EC2 N18

Methodological Guideline for AI-based System Design at Operational and System Level:

Operational Approach – 218A

L2.3.2.3

L2.4.3.3

L2.4.4.3

L2.4.8.3

contact@confiance.ai | www.confiance.ai

Document reference: 218A

Contributors

	Name	Organisation	Role
Responsible for the deliverable	Kevin Mantissa	IRT SystemX	Research Engineer
Scientific responsible	Christophe Bohn	IRT SystemX	Technical Coordinator
Co-authors	Christophe Bohn	IRT SystemX	Technical Coordinator

Document control

Revision	Date	Commentary	Author
0.1	12/10/2023	Creation	Kevin Mantissa
0.2	01/12/2023	Update for review	Kevin Mantissa
0.3	15/12/2023	Update following comments	Kevin Mantissa
0.4	17/01/2024	Update following comments for final delivery	Kevin Mantissa, Christophe Bohn
1.0	24/01/2024	Preparation for final delivery	Kevin Mantissa, Christophe Bohn


Confiance
ai

Table of Contents

A. Introduction and abstract

A.1 General introduction to trustworthy AI challenges

Trustworthiness in Artificial Intelligence (AI) within critical systems (systems that can directly or indirectly affect human life and moral entities) is essential for its widespread adoption (by the industry, the decision makers, the general public, etc.) and poses the following significant challenges.

- First, how to design AI models, so that, by construction, they satisfy trustworthy properties (accuracy, robustness…).
- Secondly, how to characterize these AI models, for example to understand and explain their behavior and their adequacy to the operational domain.
- Then, how to implement and embed those AI models on hardware, by making them fit for the target without losing their trustworthy properties.
- Another question is, what methods of data engineering to apply in order to, among other topics, manage important volumes of data and adapt to the evolution of the operational domain.
- At system level, what verification and certification processes to consider specifically for AI-based systems.
- Finally, a federation of all these matters is necessary to build an end-to-end methodological approach, supported by a consistent engineering environment compatible with industrial practices.

These are the challenges, among others, that the Confiance.ai program addresses.

A.2 Context of the methodology

The quick progress of AI systems has highlighted several challenges that the community is currently facing. Among those challenges, some of them illustrate the limits of current System Engineering methods and best practices, that are fit for Conventional Systems, but are not able to keep up with the rapid evolution of AI-based systems and the peculiarities of such systems.

Working Groups for Standardization and Engineering Systems aim to resolve the challenges imposed by this quick evolution, by updating current System Engineering standards - e.g. ISO 15288:2023 (ISO/IEC/IEEE 15288, 2023) - or by releasing new ones tailored for AI – e.g. ISO 5338 (ISO/IEC DIS 5338, 2023). Standards specific to industrial domains are also being developed, such as the current work-in-progress standard for aeronautical safety domains implementing AI: SAE ARP6983/EUROCAE ED-324 (*SAE ARP 6983*, 2023).

However, what is needed to better apprehend AI from a System Engineering viewpoint is to revisit the operational and system activities formerly applied for Conventional Systems, and to evaluate how Artificial Intelligence impacts those activities.

While major tech actors of the AI field try to incorporate some System Engineering framework into their AI Engineering, the approach described in this deliverable is at the other end. We aim to incorporate the new methods associated with AI development into the conservative System Engineering methods. In this manner, we can assess the impacts of AI on conventional methods, as well as precise the boundaries of AI engineering that are not as well-defined as other fields, due to its novelty.

In Confiance.ai context, this work on System Engineering of AI-based systems is the way to increase trustworthiness by applying on AI mainstream industrial processes. This allows to make AI more compliant with Quality Assurance (including validation) and Safety Assurance.

This work aims to be integrated with EC2.14 with Confiance.ai end-to-end approach in future works.

A.3 Purpose of the methodology

A.3.1 A two-step approach

The work presented in this document is built upon the notions of Intended Purpose and Design Intent. According to the Artificial Intelligence Act, first drafted in 2021 (AI Act Draft 2021, 2021) and updated pending validation in 2023 (AI Act Draft 2023, 2023), Intended Purpose means *"the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation".*

Intended Purpose is a part of the Operational specification of an AI-based system. It could be considered as a global introduction before the analysis of stakeholders needs and lifecycle phases.

Design Intent describes design activities to address the needs synthesis at system level and translate it in a technical way in order to make it usable without ambiguity for ML and data engineering.

Both notions of Intended Purpose and Design Intent are part of architecting activities. Architecting (*ISO/IEC/IEEE 42020:2019*, 2019) includes the activities of *"conceiving, defining, expressing, documenting, communicating, certifying proper implementation of, maintaining and improving an architecture throughout the life cycle for an architecture entity".*

The aim of the following work is to guarantee the traceability of design and architectural choices at operational and system levels for AI-based systems. We focus on AI-based systems that rely on machine learning.

To address system engineering of AI-based systems, we split the methodology into two approaches:

- An Operational Approach, that aims to identify and characterize operational needs which can only be managed by an AI-based system (could not be reached by a conventional software)
- A System Approach, that aims to gather system-level artifacts that are required for the AI-component implementation and the coverage of operational needs

To simplify the following work, assumptions are made and explained in section A.10 below.

This document will focus on and detail the Operational Approach only. For the System Approach, please refer to the corresponding deliverable in 218B (Mantissa & Bohn, 2024). We strongly advise the readers to finish the Operational Approach document before reading the System Approach deliverable. Indeed, both deliverables have to be considered as a whole.

A.3.2 Rationale for the Operational Approach

The Operational Approach gathers all activities related to Operational design required for AI-based System Specification.

The aim of Operational Design is to identify the stakeholders and the operational phases of system lifecycle, in order to perform a synthesis of stakeholders needs, for each operational phase.

In this context, the Intended Purpose as required by the AI Act for AI-based systems should be part of the Operational Specification.

The incentives that lead to develop AI-based systems are found at the operational level:

- High dimension of inputs to consider,
- Difficulty to specify the intended behavior for each combination of inputs,
- Need for the system to take autonomous decisions instead of a human operator,
- Ability of the system to treat situations never encountered before.

We can notice that in the typology of systems above, the conventional software-based systems reach their limits in terms of performance.

Thus, for AI-based systems, we encounter the following consequences for operational design:

> The complex nature of the problem to solve makes it impossible to specify each individual case that the system will encounter with a requirement.
>
> To answer the need, there is no bijection between the inputs and the outputs: multiple solutions can exist for a single situation. It results in a specific difficulty to describe and specify operational design.
>
> Autonomous levels lead to change the responsibilities boundaries between human operators and systems: the actions in human responsibility which are not described in conventional operational specifications need to be described when it becomes system's responsibility (with performance description). This leads to an increase in operational design complexity.
>
> The ability of AI-based systems to treat unspecified situations (generalization properties) is difficult to describe in an operational specification.

The possibilities offered at operational level by AI-based systems combined with the use of these innovating technologies impose to consider system limitations and specify clearly the trade-offs between stakeholders needs and technical constraints in the Operational Specification. The definition of the trade-off as a design choice implies all engineering teams. Once it is defined, each engineering team has to identify implication for operational part, system part and AI implementation part.

A.4 Confiance.ai scientific challenges addressed by the document

The deliverable addresses the following scientific challenges of Confiance.ai:

- Establish a methodology for defining the desired behavior of the trusted AI system

Besides, this deliverable also addresses questions inspired from deadlocks identified in EC2 project:

- **VR05** – How to design AI-based systems?

- **VR04** – How to properly describe operational specification regarding AI components issues?
- **VR11** – How to design AI-based systems that are User-Experienced oriented (human factor)?
- **VR19** – How to ensure the link with Functional Safety? (Quality Assurance part: we indicate the activities to perform, then we verify that those expected activities have been indeed performed)

A.5 Target audience

This deliverable (218A) and the deliverable dedicated to the System Approach (218B) target all types of profiles of the Confiance.ai program. It includes profiles with a System Engineering background. Indeed, with these approaches, the aim is to study how to appropriately incorporate AI in Conventional System Engineering methods and industrial context. Besides, it includes data and AI engineers as well, in order to give them a better understanding of what Operational and System Approaches could bring to AI implementation. This will empower them to absorb a global system view which encompasses the data viewpoint and the learning viewpoint. This approach supports the ability of the results to match with the needs, which is a condition for trustworthiness.

A.6 Glossary

Term	Signification
AEB	Advanced Emergency Braking
AI	Artificial Intelligence
ATO	Automatic Train Operation
ATP	Automatic Train Protection
CI/CD	Continuous Integration / Continuous Delivery
CONOPS	Concept of Operations
DDT	Dynamic Driving Task
EASA	European Authority for Aviation Safety
EE	Electrical Electronic
EU	European Union
GDPR	General Data Protection Regulation
GMG	Goal Mis Generalization
GoA	Grade of Automation
GPT	Generative Pre-trained Transformer
HAX	Human AI eXperience
HMI	Human Machine Interface
HW	Hardware
IEC	International Electrotechnical Commission

Term	Signification
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
MDR	Medical Device Regulation
ML	Machine Learning
NLP	Natural Language Processing (NLP)
ODD	Operational Design Domain
OEDR	Object and Event Detection and Response
PoC	Proof of Concept
QCDP	Quality Cost Delivery People
SAE	Society of Automotive Engineer
SDD	Software Design Description/Document
SEBoK	System Engineering Book of Knowledge
SOTIF	Safety of The Intended Functionality
SW	Software

A.7 Summary of limitations and perspectives

The efforts put in this document are focused in the formalization of encountered issues regarding AI-based system Operational Approach with leads to solve them as a methodological guideline, instead of providing applicable solutions. In consequence, this work could serve as a basis for future works, such as:

- Multidisciplinary collaborative works on each part of this document with various engineer profiles, like system engineering, data engineering, machine learning engineering, quality assurance, etc.
- Proposals of an engineering process for Operational Approach detailing activities and tools;
- Application of the detailed approach on an industrial use case.

A.8 Document organization

To address the questions about Operational Approach raised in section A.3 above, we will divide this work in the six following topics, which will be detailed further in the body of the document:

- Part 1: Characterization of the intended purpose
- Part 2: Operational Design from reference systems
- Part 3: Bottom-up approach – trade-offs between technical limitations and opportunities impacts at operational design level
- Part 4: Commonality of operational methodology for System Engineering and AI Systems
- Part 5: Focus on responsibility boundaries between system and users

- Part 6: Specific deadlocks for operational design specification of AI-based systems

These topics constitute a basis to ensure the consistency of the Operational Specification for an AI-based system. They do not aim to form a detailed engineering process yet.

A.9 How to use the document

This document is a guideline rather than a proper process (that could result from future works).

In the Operational Approach, we give key elements that should be considered to realize an operational specification in order to identify the synthesis of the needs in preparation for its implementation in a given AI-based system.

Between the synthesis of needs and the implementation, the System Approach detailed in 218B (Mantissa & Bohn, 2024) will achieve the translation of operational needs into technical elements in the System Specification (more details in Part 2 of deliverable 218B).

That is why the whole scope of system activities for AI-based systems engineering is covered by both deliverables.

A.10 Assumptions regarding this deliverable

In order to clarify the methodology explanation, we make the following assumptions but the resulting method can be generalized outside of these hypotheses:

- We consider an industrial project with clear milestones and clear goals defined by the product team:
  - The detailed activities are related to industrial project objectives with QCDP (Quality, Cost, Delivery, People) engagement,
  - The considered process assumes that the operational objectives can be achieved thanks to preliminary studies results,
  - Resulting industrial process aims to guarantee the achievement of the project, in accordance with its objectives

  In consequence, no preliminary work and no specific iteration is considered in our approach.
- The studied activities focus on design and architectural aspects: architectural choices and renunciation/opportunities (trade-offs).
  Detailed specifications, data specifications, datasets and testing activities should be considered out of scope of this activity.
- In accordance with Confiance.ai scope, the studied system is a critical industrial system. This kind of system requires a clear development process to grant quality assurance and safety.
- We focus on subsystems which can't be implemented by a conventional software development, and require AI-based implementation, as described in section A.3.2.

B. Part 1 – Characterization of intended purpose

Introduction to part 1

The Intended Purpose of AI-based systems is a notion introduced by the AI Act in 2021. It refers to "*the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation".*

Before delving deeper into how the Intended Purpose is key in order to specify AI-based systems at operational level, we will first use the section B.1 to explore where the Intended Purpose originates from, and what lessons could be learned and applied on AI.

It shall be noted that the notion of **Intended Use** can be used instead of Intended Purpose.

