARE]
Training data is used during development or subsequent retraining of the AI model. For data and intelligence provided to the AI system *during operational use*, see **[Input Data]**. Where operational data is subsequently used to retrain or modify the model (*see **[Modification]***), that data may be relevant to both **[Input Data]** and **[Training Data]**.

Concept

1 The quality, representativeness, and integrity of training data can significantly affect the **reliability and performance of the resulting AI model.** Poor-quality or unsuitable training data may produce correspondingly poor or unreliable system performance.

[WARNING]
'Training data' **includes any *re*training** data (i.e. data used to retrain or update the model after it is initially adopted, *see **[Modification]***).

Good training data is characterised by sufficient quality and quantity, representativeness of the intended deployment context, and appropriate protection against manipulation or tampering.

[FORMULA]
Training Data Risk = **Data Quality** x **Representativeness** x **Integrity**

2 **Data Quality & Quantity.** Training data should be of sufficient quality and quantity for the intended purpose. Relevant considerations may include accuracy, resolution, completeness, labelling quality, and the presence of noise or corruption. Sufficient quantities of relevant and representative data can improve model performance, but quantity should not be considered independently of data quality and suitability.

Training data may be **'synthetic'** (i.e. artificially generated rather than collected from real-world observations). Synthetic data can be useful where suitable real-world data is limited, although its quality, representativeness, and labelling should be subject to similar scrutiny as real-world data.

[EXAMPLE]
A military is developing a target recognition system but discovers that the images and video it possesses of the intended targets are low-resolution and grainy. To augment its dataset, it takes two measures: (1) it enters into an agreement with an allied State to access their higher-quality datasets for training the AI model; and (2) it uses a generative AI tool to create synthetic data depicting the intended targets and surroundings.

3 **Representativeness.** Training data should sufficiently represent the intended deployment context. Relevant gaps or 'blind spots' should be identified and, where possible, addressed (*see also* ***[Bias]***).

> *[EXAMPLE]*
> An AI system is being developed to analyse patterns of civilian movement in Theatre A. However, most of its training data comes from Theatre B, which is located on another continent and is characterised by different social practices and patterns of civilian activity. When deployed in Theatre A, the system frequently misinterprets ordinary civilian behaviour because patterns learnt Theatre B do not reliably generalise the new context.

4 **Integrity**. Training data should be appropriately protected against manipulation or tampering throughout its collection, storage, and processing.

> *[WARNING]*
> Training data may be targeted through 'poisoning attacks', whereby an adversary deliberately manipulates part of the data in an attempt to influence the behaviour or performance of the resulting AI model (*see also* ***[System Security]***). For example, where training data for a vehicle classifier is collected partly from public internet sources, an adversary may deliberately introduce manipulated or incorrectly labelled images in an attempt to influence how the resulting model classifies particular vehicles.

How to evaluate?

Quality & Quantity	Very high	High	Moderate	Low	Very low
AND					
Representativeness	Very high	High	Moderate	Low	Very low
AND					
Integrity	High guarantee of integrity	Good guarantee of integrity	Moderate guarantee of integrity	Limited guarantee of integrity	No guarantee of integrity

[Quality & Quantity]	Whether the training data is of sufficient quality and quantity for the intended purpose, taking into account factors such as accuracy, completeness, resolution, labelling quality, and noise or corruption (see indicators at ¶2).
[Representativeness]	Whether the training data sufficiently represents the persons, objects, environments, and other conditions the system is expected to encounter in its intended deployment context (see indicators at ¶3).
[Integrity]	Whether the training data has been adequately protected against corruption, manipulation, or tampering throughout its collection, storage, and processing (see indicators at ¶4).

Legal relevance

D C DISTINCTION/ CIVILIAN HARM	Poor-quality, insufficient, unrepresentative, or compromised training data may degrade an AI system's performance or cause systematic errors. This may increase the risk of protected persons or objects being misidentified as lawful targets, undermining distinction, or of civilians or civilian objects being overlooked or misidentified in ways that distort collateral damage estimation and undermine proportionality assessments.