Overview of part 1

This section is organized in the following manner:

- We first introduce the Intended Purpose and its origins from the medical field before focusing on its use in major European regulation for AI;
- Then, we give some examples of Intended Purposes for existing or work-in-progress AI-based systems.
- Finally, we discuss the added value and expectations of the Intended Purpose within the Operational Specification of AI-based System, and warn on pitfalls to avoid during design.

B.1 Origin of Intended Purpose

B.1.1 EU MDR

B.1.1.1 Definition of Intended Purpose in the medical field

The European Union Medical Device Regulation (EU MDR 2017, 2017) is a regulation instated by the European Union in 2017 for the medical domain. It mainly refers to condition of use of devices (usually reliant on software components) in a domain that is heavily regulated and monitored.

It shall be noted that the EU MDR is the first European regulation to have proposed a definition for the Intended Purpose, regarding medical devices.

The definition for the medical domain is the following: *"the use for which a device is intended according to the data supplied by the manufacturer on the label, in the instructions for use or in promotional or sales materials or statements and as specified by the manufacturer in the clinical evaluation"***.**

Thus, this key notion heavily inspired the version for the AI Act while adding some specificities. The definition of Intended Purpose from EU MDR is more specific to the medical domain, but it could still

incorporate devices relying on AI. While in the meantime, the definition of Intended Purpose from AI Act is supposed to be applicable to all relevant domains.

Unlike in health applications, not all AI applications are safety-critical. However, even for non-critical applications, it is essential to have a proper understanding of what the system can do, from a user perspective, in order to avoid misuses and misunderstanding that could prove fatal in future critical implementations of AI.

Besides, the reuse of Intended Purpose from EU MDR could also benefit the domain, as AI is being used in more and more medical devices, in replacement of conventional software components.

It should be noted that while the Intended Purpose is a key notion for AI and is heavily inspired from an EU medical regulation first drafted in 2017, search results in academic and industrial literatures mainly redirect to the Intended Purpose for medical activities. This shows that the notion applied to AI is still rather new and has yet to be explored further.

B.1.1.2 Added value of Intended Purpose

The EU MDR suggests that a proper Intended Purpose helps to provide the following elements on medical devices:

- Whether the product being considered fits the definition of a “medical device” and therefore whether or not the regulation applies;
- the basis for the classification of the future planned device into one of the four [risk] classes of device, as required by Article 51 of EU MDR;
- A core text which is needed for the future labelling, instructions, promotional or sales materials, the clinical evaluation and the technical documentation.

To put it simply, the Intended Purpose clarifies the position of a given device regarding applicable regulation, serves as input for hazard analysis and classification and is necessary to support the quality assurance process related to the device.

B.1.1.3 Responsible of the Intended Purpose

In terms of who has the responsibility of defining the intended purpose of a device, the EU MDR insists that it shall be defined by *“medical professional, ideally someone with experience of medical writing”*, destined to *“the intended user group, medical professional or patient, using appropriate medical language”.*

Usually, the manufacturer of a medical device is expected to be responsible for devising the Intended Purpose of said device. This is important as it will play a role in defining whether or not the product must be qualified as a medical device.

As for AI-based systems, the main responsible of the Intended Purpose is supposed to be the system provider, as indicated by the AI Act.

The Intended Purpose should serve as a support to classify the AI-based system as a critical application or not. Consequently, additional constraints shall be applied on the provider who defines the Intended Purpose, similarly to the medical field.

B.1.1.4 Content of an Intended Purpose in the medical field

Although the EU MDR as a European regulation is not expected to deliver a complete and thorough vision on the manner to define an Intended Purpose, the regulation still provides some key elements on what should be present in it. This could serve as an inspiration for Intended Purpose in AI applications.

First, the Intended Purpose is expected to be concise, no longer than "*two or three sentences" and should "focuses on what the device is intended to be used for".*

Secondly, the Intended Purpose is expected to give value to the Technical Documentation it will be incorporated in. Besides, it should also remain consistent with two other items of this Technical documentation:

- "*the intended patient population and medical conditions to be diagnosed, treated and/or monitored and other considerations such as patient selection criteria, indications, contra-indications, warnings"*;
- *"the principles of operation of the device and its mode of action, scientifically demonstrated if necessary"*.

B.1.1.5 Expected properties of the Intended Purpose in the medical field

The Intended Purpose bears several properties that ensure that it is properly defined, both in its short statement and more extensive description:

- Precise: It should explicitly mention any limitations and be as precise as possible in order to prevent edge cases or misclassification in the wrong risk class
- Defined early: It should be defined as early as possible in the development process, as it is the basis of key design and qualification decisions.
- Clear: It is expected to be clear to prevent misunderstanding from end users that could lead to potential risks or misuses.

B.1.2 SaMD UK Government Guideline (compliance with UK MDR 2002)

B.1.2.1 Context

In March 2023, the UK Government released a guideline dedicated to Software as a Medical Device (SaMD): (*Guidance: Crafting an Intended Purpose in the Context of SaMD*, 2023). It aims to provide a guideline regarding the definition and use of the Intended Purpose for SaMD, in addition to what is proposed by the UK Medical Devices Regulations (UK MDR 2002, 2002), and with the stronger presence of AI-based systems in the medical field. The UK MDR is the national equivalent to the EU MDR.

It provides some examples that are AI-oriented (which leaves the door open for its future extension to AI-based systems).

It defines what an intended purpose must contain, its criticality in an application risk level, and feedbacks and known pitfalls to avoid when defining intended purpose.

A workflow is also proposed to sum up the main steps for building an initial Intended Purpose for SaMD:

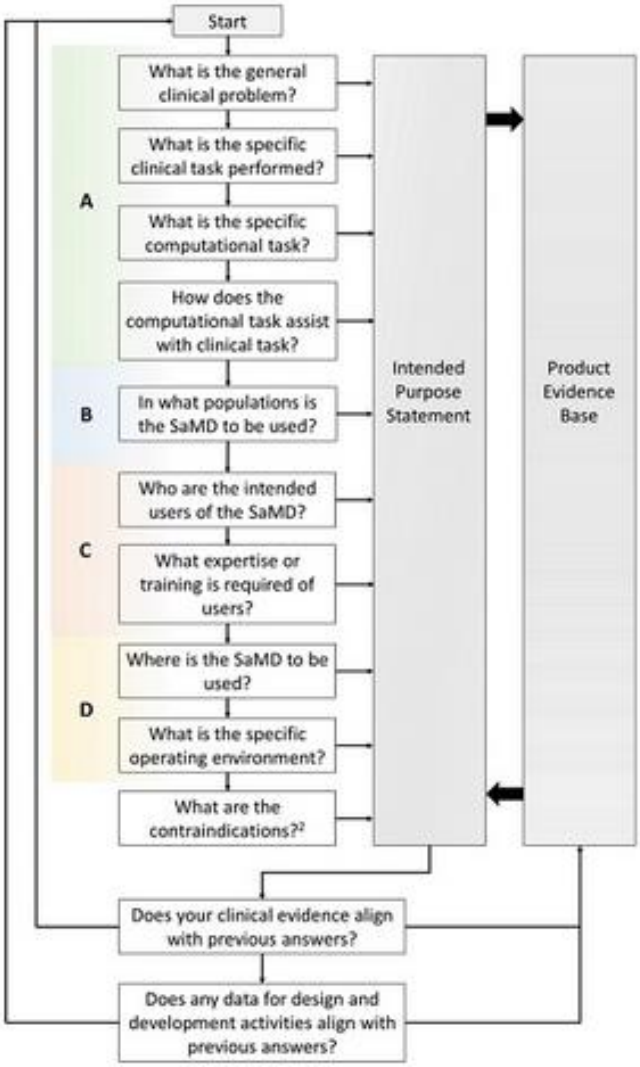


Figure 1 Workflow for Intended Purpose in SAMD Guideline

Based on this workflow and the whole guideline, we produced a mind map in annex I.1 below. This mind map sums up every wanted characteristics, related notions and expectations for the Intended Purpose for SaMD that could be relevant for AI-based systems as a whole.

B.1.3 AI Act

B.1.3.1 The Intended Purpose in the AI Act

The AI Act is a tentative EU regulation currently under discussion and for which a first draft has been published in 2021. The latest version available was released during 2023 (AI Act Draft 2023, 2023).

It aims to provide a framework to the design, commercialization and use of AI-based systems in the European markets or impacting European citizens. AI-based systems are defined in the context of the AI Act as *"a machine-based system that is designed to operate with varying levels of autonomy and that can,*

for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions, that influence physical or virtual environments" (AI Act Draft 2023, 2023, art. 3 §1 point 1).

It is fundamental as it proposes tools and methods to ensure that AI-based systems will be compliant with European regulation. It is the main resource for discussing about Intended Purpose for AI-based systems currently.

The current available draft of AI Act (2023) sets a particular focus on the Intended Purpose (more than forty occurrences), which shows the importance supported by this notion.

The following view shows how the Intended Purpose is incorporated within a corpus of expected documents and information by the AI Act.

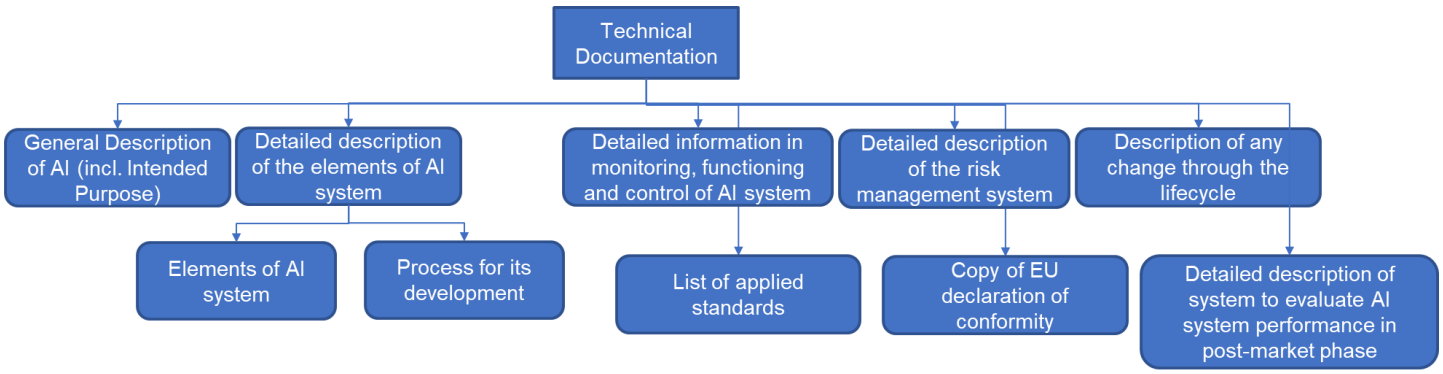


Figure 2 Expected structure of Technical Documentation in AI Act, including Intended Purpose

B.1.3.2 Notable mentions of the Intended Purpose in AI Act

The AI Act expects the Intended Purpose to be part of a technical documentation that will be submitted when considering commercialization of AI-based Systems to be deployed on European markets

We list below some key articles and mentions of the Intended Purpose that appear in the AI Act:

- **§5.2.3 High Risk AI Systems:** Intended Purpose of an AI-based system is used for its classification as high-risk (or not), combined with the type of function performed by the system
- **Requirement 66:** Whenever the Intended purpose of the system changes, a new conformity assessment will be required
- **Article 7**: Intended Purpose is a mandatory element to examine
- **Article 8:** Verification of Compliance with AI Act should rely on analysis of the Intended Purpose and the risk management system required to set up for the system
- **Article 9:** "*Testing procedures shall be suitable to achieve the intended purpose of the AI system.*"
- **Article 13:** Intended Purpose of an AI-based system to support the understanding in the description of the performance limits and the characteristics of the system and to provide transparency
- **Annex IV:** Intended Purpose is part of the Technical Documentation and is among the first expected items. The Technical documentation also includes a general description and detailed description of the AI-based system, process, and other sections.
- **Annex VIII:** When an AI-based system is registered as high-risk, the Intended Purpose is expected in the registration documentation for commercialization on the EU market.

B.1.3.3 Current limitations of AI Act regarding Intended Purpose

Although the AI Act is expected to be a key resource to drive the use of Intended Purpose when devising AI-based systems. It is currently met with some limitations:

- The AI Act does not suggest method to build Intended Purpose,
- The AI Act Intended Purpose definition relies very heavily on EU MDR regulation for medical devices. Thus, adaptation may be needed compared to its original context.

It is expected that the notion of Intended Purpose for AI will gain in popularity as the final version of the AI Act gets closer.

B.2 Examples of Intended Purpose

Due to the novelty of considering the notion of Intended Purpose for AI-based systems, it shall be noted that this notion is not yet applied massively despite the rapid progression of AI-based system. To illustrate this paradox, let's provide some examples.

B.2.1 ChatGPT

"Intended Purpose" of ChatGPT by OpenAI

ChatGPT is a generative AI model, by OpenAI, that started a rapid progression in the domain at the end of 2022, due to its quick deployment to non-experts end users and the generalization of its uses in daily activities, either professional and personal. It relies on Generative Pre-Trained Transformers, and can serve as an AI-based chatbot, as well as other features with each major update since 2022.

On its website, OpenAI (founder of ChatGPT) presents its generative AI model in the following manner:

- ChatGPT is presented as *"a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response."* (Introducing ChatGPT, 2022)
- "*OpenAI's GPT (generative pre-trained transformer) models have been trained to understand natural language and code. GPTs provide text outputs in response to their inputs. The inputs to GPTs are also referred to as "prompts". Designing a prompt is essentially how you "program" a GPT model, usually by providing instructions or some examples of how to successfully complete a task. GPTs can be used across a great variety of tasks including content or code generation, summarization, conversation, creative writing, and more"* (OpenAI Platform, n.d.).

These 2 definitions seem to be the tentative proposal of Intended Purpose for ChatGPT by OpenAI. The sourced pages also present direct examples of what ChatGPT can do, in comparison with its sibling model InstructGPT.

This page also evokes the limitations of ChatGPT (*"plausible sounding but incorrect answers",* etc.;).

How ChatGPT presents its limitations to the end users

OpenAI are aware that ChatGPT is prone to fail, so they suggest a certain number of best practices to ensure that the user gets the best results possible. Among those strategies, they suggest:

- Write clear instructions
- Provide reference text
- Split complex tasks into simpler subtasks
- Give GPTs time to “think”
- User external tools
- Systematic changes testing

OpenAI uses many examples to illustrate the capabilities and limits of its ChatGPT.

Discussion on the “Intended Purpose” of ChatGPT

A closer review of the proposed Intended Purpose for ChatGPT can raise several criticisms:

- The end users are not clearly informed on what ChatGPT can do beyond examples. Although they are welcome by a list of examples, capabilities and limitations on the main page, this list is not exhaustive and it relies on the supposed user awareness of what an AI conversational agent is and what it can globally do;
- Detailed documentation exists but the user has to look for it;
- Currently, users are likely to hear of ChatGPT through medias and word of mouth, which may give them a wrong mental model of what the system can and cannot do;
- The web page introducing ChatGPT is very succinct regarding its Intended Purpose. And it does so by comparing it to another system of OpenAI, that non-expert users are unlikely to know of.

The example of ChatGPT is interesting as it illustrates an AI-based system that became massively popular, but at the cost of a proper end users understanding. A better defined Intended Purpose, massively communicated, could have prevented misuses and misunderstandings using ChatGPT tool.

B.2.2 Welding Use Case Proposal of Intended Purpose

In the Batch 1 deliverable for Welding Use Case (Confiance.AI EC6, 2021) , §1.1.2, a first description of the use of the Welding AI Visual Inspection system that could be considered as an Intended Purpose is the following one:

> “*The AI-based system aims to assist a human operator responsible for the quality control of welding parts classification. It should enable the automatic validation of most conform welding parts, in order to display to the human operator only the parts where classification needs specific human expertise. The human operator is the only one that is able to reject a welding part assessed as non-conform.”*

This first proposal was expressed without the context of the Intended Purpose expected from the AI Act. This proposal is revisited in the context of this deliverable to suggest a more formal Intended Purpose for the Welding Use Case (but is not converged with the Use Case Owner):

> *“The AI Visual Inspection system must automatically validate the welding parts for which it is confident in the welding quality, while leaving the responsibility to a human operator to validate welding parts where the AI system cannot perform a classification or welding parts that are not conform”.*

This example illustrates how different interpretations of the Use Case can lead to design choices which lead to different implementations of the AI-based system.

B.3 Intended Purpose: Expectations and pitfalls

B.3.1 Expectations for Intended Purpose in AI?

B.3.1.1 What should we expect from Intended Purpose in the context of AI?

Based on the previous examples, the following statements should be emphasized for Intended Purpose description:

- Described in a concise statement: three to four lines in order to quickly explain what the system can do, what it acts upon (intended population or subject), who is supposed to use it (intended user) and in which context it is expected to operate (intended environment, or operating conditions);
- A more detailed description could take place in the Operational Specification, with additional elements such as:
  - counterindications (on intended user/environments/population);
  - limitations of the system (potentially supported by confidence attributes?).

B.3.1.2 Potential links between Intended Purposes and Operational Design Domain (ODD)?

From the SaMD Guideline proposed by the UK government, the Intended Purpose in the medical field describes 4 key elements**:**

- Intended Population;
- Intended Users;
- Intended Use Environment;
- Structure & Function of the device.

In the meantime, the notion of ODD, that originates from the driving automation taxonomy SAE J3016 in automotive, is a concept used to describe the environment of operation and operating conditions for a given automated feature.

An ODD describes the *"operating conditions under which the system implementing the automated feature is designed to function (…)" (SAE J3016)*. It characterizes a given operating environment and relies on the identification of elements (with attributes and sub-attributes) that are part of the environment the studied feature will evolve in. It can be used to indicate which elements are not considered for the operation of the system, due to restrictions on the feature: counterindications, etc.

It is highly advised to have a preliminary vision of the Intended Purpose of the feature at the beginning of System Design. For each iteration during Design phase, the consistency between the Operational Specification, associated ODD and Intended Purpose shall be maintained. This consistency includes the definition of automation levels, which structure the repartition of responsibilities between human operators and automated systems (this is explored in section F).

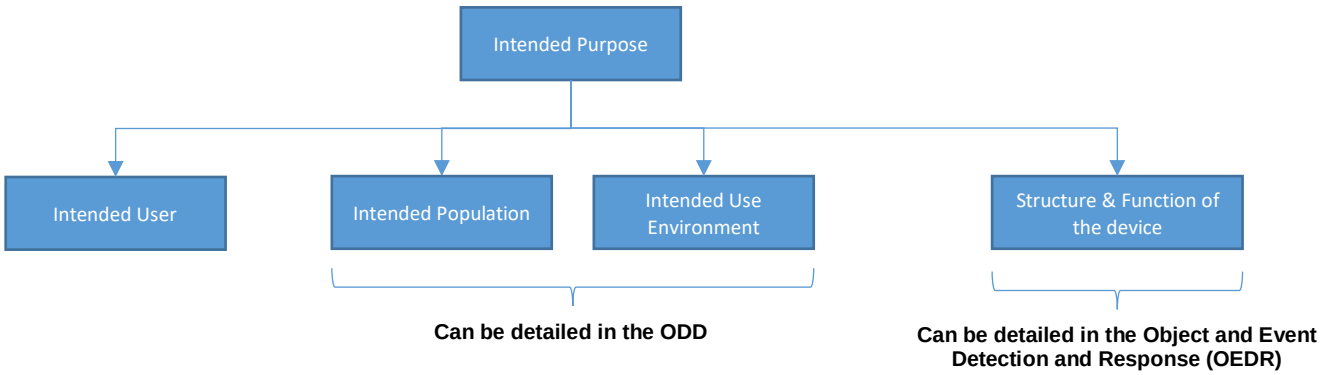


Figure 3 Links of Intended Purpose components with ODD and OEDR

It could be suggested that the notion of **Intended Population** (subjects to observe/monitor, with counterindications) as well as the notion of **Intended Use Environment** (with counterindications) are detailed further in the ODD. In addition, aspects related to the **function of the device** could be described in the expected Object and Event Detection and Response (OEDR), a function that manages the monitoring of the operating environment and the application of an appropriate response when an event arises.

Suggested workflow for specification of Intended Purpose Statement for AI-based Systems

The workflow below takes inspiration from the UK Government Guideline on SaMD (Figure 1) to propose key questions and elements that will help to define the key elements expected for the Intended Purpose statement.

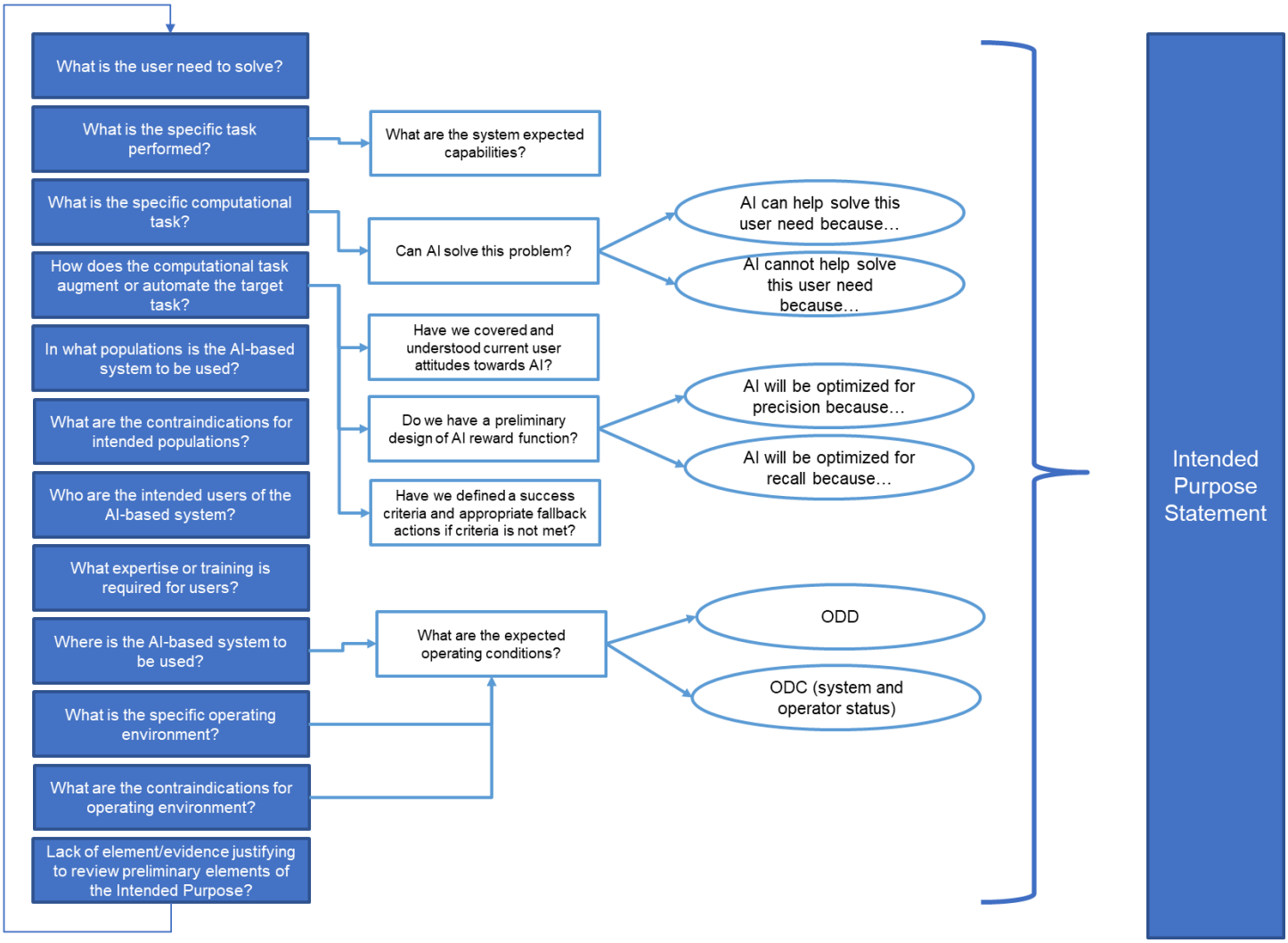


Figure 4 Suggested Workflow for building Intended Purpose Statement

This viewpoint gives a first vision on how to formulate the Intended Purpose statement specifically for AI-based systems, but deserves to be completed through an existing Operational Specification that refines the synthesis of needs for all stakeholders, as explored in the following parts of the document.

B.3.2 Pitfalls of the Intended Purpose

B.3.2.1 Pitfalls to avoid

A proper definition of the Intended Purpose can be challenging. It is important to avoid some common pitfalls while specifying the Intended Purpose as they can have long-lasting impacts later in the development processes.