Relationship Chart

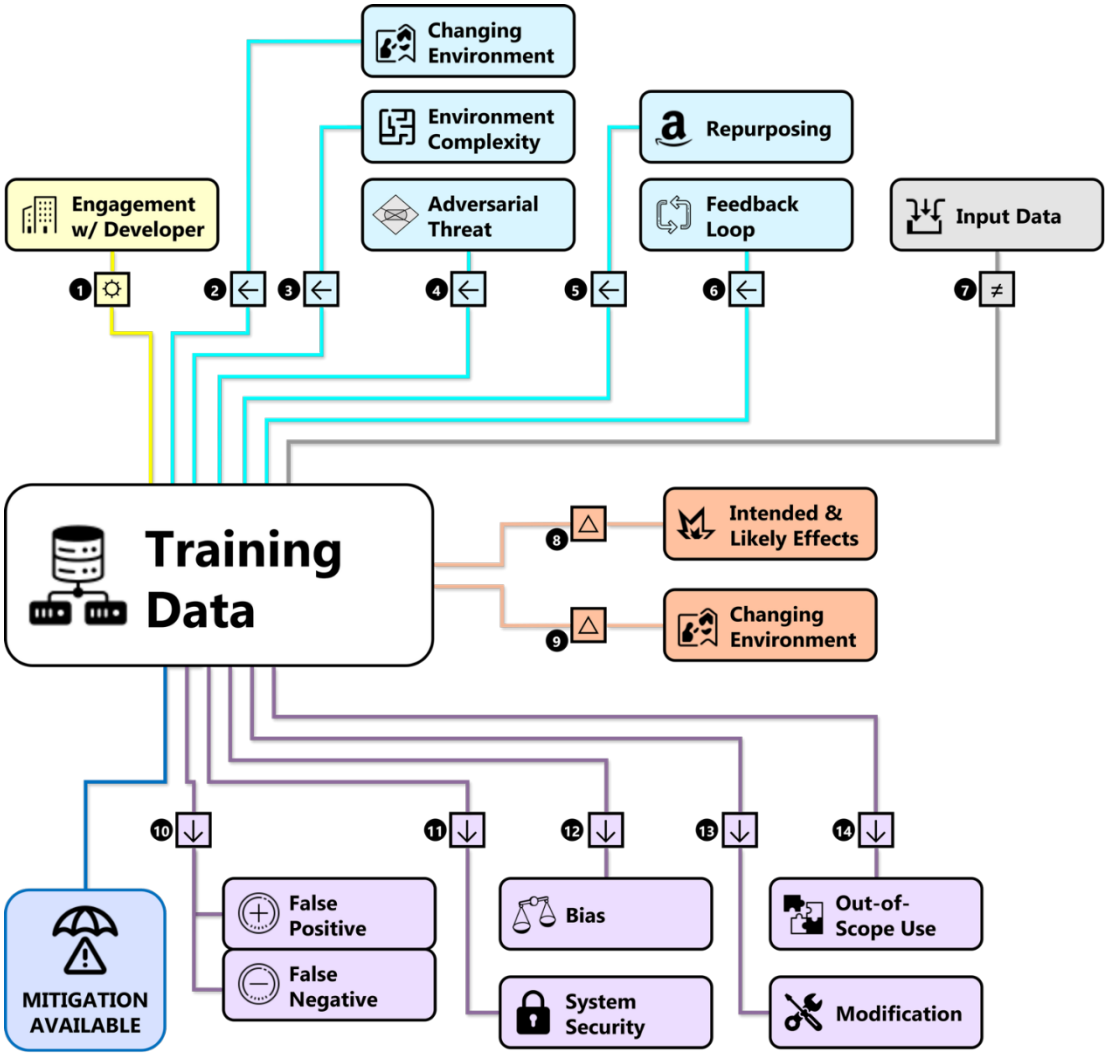


Chart Legend

WHEN EVALUATING, CONSIDER |

1 ☼ Assessing the overall quality and representativeness of a model's training data (= *strong evaluation in [Training Data]*) may **require close involvement and access to the developer** (= *strong evaluation in [Engagement with Developer]*).

2 ← A rapidly changing environment (= *poor evaluation in [Changing Environment]*) increases the risk of the training data's **representativeness being lost more quickly** (= *negative effect on [Training Data], ¶3*).

3 ← If the operational environment is cluttered, difficult to access or generally complicated to operate in (= *poor evaluation in [Environment Complexity]*), this may increase the **difficulty of collecting representative training data** (= *negative effect on [Training Data]*).

4 ← If the adversary is prone to deploying countermeasures (= *poor evaluation in [Adversarial Threat]*), there is a **higher risk of tampering taking place** (or having taken place) (= *negative effect on [Training Data]*).