Among the common Intended Purpose pitfalls to avoid for AI, there is:

- Definition of a vague intended purpose, that leaves room for interpretation
- The risk of function creep: giving to the feature a scope that goes beyond its initial Intended Purpose

- The risk of Multi-purpose devices, which can lead to a confusion regarding what is the exact Intended Purpose of the feature
- The lack of evidence to support that the purpose can be achieved, which can be critical regarding producing pieces of evidence that the feature satisfies its Intended Purpose

B.3.2.2 Business risks

These common pitfalls can lead to the following business risks:

- Serious failures (error in diagnosis by the AI-system due to using the AI-based system on a counter-indicated population of patients, etc.);
- Challenges in setting a proper Quality Management System, that impacts project delays and costs;
- Non-compliance with regulation (AI Act, etc.), which can lead to legal risks;
- Consumer quality risks: warranty costs and customer dissatisfaction;
- Risks for the system improvement: lack of visibility on how changes to the system may impact its Intended Purpose and related engineering items (expected justification and evidence).

B.4 Conclusion on Part 1

In this Part 1, we introduced the Intended Purpose with its origins and some examples. It is a key notion that supports the Operational Approach by explaining what the designer wants to achieve with the system. The Intended Purpose spans throughout the whole Life Cycle of the AI-based system, as it trickles down to the lower layers of engineering activities. While it is a novelty in the domain of AI, its use in another complex open field such as the medical domain shows that it provides essential information in order to design the system “as intended”. The Intended Purpose enables a proper communication to the system users in order to inform them of its intended use.

At Operational level, the place of AI is recontextualized within a larger system, and provides the AI suppliers (data scientists, data engineers, etc.) with a higher view of the expected feature that is not restricted to the AI component. Accordingly, for AI as for other complex engineering domains (mechanical, software, etc.), it consists in considering needs before technical solutions.

The Intended Purpose are supported by four main components: Structure & Function of the Device, Intended User, Intended Population, Intended Use Environment. Some of these components can be associated to well-known notions related to automated driving such as OEDR and ODD. In any case, the expected behavior and the external elements contributing to achieve the mission should be described.

Building an Intended Purpose represents several challenges: ensuring that the Intended Purpose is concise, clear, and defined with an adequate scope. But it is a primordial step for the Operational Approach in order to ensure that the operational needs of the AI-based system are converged among all stakeholders in order to guarantee that implementation remains consistent with the expectations.

C. Part 2 – Operational Design based on reference systems

Introduction to part 2

In section B, we focused on the role that the Intended Purpose can play in the Operational Specification of AI-based Systems. In addition to building a proper Intended Purpose of the AI-based system, the Operational Approach for AI-based Systems could benefit from relying on reference systems for development of new systems. Reference systems are existing systems that use similar technology, pursue similar goals or are related to the same application as the newly developed system. They can help the development by providing field feedbacks, best practices and known limitations from previous systems, as well as arguments for carry-over of existing implementations.

Overview of part 2

We mainly provide leads rather than concrete solutions in this section, but it has been identified as a key step for future work.

In this section:

- We give quick leads on the role that reference systems can play in the Operational Specification;
- We go on by suggesting how those reference systems should be described;
- Additionally, we also cover what kind of elements should be captured for operational design;
- Finally, we discuss how reference system could help to further improve the system.

We focus on the operational aspect of using reference systems from a technical viewpoint, while the section F deals with operational aspect from a human-AI collaboration viewpoint.

C.1 Role of reference systems in Operational Specification

Identification of reference system(s) from an application and technological viewpoint

For designing a new system, whether as an update of an existing system or as an innovation, we always rely on some sort of reference systems. This is the way to optimize the engineering efforts and limit industrial risks and guarantee quality of the results. Reference systems can be:

- Previous products with features covering a reduced scope,
- Products that were fully reliant on human operators with the same delivered service,
- Proof of Concept (PoC) realized during pilot phases of the project for some parts or the whole system.

Reference systems can be considered from the **technological viewpoint** or from the **application viewpoint**. Several reference systems can be necessary to cover the scope of the newly developed AI-based system.

For technological viewpoint, we consider components (sensors, actuators, etc.) or proven-in-use technology that will be implemented in the newly developed systems. We can distinguish the **hardware and software components for sensing and actuating,** the **type of architectural implementation,** as well as the **type of technology** used for processing (AI-based, not AI-based).

For application viewpoint, we consider a system evolving in the same operational context with related stakeholders (automotive, railways, aeronautics, etc.). It often relies on systems providing the same type of services realized by human or with a lower level of automation. It allows to consider the shared situations that both systems will experiment with some feedback already available. Stakeholders can also be expected to express their needs for the new system when compared to the reference systems, in order to improve the relevance of those needs.

The identification of one or more reference systems is key to identify what we could expect or not from a new system development. They are essential to perform impact analyses to assess how higher level of automation may impact the design of the AI-based new system development, what are the known topics that are already under control and what could be current challenges that may impact the success criteria. The Operational Specification must precise what is carried over from reference systems and focus efforts of description on new elements not covered by reference systems, without overlooking the interaction between the new and carried-over elements.

C.2 Description of reference systems (operational and architectural viewpoints)

We need to answer the following questions to identify if a reference system is relevant for the new design:

- What is the type of reference system (applicative, technological)?
- What is the level of knowledge on the system (internal or external, developed by the company or worked on by contractors)?
- What is useful from the reference system? What is not applicable?
- Is there an operational specification existing for the reference system, or do we have to reform the main elements of operational specification for this system? Are the stakeholders of the reference systems available for the new design?

C.3 Capture methodology of relevant elements for operational design

We present the first vision of this methodological step to be experimented and improved in future works:

- Select the relevant reference systems;
- Identify the gaps and commonalities between the reference systems and the target system;
- Check how the reference systems can help to identify the needs for the Operational Specification of the new AI-based system;
- Gather the operational specification elements (stakeholders, use cases, etc.) or feedback elements that can be useful, for each relevant item;

- Check the impacts of reused elements from reference systems on the scope of the new system design not covered by reference systems.

C.4 Capture the new system improvements (automation level, usage) in comparison to reference system(s)

In the section E of EC6.8 of ODD deliverable (Confiance.AI EC6, 2022), dedicated to ODD Engineering Process, we proposed a three-step analytical approach enabling to build the Operational Design Domain (ODD) for a newly-developed system. In this process, the first step is related to the identification of topics considered relevant to address in the ODD. This step is built upon the following inputs: the preliminary design of the new developed feature, and the Industrial State of the Art from which we identify several reference systems for the newly developed system.

The analysis of the gaps between reference systems and the newly developed system is particularly interesting as it highlights specific topics linked to the automation of the system. Indeed, the newly developed system might have a higher level of automation, additional functionalities and/or a greater scope of operating conditions than the reference systems. It is then important through the Operational Specification to identify what those gaps are exactly and how they can be fulfilled to ensure that the newly developed system remains consistent with its associated Intended Purpose.

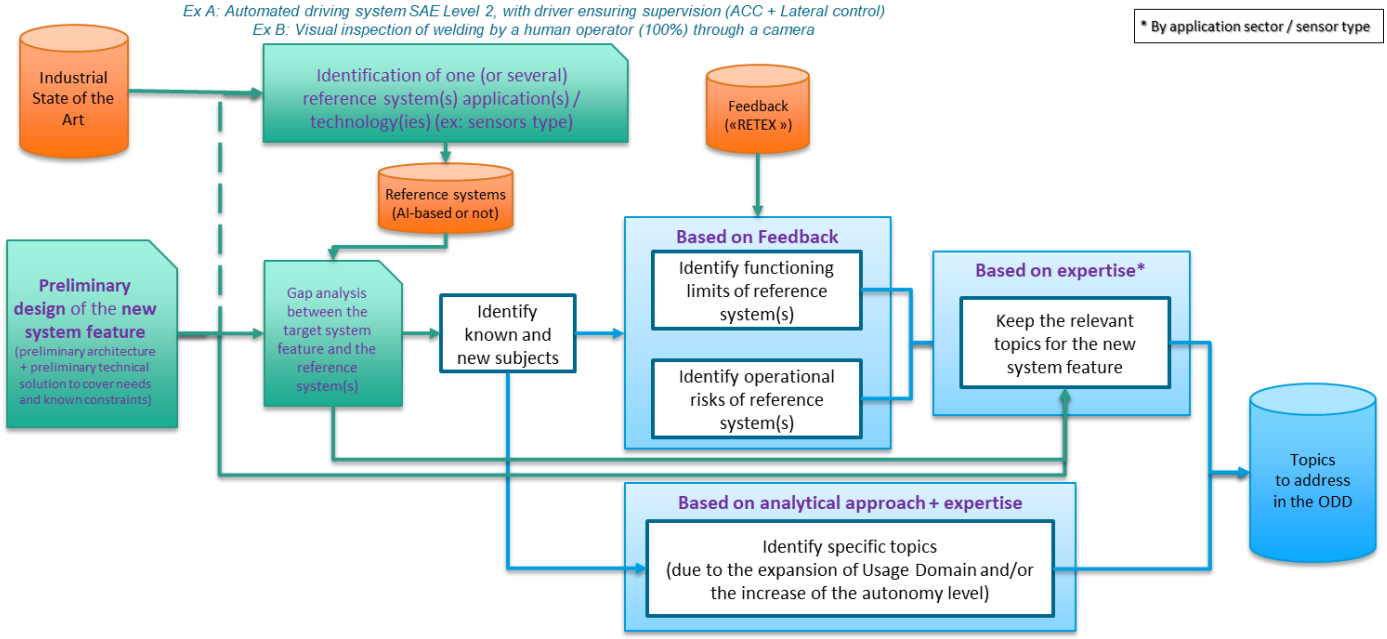


Figure 5 Step 1 of ODD Engineering Process (Analytical Approach) from EC6.8 Deliverable

NB: One of the goals of the Operational and System Approach is to define the green activity blocks of the diagram above, which was out of scope of the ODD process, but which is an important part of EC2.18 work.

C.5 Conclusion on Part 2

Using reference systems bears several advantages when starting a new development. It provides the development team with applicative and technological information that can serve as support for feedbacks and upgrades.

The goal of AI-based systems often consists in automating actions originally performed by humans. In this context, several reference systems can be used: human-based, systems with less functionalities or lower levels of automation reference systems. Those reference systems are likely to have an associated Operational Specification that can be worked upon for a new development.

Using reference systems enable to identify easily involved stakeholders and their characterized needs. The available level of information depends on the typology of the reference system: previous development of the company with high visibility on engineering tasks, or external development with restricted information.

Reference systems allow to identify more precisely, from the beginning of design activities, the level of collaboration between humans and systems and resulting automation levels. This aspect is explored further in section F of this deliverable.

D. Part 3 – Bottom-up approach: trade-offs between technical limitations and opportunities impacts at operational design level

Introduction to part 3

As seen in the section C, AI-based system development can be considered when we go over the limitations of what can be achieved by conventional systems. It results in a double innovation: an innovation in terms of use that relies on a technological innovation, thanks to AI. The service provided by the AI-based system is an innovating system for which there is not yet a matured state of the art. Thus, it is necessary to find a proper trade-off between the provided service and the constraints and risks of the technology. Once this trade-off has been found, the stakeholders need to be involved again to ensure they can adjust their needs in regards to the newly achieved trade-off.

For Operational Approach, a top-down view from stakeholders needs on system is not sufficient to raise potential operational issues related to implementation choices. From a bottom-up view, there are several constraints (technical, technological, etc.) on systems from expected implementation. These specific constraints need to be taken into consideration at operational level in order to better refine the operational needs of the stakeholders. Beyond identifying the appropriate Intended Purpose of the system through their needs, the stakeholders are also involved in refining the needs that are enlightened by the choice of technologies.

The Operational Specification aims to ensure the consistency between the synthesis of needs and the synthesis of constraints, thanks to the trade-off analysis.

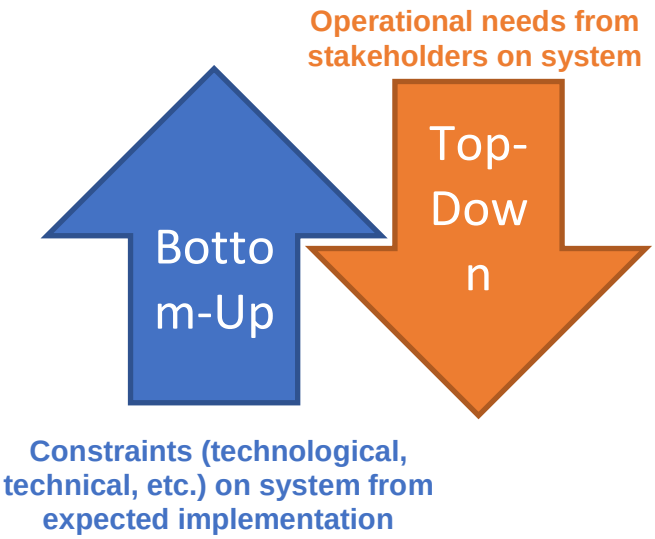


Figure 6 Benefits from combining Top-Down and Bottom-up for Operational Specification

Overview of part 3

In this section that is mainly driven by examples:

- We showcase what kind of operational impacts, limitations and opportunities should be captured in AI Operational Specification;
- Then we illustrate how those limitations and opportunities could be captured;
- Finally, thanks to existing works in the field, we evoke the manner trade-offs can be deemed necessary through gaps between the design intent and the AI-based system real implementation.