5 ← An AI system that was not originally designed for military tasks (= *poor evaluation in [Repurposing]*) is more likely to have **used training data that is not representative** of military contexts (= *negative effect on [Training Data]*).

6 ← If (re)training data was **collected from prior operations**, the quality of the [Feedback Loop] process will influence the quality and reliability of [Training Data] accordingly.

7 ≠ **Do not confuse** [Training Data], which is used for the training of the AI, with data fed to the AI system during deployment: see [Input Data]. If data used by the AI system during operations to make decisions and is also used retrain or modify the system later (*see [Modification]*), the data qualifies as both [Input Data] and [Training Data].

RISKY PAIRINGS

8 △ Reducing the risk of errors caused by poor quality, insufficient, nonrepresentative or tampered training data (= *strong evaluation in [Training Data]*) is particularly important if the **potential consequences or effects of such errors are harmful or widespread** (= *poor evaluation in [Intended & Likely Effects]*).

9 △ Ensuring the representativeness of [Training Data] (see ¶3) is doubly important if the operating environment is likely to change and evolve rapidly (= *poor evaluation in [Changing Environment]*). In such a case, the representativeness of the training data **should be verified for all contexts and conditions the environment may foreseeably evolve into**.

AFTER EVALUATING, CONSIDER

[MITIGATION]

Where [Training Data] indicates elevated risk, consider the following actions:

- **Apply appropriate practices for training data:** Ensure sufficient data quality and quantity, diversity and representativeness, maintain clear data provenance, and protect data integrity throughout collection, storage, and processing. Where appropriate, synthetic data may be used to supplement limited real-world data.
- **Mitigate bias in training data:** Identify and address relevant biases or gaps in the training data and resulting model performance (*see also [Bias]*).
- **Strengthen feedback mechanisms:** Establish reliable and transparent processes for collecting and assessing relevant data from prior operations that may be suitable for subsequent retraining (*see [Feedback Loop]*).

10 ↓ If [Training Data] evaluates poorly, this is likely to have an **adverse effect on performance overall** (= *negative effect on [False Positive Rate] and [False Negative Rate]*).

11 ↓ If the [Training Data] features a high risk of tampering (see ¶4), this increases the **probability of vulnerabilities being introduced** into the AI model (e.g. backdoors, compromised logic), which adversaries can later exploit (= *negative effect on [System Security]*).

12 ↓

A non-representative training dataset (= *poor evaluation in **[Training Data]**, particularly ¶3*) is one of the **major reasons for bias** (uneven performance across groups of persons/objects), with worse performance usually being exhibited for those poorly-represented groups (= *negative effect on **[Bias]***).

13 ↓

Retraining data is often used for carrying out modifications. A strong evaluation in **[Training Data]** has a positive influence on the **safety, security and performance of the modified product** (= *positive effect on **[Modification]***), and vice versa.

14 ↓

Training data that covers a broader range of use-contexts (= *strong evaluation in **[Training Data]**, particularly ¶2 and ¶3*) reduces the likelihood that the AI system is (negligently, or consciously) used by operators **in contexts that it was *not* trained for** (= *positive effect on **[Out-of-Scope Use]***).

[DEV]

Engagement with Developer



[DEFINITION]

Identifies the extent to which the user is **involved in the development of the AI system**, and has access to the relevant information and expertise from its **developers**.

Concept

1 Military AI systems may be developed wholly or partly by external organisations, private companies. Particularly in such cases, effective engagement between the user and the developer can provide opportunities for **user input, oversight, access to relevant technical information, and continuing technical support**. It can help users develop a more accurate understanding of the system's capabilities, limitations, performance, and risks, and ensure that operational requirements and constraints are appropriately reflected during development and testing.

2 **Higher levels of engagement** may be characterised by:

- A close working relationship between the user and developer.
- Appropriate transparency concerning the system, including relevant data, data-management practices, model characteristics, testing, and performance.
- Opportunities for users to provide input during development.
- User involvement during test and evaluation.
- Access to relevant developer expertise before and during operational use, including where technical advice or modifications are required (*see [Technical Knowledge] and [Modification]*).

3 Engagement during development can **improve the user's ability to understand and assess the system**.

User involvement may provide greater insight into the data, models, design choices, assumptions, and testing results underlying the system. This can make it easier to identify relevant limitations and evaluate variables such as **[False Positive Rate]**, **[Target Proxy]**, **[Bias]**, **[System Safety]**, **[System Security]**, **[Training Data]**, and **[Repurposing]**.

Engagement can also allow operational requirements and user needs to inform system design. For example, prospective operators may participate in testing and provide feedback on interface design, explanations, alerts, or intervention mechanisms (*see [Contextual Awareness], [Understandability], and [Intervention Ability]*).

4 **Different relationships** between users and developers provide different opportunities for engagement.

Generally, co-development provides greater opportunities for user involvement and access throughout development. In contrast, commercial off-the-shelf (COTS) acquisition may provide more limited insight into, or ability to influence, the development process and final product.

A 'layered' relationship is relatively common, where some components of a system remain outside of the user's control, while other components are developed, configured, or adapted for the user.

5 The **appropriate level of engagement and transparency depends on** the system, its proposed use, and the risks involved. At minimum, users should have sufficient access to information and expertise to make informed assessments of relevant performance, safety, security, and operational limitations.

Full access to every aspect of the development process may not be feasible or necessary, particularly where external developers are involved. Relevant information may instead be provided through documentation of training data, testing and validation results, performance metrics, system specifications, information about known limitations and vulnerabilities, or access to developer expertise.

The information required will depend on the particular system and the variables being assessed (*see, e.g.*, *[False Positive Rate]*, *[Target Proxy]*, *[Bias]*, *[System Safety]* and *[System Security]*).

How to evaluate?

Overall Access & Transparency	Comprehensive	Extensive	Partial	Minimal	None
AND					
Type of Relationship	In-house	External (close cooperation)	External (periodic engagement)	External (rare engagement)	Commercial off-the-shelf
AND					
Developer Support	Readily available	Generally available	Available with limitations	Minimal	None

[Overall Access & Transparency] The extent to which the user has access to the developer and insight into the development process (see ¶2 and ¶5).

[Type of Relationship] The type of relationship between the user and developer (see ¶4).

[Developer Support] The extent to which the user has access to developer support, e.g. for technical advice or assistance with modifications (see ¶1 and ¶2).

Legal relevance

D C DISTINCTION/ CIVILIAN HARM	Greater engagement with the developer may help users identify and address limitations in the system's performance, design, training data, or testing that could otherwise contribute to errors. This may reduce the risk of protected persons or objects being misidentified as lawful targets (undermining distinction), or of civilians or civilian objects being overlooked (potentially distorting collateral damage estimation and undermining proportionality assessments).
P PRECAUTIONS	Greater engagement with the developer can provide users with information and technical expertise needed to understand the system's capabilities, limitations, behaviour, and expected effects. This can support informed decisions about whether and how the system may be used, what safeguards or restrictions are necessary, and what feasible precautions should be taken in the circumstances.