D.1 Methodology to analyze operational impacts of AI-based systems

D.1.1 Example of SOTIF Bottom-up Approach

ISO 21448:2021 standard (*ISO 21448*, 2022, p. 21448) deals with the Safety of the Intended Functionality, in automotive. This is a complementary approach to automotive Functional Safety already covered by the ISO 26262:2018 standard (*ISO 26262-1*, 2018) which deals with the impact of failures on EE (Electrical Electronic) systems. On the contrary, SOTIF deals with insufficiencies of specifications or limitations of performances that can entail safety hazards for EE systems.

ISO 21448 relies on the main notion of triggering conditions that could entail hazardous behaviors and lead to hazardous events. It starts with an analysis of the Specification and Design of the studied system (Clause 5) and deliver several analyses to verify whether this Specification is SOTIF-compliant or not. If a potential risk is detected, the Clause 8 of SOTIF suggests several types of Functional Modifications to apply on the system in order to ensure the Specification and Design is properly updated for a new iteration of the ISO 21448 process.

This is an interesting illustration of trade-offs from a Safety viewpoint, as several actions can be undertaken to appropriately mitigate a given SOTIF risk.

One example to illustrate this situation with a potential AI-based system: a Traffic Jam Chauffeur (TJC) with SAE Level 3, that operates the vehicle on highways during traffic jam with a speed up to 60 kph. It is implemented in Mercedes Drive Pilot in Germany and some states of the United States. This type of feature relies on sensors for the perception of the road and the environment conditions.

The operation of this automated feature could be affected by meteorological conditions, such as dense fog, which can impede the perception of the system and could present a risk. Thanks to SOTIF, several proposals could be made to solve this issue:

- The automated feature should pursue its mission at a reduced speed, by relying on other sensors, if relevant (*Restriction of authority for the intended functionality for specific use cases, clause 8.3.3*).
- The automated feature cannot handle this change in operating conditions, it requests a driver takeover while operating at a reduced speed (*Removal of authority for the intended function for specific use cases, clause 8.3.3*).

This example illustrates how two alternatives could enable solving a particular issue, with their pros and cons, hence the trade-off in this particular case. It is part of a bottom-up approach as they will be used to raise elements on design, ODD, or human-machine collaboration strategies that are requested by the design, and upon which the stakeholders must take position.

Other types of Functional Modifications exist, but they do not necessarily involve the AI-based nature of the system.

D.1.2 Example of EC6.8 ODD Engineering Process expectations

An additional example could be inspired from the EC6.8 ODD Engineering Process, presented in C.4 above. Indeed, the suggestion to perform an impact analysis by using reference systems and the preliminary design of the newly-developed systems, could be applied in order to highlight operational limitations and opportunities that result from the use of AI.

D.2 Capture of technical limitations and opportunities that should impact the system design

D.2.1 Impacts of reference systems and AI on trade-offs

D.2.1.1 Example of Welding Use Case: automation trade-offs

Impact of AI on the nature of application and the associated operational needs

The Welding Use Case evoked in section B.2.2 is an interesting case due to the nature of its reference system.

In the current AI application, the AI component validates automatically mechanical parts that it is confident about, while letting the operator validates the ones that are not conform or that are unclassified. It is considered as a decision support AI-based system.

The AI-based application of Welding introduces several items and notions that could be traced in the Operational Specification of the system:

- It provides an answer to a **classification problem**: each mechanical part belongs to a single discrete category (among three: Conform, Non-Conform, Unclassified)
- The third category (unclassified) is needed for classification in the AI application as:
  - It ensures that the human operator is still involved in the classification process
  - It gives more flexibility to the trustworthiness threshold by better considering the potential unreliability of the AI agent
  - It takes into consideration the fact that the dataset is unbalanced regarding the non-conform parts (low quantities due to the assurance quality process)
  - It plays a part in maintaining the human operator trust in the AI-based system

This Confusion Matrix could be established for conventional human-based Visual Inspection of Welding parts:

		Prediction (by human operator)	
		Conform	**Not Conform**
Ground Truth	Conform	True Positive	False Negative
	Not Conform	False Positive	True Negative

Figure 7 Initial confusion matrix for UC Welding

For the current application in Confiance.ai, with the addition of the third category for “Unclassified”, it might take the following form:

		Prediction		
		Conform	**Not Conform**	**Unclassified**
Ground Truth	Conform	True Positive	False Negative	Status by human operator
	Not Conform	False Positive	True Negative	Status by human operator

Figure 8 Illustration of confusion matrix for UC welding considering the "Unclassified" category

NB: The ground truth supposes that the welding part is either conform or not conform. Unclassified as Ground truth would not make much sense in this context, unless we consider as Ground Truth the classification from a qualified human operator.

This is a concrete example of how AI-related constraints can be raised in a bottom-up approach and impact the initial operational needs.

Definition of success criteria

The Welding Use Case is also interesting as it gets several success criteria that relies on the existing Visual Inspection Use Case at Renault

For example:

- **No False Positive** (images of non-conform welding parts that are classified as conform): This relies on existing Quality processes (thus, on reference human-based systems)
- **Less than 5% of False Negatives** (images of conform welding parts that are classified as non-conform): This is directly linked to the AI application, in order to prevent an additional workload for the human operator
- **Between 10% and 15% of unclassified welding parts** (Unclassified): This is directly linked to the AI application, in order to ensure that the human operator remains vigilant at their posting

There a clear notion of trade-off in this use case with the definition of these 3 thresholds, thresholds that are related to the AI classification nature of the UC.

Thus, with other types of AI problems (regression, etc.), an interesting question would be: what are the relevant metrics enabling an adequate trade-off, that ensures a proper collaboration between human operators and AI systems? How do we identify them (through existing processes, through automation or augmentation considerations, etc.)?

Synthesis on this example

The example of the Welding Visual Inspection Use Case highlights that implementing AI on an application which was previously human-based will raise potential limits due to technological maturity (e.g. the need for an “Unclassified” category to take into consideration the limits of AI). This will cause to define new operational needs, that will be associated to the identified limits and will help to consider new transitions between operational states.

D.2.1.2 Example from automotive: automated driving trade-offs

Identifying as soon as possible how technical limitations can impact the operating states of the AI-based system is a key information to have in the Operational Specification.

We can use as an example in automotive the use case of First Responders interacting with automated vehicles: Both Waymo (with their current robotaxi deployment in SAE Level 4) and Mercedes (with their TJC Drive Pilot in SAE Level 3) have made different choices of OEDR upon detecting a First Responders vehicle (police, emergency, etc.).

In the case of the Drive Pilot from Mercedes: *“DRIVE PILOT always drives cautiously and takes extra precaution upon detection of emergency signals and/or pedestrians. When such situations are detected, the driver is asked to take over control - until they do, DRIVE PILOT reduces its speed and creates a larger gap or lateral offset”* (Mercedes Benz, n.d.). The SAE Level 3 as defined in SAE J3016 expects that a fallback-ready user takes over when requested by the automated feature. Thus, despite the system capabilities to detect first responder vehicles, the driver is expected to ensure the fallback as intended by the SAE Level 3.

In Waymo case, they describe the vehicle response in the following manner: *“If a Waymo autonomous vehicle detects that a police or emergency vehicle is behind it and flashing its lights, the Waymo vehicle is*

designed to pull over and stop when it finds a safe place to do so." (Waymo, 2023). This is also consistent with the SAE Level 4 attributed to this system, which does not request that a fallback-ready user take over the driving, since the automated system is able to manage these situations in the defined ODD.

Thus, we can observe on the same AI-based application (automated driving) how the technical limitations and design intent of each developed feature can impact the trade-offs regarding the response that is provided by the systems in a given situation. Identifying as early as possible these limitations in this situation gives leads to identify the choices of technology regarding takeover requests management.

The Operational Specification is expected to support the choices of design that can be technical or technological constraints.

D.2.2 Vision of specification and trade-offs

The team at Google DeepMind (Pedro A. Ortega et al., 2018) shares an interesting vision of the specification that is defined as *"ensuring that an AI system's behavior aligns with the operator's true intentions"*. They add that *"the challenge of specification is to ensure that an AI system is incentivized to act in accordance with the designer's true wishes, rather than optimizing for a poorly-specified goal or the wrong goal altogether."*. This illustrates issues that a lack of Operational Specification for AI-based system can potentially lead to.

They are aligned with this article from CSET (Tim G. J. Rudner & Helen Toner, 2021) that describes three visions of Specification:

- **Ideal specification** ("wishes", here the Ideal view): What task the designer of a Machine Learning system wants the system to perform. hypothetical (but hard to articulate) description of an ideal AI system that is fully aligned to the desires of the human operator. *"Hypothetical description of an AI system's objective that is fully aligned with the human designer's desires"*.
- **Design specification** ("blueprint", here the Operational & System Specification): The proxy objective the System Designer specifies to enable the system to learn to perform the task. specification that we actually use to build the AI system, e.g. the reward function that a reinforcement learning system maximizes. *"Specification actually incorporated into the system (proxy the System Designer chose to implement)"*.
- **Revealed specification** ("behavior", here the component capabilities): What the system actually does. Specification that best describes what actually happens, e.g. the reward function we can reverse-engineer from observing the system's behavior. "*An imperfect proxy of the ideal specification. (..) Is the observed behavior when deploying a ML system in the real world"*.

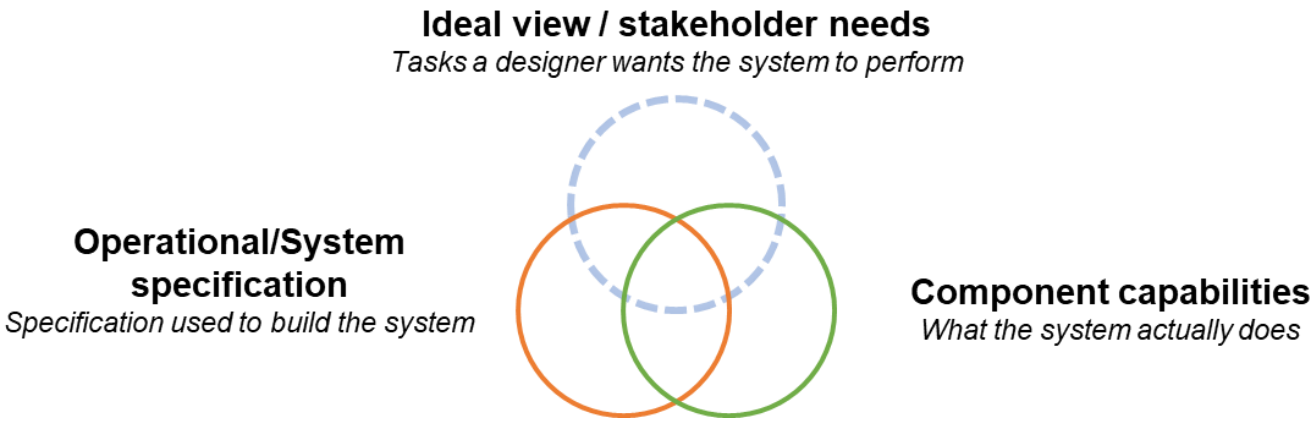


Figure 9 Venn Diagram of Ideal/Design/Revealed specifications

An Ideal Specification may be directly inspired by a reference system: we know what it does, we perceive its explicit and implicit characteristics, but we are not able to achieve a complete translation from what the reference system is, to what the new system should be.

The target of the trade-off is to find a proposal of solution that enables the best superposition of all three circles above: satisfying the stakeholders needs while considering the technological limitations.

D.3 Conclusion on Part 3

Through the examples in this section, we had a sneak peek on the importance of the Operational Approach in addressing identified limitations and opportunities in AI-based system development.

Combining a top-down approach with a bottom-up approach in the design of AI-based systems will result in discussing alternatives and design choices with their respective pros and cons: it highlights the need for trade-offs analyses.

The Operational Approach should be a synthesis of the trade-offs between the Operational Needs of the stakeholders of the AI-based system (driven from the top) and the constraints resulting from the system implementation (coming from the bottom). The stakeholder's needs must accommodate the system design, and the system design must be an answer to the stakeholders needs. Ensuring the traceability of such trade-offs will also contribute in the quality assurance of the expected AI-based system.

Another manner to consider the trade-offs of an AI-based system is through the notions of Ideal, Design and Reveled Specifications: the lack of overlap of those perceptions lead to expose blatant limitations of the systems, as well as more pernicious issues related to a misalignment between the expected goal from the human operator and the real goal pursued by the AI-based system. This will be explored further in section F.

Discussing trade-offs in the Operational Approach can span several phases of the life cycle of the AI-based system. In the following section, we will see what are those phases and how compatible they are with the life cycle of conventional System Engineering.

E. Part 4 – Commonality of operational methodology for System Engineering and AI Systems

Introduction to part 4

With the emergence and rapid evolution of AI-based systems, several big players of the AI domain are from the tech industry, with a software development and programming culture. The software environment tends to evolve differently from more traditional environment, as illustrated by the popular principles of CI/CD (Continuous Integration/Continuous Delivery) where software-based products and their updates are shipped at a high frequency in order to bring stability as soon as possible. Due to the fact that AI-based products present similar characteristics to software-based systems, we also observe a similar trend of releases and updates from AI current big players in the tech industry. However, at the same time, several initiatives from those players show a change in paradigm, as entities such as Google and Microsoft try to incorporate conventional System Engineering methods in their AI process. Google has developed in 2019 the Google PAIR Guidebook (*Google Pair Guidebook*, 2021), and Microsoft has capitalized upon their HAX Toolkit and research papers ('HAX Toolkit Project', n.d.) to provide guidelines that directly inspired their AI-based solutions (in their Office and Windows brands). This illustrates that software-based processes are not sufficient to cover aspects of AI implemented into complex systems (rather than single components) that rely on the technology to achieve a given service. This is an incentive for jobs such as Data Engineers and ML Engineers to think about the purpose of the developed AI components implemented in a given system.

The approach that we propose in this work for AI is consistent with conventional System Engineering industrial approaches. It meets naturally the bottom-up approach of big players who industrialize AI-based systems. Our aim is to find ways to incorporate aspects of AI into the Conventional System Engineering Process. Proceeding in this manner is justified by:

- The profiles of Confiance.ai stakeholders, who are mainly from non-AI backgrounds;
- The fact that System Engineering benefits from decades of experience and feedback, which makes it a robust process to lean on for transverse topics. The idea is to find how we can adapt some aspects of System Engineering with AI while retaining the rigor of the process.
- The fact that this particular topic seems rather new, as the SEBoK still considers that it is an emerging topic. This represents an opportunity to pursue.

For these reasons, we will observe in this section what are the aspects that Conventional Engineering share in common with AI, starting with Life Cycle phases.

Overview of part 4

In this section:

- We first provide leads on some existing references for operational design;

- We then rely on the SEBoK in System Engineering to go over how AI impacts the existing life cycles that are standard practices in industries with a System Engineering culture.

E.1 Identification of a referential for operational design

Referential for System Engineering: SEBoK

The System Engineering Book of Knowledge (*SEBoK*, 2023) is a guideline on System Engineering. It is built upon an important amount of industrial knowledge, methods and best practices on the main principles of System Engineering. It is supported by three entities: the INCOSE (International Council of System Engineering), the IEEE Systems Council, and the Stevens Institute of Technology. It references several key standards such as the ISO 15288 dedicated to Systems and Software Engineering. It is regularly updated and has started to incorporate emerging topics, among which can be found the rise of AI, but it remains a minor topic at the moment inside the SEBoK.

E.2 Role of Operational Specification for AI in the Life Cycle

In the following section, we will go over conventional Life Cycle phases, according to SEBoK and ISO 15288:2023, and how they are or could be impacted by the development of AI-based systems.

Although the Operational Specification is established during the Concept phase, it will impact and treat the other phases of the Life Cycle for a given system. This should be particularly important in the context of AI due to the numerous changes it brings.

E.2.1 Life Cycle Stages: generic steps

SEBoK proposes a generic Lifecycle model (Kevin Forsberg et al., 2023) that presents a linear vision of how a given system will evolve through time:

- The **Concept Definition** identifies Concepts of Operations (ConOps) and Business case, key stakeholders and the expected desired capabilities.
- The **System Definition** takes as inputs the results of the Concept Definition and analyzes the feasibility of the system definition
- The **System Realization** as a bridge between the System Definition and the System Deployment and Use where several iterations are performed to consolidate and/or amend the system design, identify potential challenges, and prepare for the deployment of the system of interest.
- The **PSU&R phases** describe the coverage of Production (including updates of the product), Support and Utilization & Retirement. The business will dedicate resources to produce, support, utilize and recycle the system on its expected lifetime.

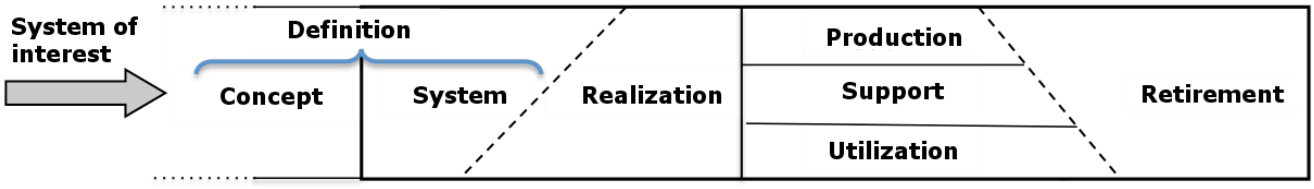


Figure 10 Generic Life Cycle Model from SEBoK

E.2.2 Role of the Operational Specification in the Life Cycle

The Operational Specification shall gather the needs and stakeholders regarding every stage of the life cycle of an AI-based system.

We can take the example of the TJC in automated driving evoked in section D.1.1 above. This automated feature is only applicable on motorways in specific countries or regions. The Operational Specification shall consider the needs for maintenance of the feature by a garage, for which the area of operation is not reachable during day of work. The needs for diagnosis and test after repairs have to be considered.

As seen in section D, the reference systems can help in addressing the different phases of the life cycle.

E.3 Identification of AI-based system stakeholders

Definition of a stakeholder

One goal of the Business or Mission Analysis from SEBoK is to identify the Enterprise Strategic Goals and the Stakeholders needs. Regarding stakeholders, the ISO 15288:2023 (*ISO/IEC/IEEE 15288*, 2023) proposes the following definition:

- *"individual or organization having a right, share, claim, or interest in a system or in its possession of characteristics that meet their needs and expectations. End users, end user organizations, supporters, developers, customers, producers, trainers, maintainers, disposers, acquirers, suppliers, regulatory bodies, and people influenced positively or negatively by a system."*.

This definition highlights various notions such as suppliers who may not be involved at the same level of development than the main developers of AI-based systems.

With the Operational Approach, the idea is to appropriately involve the right stakeholders when needed.

Why is it important to properly identify the AI stakeholders?

Through stakeholders, we take an orientation on a design choice. The stakeholders must be actors of the design, with a proper vision on the consequences of their choices.

The aim of Operational Specification is to translate Stakeholders Needs into Functional Requirements.

Needs Definition drives the design, and the synthesis of needs from multiple stakeholders may reveal contradictions. These contradictions must be analyzed to achieve trade-offs, in order to decide to create a consistent convergence of Needs.

As for System Engineering, the actors of the development process are not considered as stakeholders. However, new specific stakeholders related to the AI-based systems shall be considered: ethics, data quality, etc.

Iterative approach with stakeholders

In order to express a set of needs, a scope shall be defined. A preliminary Intended Purpose shall be built with the main stakeholders of the system under development. It allows to identify other stakeholders and

make them express their needs. After a first synthesis, an iterative process occurs to enable each stakeholder to align and detail their needs.

E.4 Identification of needs for AI-based systems

The identification and satisfaction of needs may be more complex on AI-based systems due to the hardships to properly define the needs, and to define success criteria that achieve the expected performances.

The needs allocated to a given system follow several steps of maturation and transformation, from their capture, to their translation into requirements (for stakeholders, systems, etc.), in order to have a system implementation aligned with the Operational Specification. SEBoK has proposed a Cycle of Need during System Engineering development process (Alan Faisandier et al., 2023). We tweaked this cycle in Figure 11 to include new elements related to AI-based systems and how they could be impacted (suggestions in red):

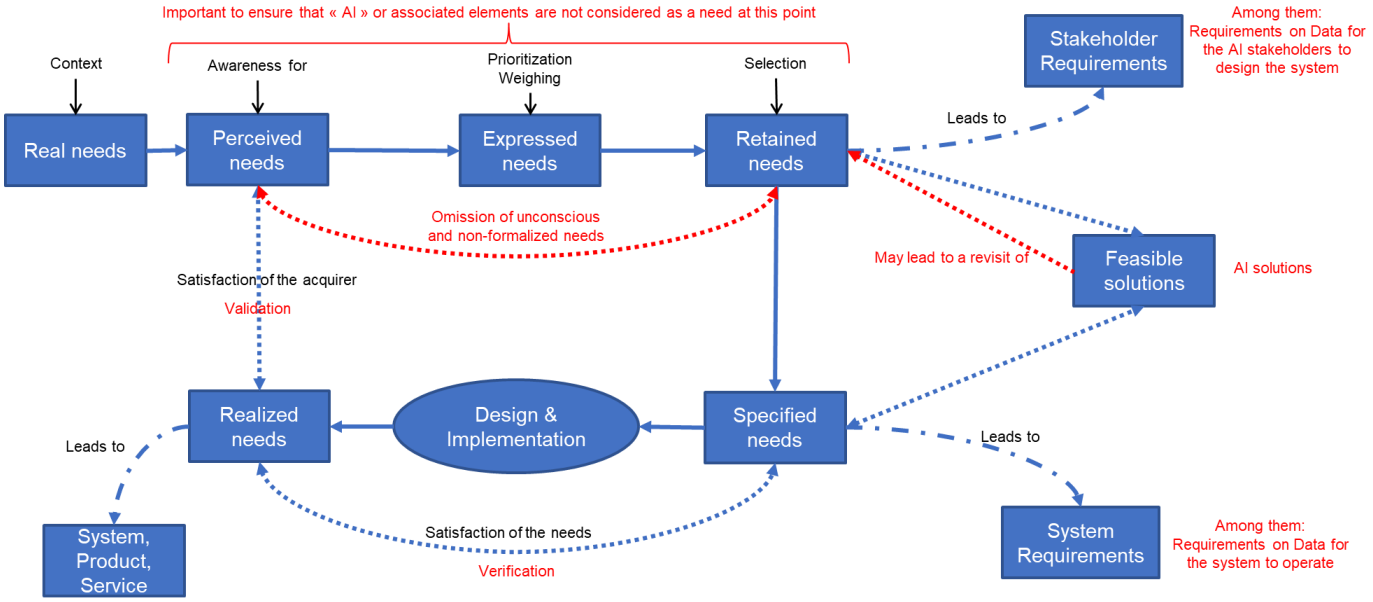


Figure 11 Modified version of Cycle of Need to include AI considerations (from Faisandier)

- "Perceived Needs" can include needs that the user feels (physically or through other means) but that they may not convey properly or that they do not perceive consciously despite its importance;
- Thus, "Perceived Needs" may be larger than "Expressed Needs", which may also be larger than the "Retained needs";
- In this context, a gap must be considered between perceived needs and retained needs. It translates a potential omission of unconscious or non-formalized needs that may be experimented only after a first implementation of the system.

There is also the question on how to align the user's mental model ("Perceived needs") with the Intended Purpose of the system which may influence the "Retained needs", to ensure that the "Realized needs" cover this mental model.

Finally, this Cycle of Needs also raises the question of how to consider data, and how to choose the relevant data for a given AI-based system.

E.5 Needs regarding data engineering in the Life Cycle

Data serves as the basis for all AI-related activities, from gathering of raw data to training of AI models and use of data to draw up metrics in order to deploy systems. This means that data has to be managed throughout the life of its associated AI-based systems, and beyond. It adds more constraints for the management of data and needs to guarantee that the designed system does not drift from its initial Intended Purpose throughout its Life Cycle, which is a requirement from the AI Act (Requirement 49 “High-risk AI systems should perform consistently throughout their lifecycle”).

Besides, use of AI-based systems likely to gather personal data is regulated in Europe with the General Data Protection Regulation (GDPR, 2016). It shall imply needs during the whole life cycle of the products: design, production, transportation, use and retirement.

E.6 Conclusion on Part 4

Initially, we consider the System Engineering paradigm, used in the industry, which consists in starting from the elicitation of needs and refining them in order to instantiate technical solutions. Major players of the AI industry take the other way by starting to incorporate System Engineering in their software culture. The committed stance of this deliverable is to present the joining points of this global approach. It includes the changes brought by AI in a System Engineering culture. System Engineering is supported by an established set of industrial standards and best practices that obtained a consensus. In the meantime, AI-based systems allow to address complex situations with a lower level of efforts in comparison to the first level of results. The resulting methodological guideline shall improve the maturity of AI engineering process, without reducing the benefits of this specific technology, in particular for functionalities that cannot be performed by conventional systems.