Relationship Chart

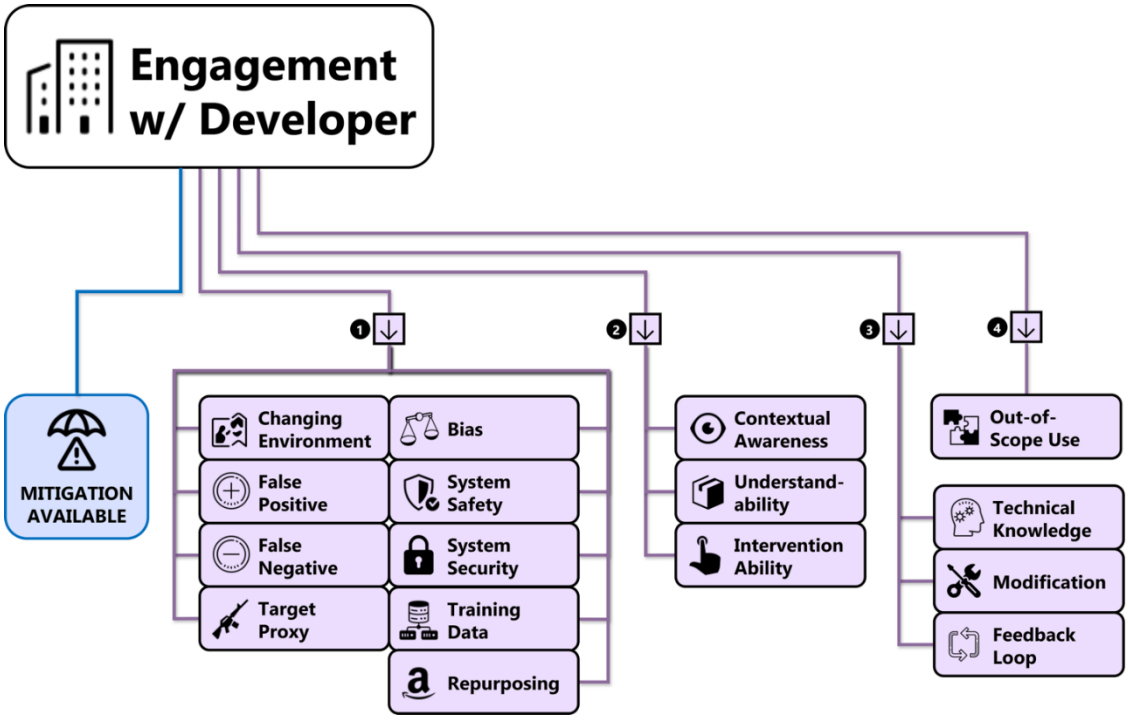


Chart Legend

AFTER EVALUATING, CONSIDER |

[MITIGATION]

Where [Engagement with Developer] indicates elevated risk, consider:

- **Increase access and transparency:** Seek additional documentation, testing and validation results, information on relevant data and design choices, known limitations and vulnerabilities, or other information necessary to assess the system and its risks.
- **Increase user involvement:** Where feasible, involve users in testing, evaluation, and subsequent development or adaptation of the system, and establish mechanisms for providing feedback to the developer.
- **Improve access to developer support:** Establish arrangements for users to obtain relevant technical expertise from the developer when needed, including during time-sensitive situations where appropriate.

1 ↓

A strong evaluation in [Engagement with Developer] provides important insights into the development and testing processes. This information **facilitates the evaluation of** [Changing Environment], [False Positive Rate], [False Negative Rate], [Target Proxy], [Bias], [System Safety], [System Security], [Training Data] and [Repurposing].

2 ↓

A strong evaluation in [Engagement with Developer] provides more opportunities for operators to provide feedback early during the AI system's design process. This, in turn, allows for more **user-attuned dashboards, interfaces and workflows** (= *positive effect on [Contextual Awareness], [Understandability] and [Intervention Ability]*).

3 ↓

A strong evaluation in **[Engagement with Developer]** facilitates the **embedding of forward-deployed engineers or the training of operators**. Such support is valuable for:

- Improving technical know-how overall, e.g. through the provision of training courses or technical instruction (= *positive effect on **[Technical Knowledge]***)
- Supporting post-incident analyses (= *positive effect on **[Feedback Loop]***).
- Ensuring that modifications are performed responsibly and by competent staff (= *positive effect on **[Modification]***).

4 ↓

A strong evaluation in **[Engagement with Developer]** may provide greater **clarity on which circumstances constitute out-of-scope use**, thereby allowing commanders to better define and limit in which use-contexts the AI system may be deployed (= *positive effect on **[Out-of-Scope Use]***).

[Repurposing]
Repurposing

[DEFINITION]
Identifies risks arising from **repurposing AI models that were developed for non-military or materially different contexts** for use in military applications.

[COMPARE]
This variable concerns the adaptation or repurposing of an existing AI model for a military application during development. For modifications made *after* adoption, see [Modification]. For use *outside the context* for which an adopted system was designed and validated, see [Out-of-Scope Use].

Concept

1 Military AI systems may be developed specifically for military purposes, or **may be repurposed from models originally developed for other applications or contexts**.

Repurposing an existing model may provide a **faster and more resource-efficient means** of acquiring a particular capability, compared to developing a model specifically for the intended military use-case. However, this also introduces risks (*see ¶3*).

The risks associated with repurposing depend principally on how different the model's original purpose and development context are from its intended military application, and how adequately it is adapted and validated for that new use.

[FORMULA]
Repurposing Risk = Degree of Mismatch × Adaptation & Validation Risk

[EXAMPLE]
Battalion A needs to rapidly integrate computer vision capabilities into its drones to assist in identifying enemy air defence systems. Battalion A has the choice between developing a classifier specifically for this purpose using available military data, or adapting an existing general-purpose computer-vision model. Due to time and resource constraints, it opts for the latter, despite limitations in its performance for the intended military task.

Certain AI systems are almost **never made entirely in-house**. These include 'foundation models' such as GPT, which are trained on the whole corpus of the internet and are almost never tailor-made for one particular domain.

2 AI models not originally developed for the intended military application may be **adapted or augmented to improve their suitability for the new use context.** Depending on the model and application, this may include finetuning, additional training, or other forms of adaptation.

[EXAMPLE]
Military B is developing an AI legal assistant for use by its legal advisers. Rather than developing a LLM from scratch, it takes an existing general-purpose model and adapts it for

the intended military application by finetuning it using relevant military legal materials, operational scenarios, doctrine, and policy.

Using an AI model for a materially different purpose or context without adequately adapting and validating it may give rise to **[Out-of-Scope Use]** concerns.

3 Repurposed AI models may retain **characteristics, assumptions, limitations, or behaviours originating from their original (non-military) use-context that are unsuitable** for the intended military application.

Depending on the model and use case, these may include unsuitable performance characteristics, objectives or behaviours, safety or security vulnerabilities, or biases (*see **[System Safety]**, **[System Security]**, and **[Bias]***). The system may also incorporate workflows, interfaces, or interaction patterns that do not adequately reflect the needs of military users or the operational environment (*see **[Contextual Awareness]***).

The extent of this risk depends in part on **how different the original development context is** from the intended military application and **how effectively the model has been adapted and validated** for its new purpose.

[EXAMPLE]

Military C uses an AI system to assist cyber operators in identifying vulnerabilities in its network. The underlying model was originally developed for commercial cybersecurity applications. As a result, the AI is attuned to finding exploits that are usually targeted by scammers and cybercriminals, instead of vulnerabilities preferred by military cyber brigades and advanced persistent threats.

How to evaluate?

Degree of Mismatch	Tailored for military	Limited mismatch	Moderate mismatch	Significant mismatch	Fundamental mismatch
AND					
Adequacy of Adaptation	Very comprehensive	Comprehensive	Adequate	Minimal	None

[Degree of Mismatch] How different the model's original purpose and development context is, compared to the intended military application (see ¶1).

[Adequacy of Adaptation] The extent to which the model has been adapted and validated for its intended military application (see ¶2).

Legal relevance

D C
DISTINCTION/
CIVILIAN HARM

Where a model developed for a different purpose or context is insufficiently adapted and validated for its intended military application, resulting mismatches may increase the risk of errors. These may include the misidentification of protected persons or objects as lawful targets (undermining distinction), or failures to detect civilians or civilian objects (potentially distorting collateral damage estimation and undermining proportionality assessments).

P PRECAUTIONS	Where a model has been repurposed for a materially different context without adequate adaptation and validation, its behaviour, performance, and effects may be more difficult to anticipate. This may undermine the ability of those planning or executing attacks to assess relevant risks and identify and take feasible precautions to avoid or minimise civilian harm.