In this context, studying conventional System Engineering Life Cycle stages is a good opportunity to identify what activity has to be performed for Operational Design of AI-based systems. In a methodological point of view, stakeholder identification, capture and synthesis of needs and consideration of the whole life cycle of the system remain applicable.

F. Part 5 – Focus on responsibility boundaries between system and users

Introduction to part 5

The previous activities of the Operational Approach enabled to build a first impression on the stakes of building a proper Operational Specification for AI-based systems.

The Part 6 builds on the synthesis from previous activities. It focuses on how AI redefines the human-system interactions and what must be considered to ensure that the role of the end user is properly defined in the Operational Specification.

The use of AI highlights the following: what was once covered by the human capabilities and out of scope of the operational specification becomes mandatory now that AI-based systems are likely to take over some human responsibilities.

Redefining the boundaries between human & AI-based systems, starting from human-based reference systems, is key to identify what kinds of services AI technology will help to provide and what kind of products can emerge from these analyses.

Overview of part 5

In the following section, we present how the Operational Specification of AI-based systems is impacted by new activities related to the delegation of functions from human operators to the system:

- We first explore what the transition from human-based to automated systems entails, and how it justifies the need for defining automation (and AI) levels;
- Then, we illustrate through examples how some industries use human-system collaboration frameworks and how they can help to define an AI-based application;
- In a third part, we warn about the use of automation levels at AI component levels and how they do not replace a proper classification at system feature level;
- Finally, we give some recommendations on the way to define shared responsibilities in Operational Specification for AI-based systems.

F.1 The foundation for human and system shared responsibilities

F.1.1 Transition from human reference systems to AI-enabled systems

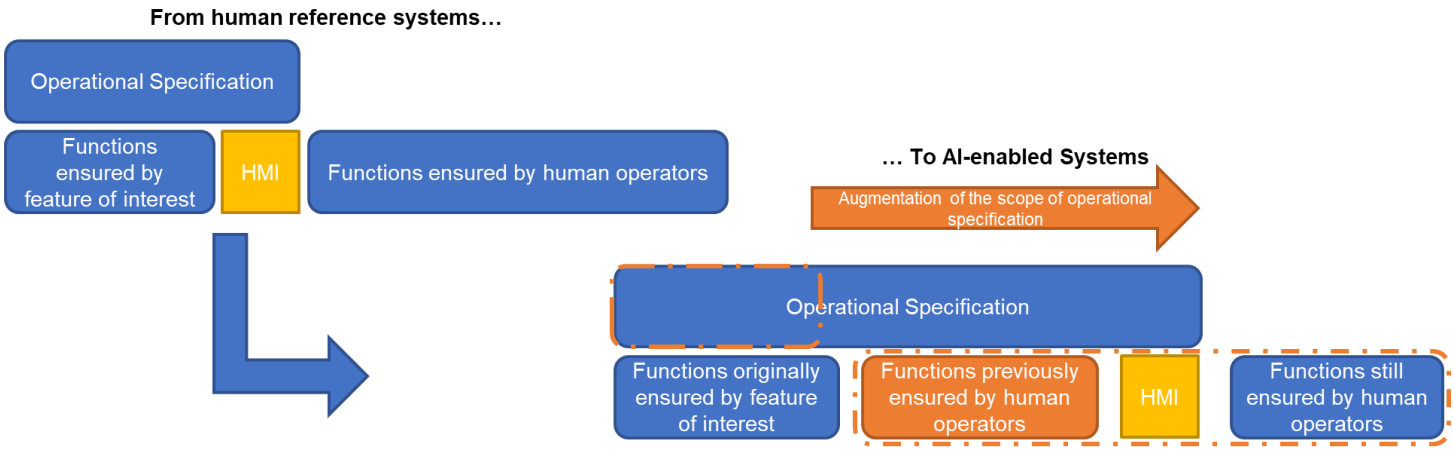


Figure 12 Evolution of Operational Specification from human reference systems to AI-based systems

Going from conventional systems (generally human-operated) to automated (potentially AI-enabled) systems bears several impacts on the engineering items:

In Conventional system, we can consider that the Operational Specification covers the Functions ensured by the feature of interest/under development, as well as needs regarding the interface between humans and systems. The functions ensured by human operators are already known (human operators are either informed, trained, or qualified for them) and are not as thoroughly described.

However, in automated systems, we can notice a shift: with the delegation of functions to the system previously performed by human operators, the Operational Specification has a bigger scope to cover. Functions that were not extensively explicit due to being under human hand need to be thoroughly described for the system to be able to operate properly. Thus, the main challenge of automation (independently of the use of AI or not) is to clearly identify which functions/tasks are still ensured by human operators, which ones can be delegated to the system (automated, and how the latter should be described.

These functions previously ensured by human operators that are then delegated to the system reveal the following challenges:

- There is often a lack of knowledge and description of these intuitive functions assumed by humans;
- Additional efforts must be granted on the Operational Specification of AI-based system to process the delegation of activities;
- How can these functions be captured and translated into requirements for AI-based systems?

In order to provide a framework for delegation from human to systems, a common method consists in establishing a classification of levels of automation for a given feature. The following section F.2 below will detail some examples.

F.1.2 Why do we need automation levels?

The classification framework of SAE J3016 is the main reference for automated driving. Indeed, the associated automation levels bear several characteristics that make them efficient tools for design:

- They are a direct consequence from the needs to delegate more and more responsibilities from a human operator to a system operator;

- They are a core component of a feature, as they symbolize a design intent;
- They give a framework on how to delimit human & system responsibilities: it gives a first view on what the system can and cannot achieve on its own;
- They give leads on some design orientations at system levels;
- They are mandatory to talk about the Operational Design Domain (ODD).

To classify a feature according to automation levels, we must first define what this feature aims to achieve and what are the expected functions we want to automate for a given activity.

For this purpose, we will present an example of driving automation in the automotive domain. Besides, we will also explore how automation is performed in another domain: the train operation in railways.

F.2 Feature level automation: examples of automation levels in industry

F.2.1 SAE J3016 in automotive

F.2.1.1 Presentation of SAE J3016

SAE J3016 proposes a taxonomy for the automation of driving and is focused on automation of driving features. It proposes 6 classes of automation depending on who assumes the responsibility of parts or all of the driving activities (see Figure 13). It is not a Safety standard nor an AI standard for automotive.

There is a single mention of ‘artificial intelligence’ in the whole document. AI is only a mean among others to automate the system.

It also defines key concept such as ODD and OEDR.

A feature is defined, in SAE J3016 for automated driving, as *“a Level 1-5 driving automation system’s design-specific functionality at a given level of driving automation within a particular ODD, if applicable”*. A feature should be considered as an answer for a given service to the user that aims to be automated. In automotive, this corresponds to the service associated to driving (or operating a vehicle).

	Level	Name	Narrative Definition	DDT		DDT Fallback	ODD
				Sustained Lateral and Longitudinal Vehicle Motion Control	OEDR		
Driver Performs Part or All of the DDT							
Driver Support	0	No Driving Automation	The performance by the driver of the entire DDT, even when enhanced by active safety systems.	Driver	Driver	Driver	n/a
	1	Driver Assistance	The sustained and ODD-specific execution by a driving automation system of either the lateral or the longitudinal vehicle motion control subtask of the DDT (but not both simultaneously) with the expectation that the driver performs the remainder of the DDT.	Driver and System	Driver	Driver	Limited
	2	Partial Driving Automation	The sustained and ODD-specific execution by a driving automation system of both the lateral and longitudinal vehicle motion control subtasks of the DDT with the expectation that the driver completes the OEDR subtask and supervises the driving automation system.	System	Driver	Driver	Limited
ADS ("System") Performs the Entire DDT (While Engaged)							
Automated Driving	3	Conditional Driving Automation	The sustained and ODD-specific performance by and ADS of the entire DDT with the expectation that the DDT fallback-ready user is receptive to ADS-issued requests to intervene, as well as to DDT performance-relevant system failures in other vehicle systems, and will respond appropriately.	System	System	Fallback-ready user (becomes the driver during fallback)	Limited
	4	High Driving Automation	The sustained and ODD-specific performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will need to intervene.	System	System	System	Limited
	5	Full Driving Automation	The sustained and unconditional (i.e., not ODD-specific) performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will need to intervene.	System	System	System	Unlimited

Figure 13 Automation levels in automotive from SAE J3016

F.2.1.2 Mandatory information: Main function of the system under interest

The SAE J3016 was introduced as a catalyzer to help develop a common understanding about the landscape of automation in automotive, in order to achieve fully automated driving.

But what does ‘driving’ or ‘operating a vehicle’ means in the first place?

Operating a vehicle (aka Driving) can be summarized as a set of activities, performed by an agent (human or system), destined to perform the entire DDT of said vehicle.

The purpose of a vehicle is to transport humans and/or objects from point A to point B, by using dedicated infrastructure and following defined rules (Road Code, etc.).

Operating a vehicle can be achieved in several fashions that can give birth to different solutions at operational levels, hence the importance to properly define and decompose what operating a system means and how we expect the system and human operators to cooperate.

In automotive, to achieve automation of driving, we rely on a commonly known reference system: driving ensured by a human driver.

F.2.1.3 The various types of functions needed to operate a system

Operational, Tactical and Strategic Functions

By decomposing the unitary actions needed by the human driver to operate a vehicle, we can identify “functions” that serve as a basis to establish a sharing of responsibilities between a human operator and an automated system.

Currently, in the automotive domain, automation (by AI or other technologies) is only applied to operational and tactical functions of a system. Indeed, strategic functions, by essence, is still performed by human action as they result from a human skill to plan and see ahead of time in order to accomplish a trip.

- Operational functions are functions that needs split-second and sustained reactions in real time.
- Tactical functions are functions needed to maneuver the vehicle in real time.
- Strategic functions are differed functions that enable to choose where, when and whether the system should be operated.

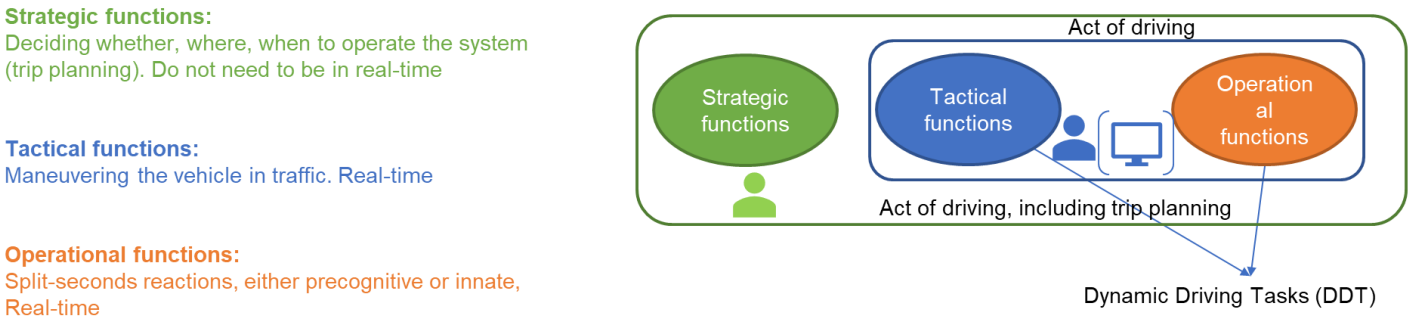


Figure 14 Types of functions needed to operate a system, according to SAE J3016

Tactical and Operational Functions: DDT

In SAE J3016, operational and tactical functions needed to operate the system are called Dynamic Driving Tasks (DDT). DDT constitute the mandatory (and often intuitive) actions that a human driver must perform to operate a vehicle in a nominal fashion. The achievement of these functions is made possible by the knowledge and experience of the human driver (learnt through the cursus needed to obtain a driver's license, which assesses one's ability to drive a vehicle safely). Nevertheless, we can perceive how some functions can be delegated to a system.

In automated driving, DDT are performed by the system, the human operator, or a combination of both. When a DDT is ensured by the system, we can envision the possibility that this could rely on AI-based technologies.

The concept behind Dynamic Driving Tasks (or similar notion) seems necessary to define what is needed to operate the system, first as a human operator, then as an automated system.