Relationship Chart

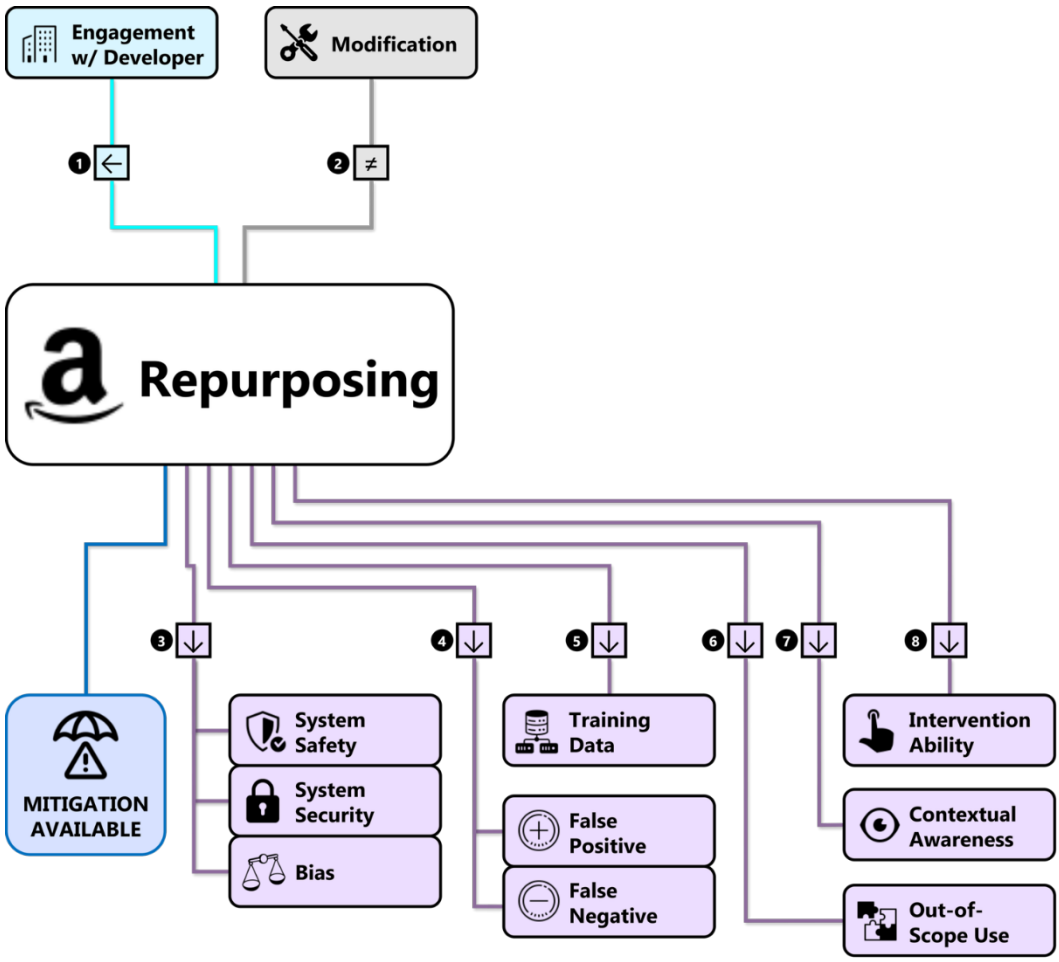


Chart Legend

WHEN EVALUATING, CONSIDER |

1 ← Greater involvement in the development process (= *strong evaluation in [Engagement with Developer]*) may offer more opportunities to **develop a tailored-to-military system**, instead of needing to rely on general-purpose or minimally finetuned models (= *positive effect on [Repurposing]*).

2 ≠ **Do not confuse** [Repurposing], which looks at changes made to the AI model *during development*, with changes made *after adoption* (often in response to changing circumstances): see [Modification].

[MITIGATION]

Where [Repurposing] indicates elevated risk, consider the following:

- **Reduce context mismatch**: Where feasible, select or develop models with characteristics, data, and capabilities suited to the intended military application, reducing the degree of repurposing required (*see [Training Data]*).
- **Rigorous adaptation**: Where an existing model is repurposed, adapt it appropriately for the intended military application. Particular attention should be paid to assumptions, patterns, biases, or behaviours originating from its original development context that may not transfer appropriately to the new use.
- **Rigorous testing and validation**: Test and validate the repurposed model under conditions representative of its intended military application to identify performance limitations, undesirable behaviours, or vulnerabilities.

3 ↓ Repurposed models (= *poor evaluation in [Repurposing]*) may **carry over robustness problems, security vulnerabilities, and undesirable biases** from their original version (= *negative effect on [System Safety], [System Security] and [Bias]*).

4 ↓ For repurposed models (= *poor evaluation in [Repurposing]*), their **performance may deteriorate if the model is used in a military context** (= *possible negative effect on [False Positive Rate] and [False Negative Rate]*). If the system's performance has previously only been measured in a general-purpose or civil context, re-testing in the new military use-context is advised.

5 ↓ For repurposed models (= *poor evaluation in [Repurposing]*), there is a high likelihood that the **data that was used during training is overly generic or misaligned** (= *negative effect on [Training Data]*), as the data was not originally collected and tailored for military contexts.

6 ↓ For repurposed models (= *poor evaluation in [Repurposing]*), there is a higher likelihood that the military tasks/situations for which the system will be deployed **fall outside of the system's intended use-context** (= *negative effect on [Out-of-Scope Use]*).

7 ↓ The **way information is presented** in civil or general-purpose AI applications (= *poor evaluation in [Repurposing]*) may not be well-suited for military contexts (= *possible negative effect on [Contextual Awareness]*).

8 ↓ **Methods/channels of intervention** used for civil or general-purpose AI applications (= *poor evaluation in [Repurposing]*) may not be well-suited for military contexts (= *possible negative effect on [Intervention Ability]*).