F.2.2 IEC 62267 in railways with Grades of Automation

F.2.2.1 Context of IEC 62267

Railways field is also involved in automation, which led to the creation of Grades of Automation (GoA) from IEC 62267:2009 (*IEC 62267*, 2009) and IEC 62290-1:2014 (*IEC 62290*, 2014). Similarly, to SAE J3016, the IEC 62267 standard decomposes the activity of "operating a train" to determine:

- What are the basic functions mandatory to operate a train and usually performed by a human operator (GoA0)?

- What are the expected levels of automation that can be reached and what are the shared responsibilities of human and system operators?
- What are the mandatory implementations in order to achieve a certain level?

F.2.2.2 Operating a Train

The IEC 62267 defines what are the expected basic functions mandatory to operate a train. It relies on several functions such as:

- Some Operational and Tactical functions (“Dynamic Tasks”) such as:
  - Setting the train in motion on its track (longitudinal management);
  - Stopping the train at defined stations (longitudinal);
  - Ensuring Door closure conditions;
  - Ensuring Operation in event of disruption.
- Expected Strategic functions (still ensured by human operators):
  - Planification of station stops;
  - Management of traffic regulation.

For train operation, the reference system is On-sight train operation (GoA level 0). From this level, it is easier to identify what can be automated in regards to what is already achievable through human operation.

The Figure 15 below further detail the classification of automated train operation in regards to the shared responsibilities for each basic function:

Basic functions of train operation		On-sight train operation	Non-automated train operation	Semi-automated train operation	Driverless train operation	Unattended train operation
		GOA 0	GOA 1	GOA 2	GOA 3	GOA 4
Ensuring safe movement of trains	Ensure safe route	**Operations staff** (points command/control in system)	**Systems**	**Systems**	**Systems**	**Systems**
	Ensure safe separation of trains	**Operations staff**	**Systems**	**Systems**	**Systems**	**Systems**
	Ensure safe speed	**Operations staff**	**Operations staff** (partly supervised by system)	**Systems**	**Systems**	**Systems**
Driving	Control acceleration and braking	**Operations staff**	**Operations staff**	**Systems**	**Systems**	**Systems**
Supervising guideway	Prevent collision with obstacles	**Operations staff**	**Operations staff**	**Operations staff**	**Systems**	**Systems**
	Prevent collision with persons	**Operations staff**	**Operations staff**	**Operations staff**	**Systems**	**Systems**
Supervising passenger transfer	Control passenger doors	**Operations staff**	**Operations staff**	**Operations staff**	**Operations staff** or **Systems**	**Systems**
	Prevent injuries to persons between cars or between platform and train	**Operations staff**	**Operations staff**	**Operations staff**	**Operations staff** or **Systems**	**Systems**
	Ensure safe starting conditions	**Operations staff**	**Operations staff**	**Operations staff**	**Operations staff** or **Systems**	**Systems**
Operating a train	Put in or take out of operation	**Operations staff**	**Operations staff**	**Operations staff**	**Operations staff**	**Systems**
	Supervise the status of the train	**Operations staff**	**Operations staff**	**Operations staff**	**Operations staff**	**Systems**
Ensuring detection and management of emergency situations	Perform train diagnostic, detect fire/smoke and detect derailment, handle emergency situations (call/evacuation, supervision)	**Operations staff**	**Operations staff**	**Operations staff**	**Operations staff**	**Systems**

Figure 15 Grades of automation from IEC 62267

From the figure above, we can highlight the following elements (and differences with vehicle automation):

- GoA is applicable to a railway System of System (including train borne and wayside levels). This means that the attribution of GoA also relies on other features (Automatic Train Protection and Operation, known as ATP and ATO, are required for some GoA) and processes (supervision by humans in supervision centers);
- There is already a first dimension of technical implementation and operational process in this scale of automation ➔ Going from On-Sight GoA 0 to Non-Automated GoA 1 implies the use of ATP equipment;

- The guided nature of these systems simplifies some aspects of the automation of train operation, compared to the more open nature of operation for automated vehicles.

Thus, it appears that identifying the basic functions for a system operation is mandatory in order to build a relevant framework for shared responsibilities between human operators and AI-based systems.

F.2.3 Need for Domain-specific automation levels

The examples of automotive and railways also highlight the need for domain-specific automation levels framework.

While the global objective of building a domain-specific classification remains the same, each domain has its own operating challenges and specific human-system collaboration in existing workflows. Building such a framework is also ideal to study human-based reference systems and the impact of automation on such systems.

Establishing a proper automation levels framework is key to properly define the role that an AI-based system will play in a given feature.

F.3 AI component automation levels: appropriate expectations

F.3.1 Introduction to EASA AI Levels for AI components

In their Concept Paper for AI (*EASA Concept Paper: First Usable Guidance for Level 1 & 2 Machine Learning Applications*, 2023), the European Authority for Aviation Safety (EASA) framework of classification of AI levels **for AI-based components** (implemented in systems to achieve a given feature), mainly destined to aeronautics.

This framework is **high-level** and **component-centric** (hardware and software)**.**

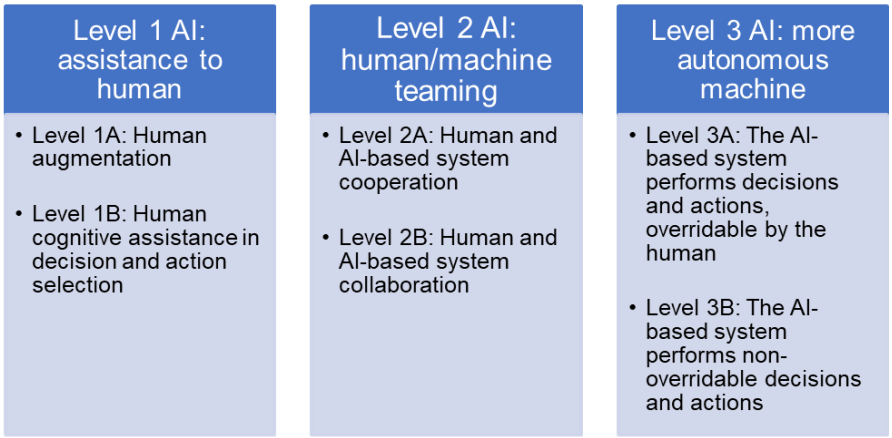


Figure 16 EASA proposal for AI levels

F.3.2 EASA AI levels and Sense Plan Act

The concept paper and suggested AI level framework from EASA detail how each level impact the existing human based process and the achieved human-AI interactions compared to existing ones.

This highlights the need to perform thorough impact analyses of AI as a whole before considering the implementation in a given system.

There is also a notion of end user authority on the AI system, that results from each level: the higher the level of AI, the less control the end user has on the AI-based system.

A a vision of AI-based Components that relies on the **Sense – Plan – Act** paradigm can be deduced from the EASA framework, that we illustrate in the following table.

EASA AI Levels	Information Analysis and Acquisition (Sense)	Decision (Plan)	Action Implementation (Act)
1A Human Augmentation	X		
1B Human Assistance	X	X	
2A Human-AI cooperation	X	X	X
2B Human-AI collaboration	X	X	X
3A Supervised advanced automation	X	X	X
3B Autonomous AI	X	X	X

F.3.3 Example of classification for Advanced Emergency Braking (AEB) system

The AEB system in automotive could be considered as an EASA Level 3A AI Component. However, the feature enabling Emergency Braking that is implemented by the AEB System is currently considered as SAE Level 0 according to SAE J3016: *"Active safety systems, such as electronic stability control (ESC) and automatic emergency braking (AEB) (…), are excluded from the scope of this driving automation taxonomy because they do not perform part or all of the DDT on a sustained basis, but rather provide momentary intervention during potentially hazardous situations".*

This is due to the SAE taxonomy considering the classification at an operational level, while the scope of EASA classification is reduced to an AI Component without context.

This example highlights that the EASA classification is not sufficient to cover the operational aspects of a given AI-based feature, since they are intrinsically generic. Different abstraction levels and specifics of each classification system may coexist, but they must be considered with regard to their intended use. Indeed, SAE J3016 deals with design-specific functionalities (features) in automotive while the EASA concept paper deals with AI-based systems and components, with a clear orientation towards aeronautics.

F.4 Steps for characterization of AI-based systems using classification levels

F.4.1 Defining shared responsibilities at feature level, via the Operational Specification

The previous section highlights the precaution that comes with trying to synchronize various frameworks of AI classification. An article from Digital Avionics Systems Conference (Schweiger et al., 2021) tried to match the classifications from automotive (SAE J3016), EASA and Human Factors. However, several examples can be found that challenge this matching. Indeed, we consider that a SAE Level 0 Feature (like AEB) could rely on EASA 3A Level AI-based component, so there is no strict match between classifications.

It also reinforces the need to act in order:

- First focus on the definition of the feature and its intended purpose in the Operational Approach, with an operational view of human and system shared responsibilities,
- Then, define the AI-based components implementing the feature and their associated EASA AI level required to implement said features, with shared responsibilities defined at system/component level.

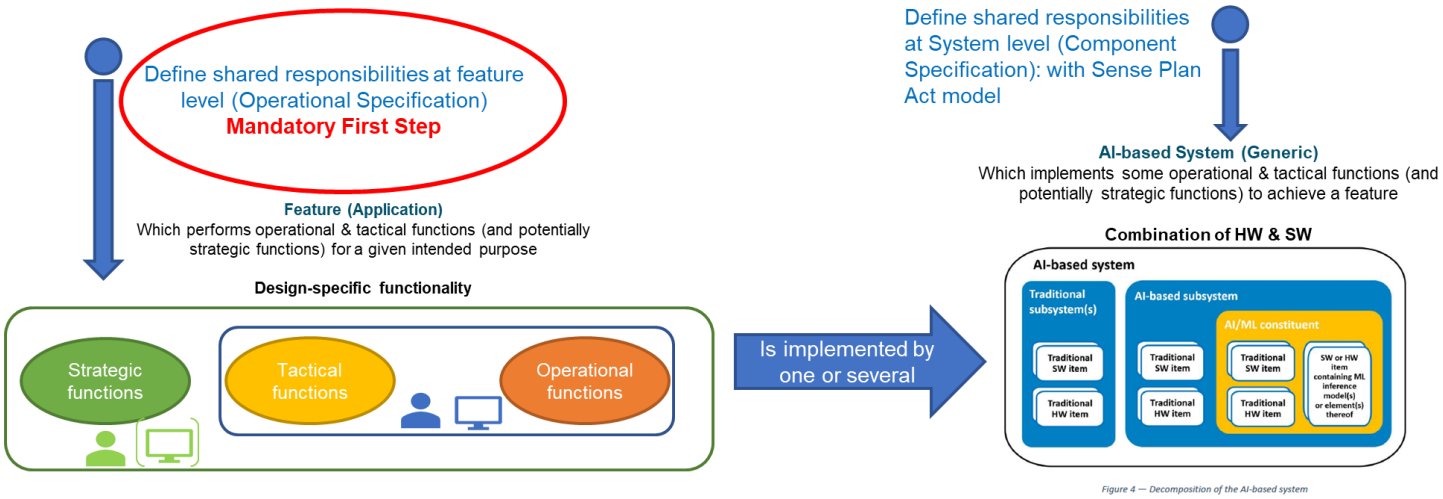


Figure 17 Proposal on how to consider shared responsibilities for automation of systems

F.4.2 Defining Human – System shared responsibilities using an appropriate scale

To define a proper framework of shared responsibilities enabling the automation of a given system, the following questions can help to guide the thought process

- Is the system already supporting some form of automation?
- Is the system implementing a feature already performing parts or all of the basic operational, tactical or strategic functions?
- Can some of these functions be achieved by using AI-based technologies only? (which ones and how?)
- What are the objects and events of the environment that need a response, i.e. what is the corresponding OEDR? How do we ensure that OEDR covers the expressed needs?

To build such framework, we need to keep in mind that the following elements are:

- A reference system, preferably human-based in order to establish a scale if there is none yet. It helps to highlight the basic functions mandatory for the system operation;
- The Intended Purpose of the System feature.

The resulting levels of shared responsibilities between human and system operators may rely on potential processes, technological advances or potential implementations

We can consider two kinds of AI/automation classification:

- A specific automation classification considering the specifics of some system features: a domain specific classification (beyond automotive and railways;
- A generic AI classification oriented towards system/component (the EASA classification could serve as a basis).

F.5 Conclusion on Part 5

Responsibilities boundaries between system and users have to be considered as a transversal viewpoint that could highlight all the five other parts of the document.

A clear vision of what operating the system means is required before defining a potential automation level with the corresponding repartition of responsibilities between the system and the human operator. The use of a reference system can help in both of these design phases. From non-automated or assistance reference systems, we can extract existing process based on human operator that help to identify the main functions and processes performed by humans and that can be potential candidates for (higher) automation.

The design choice shall guarantee, at operational level, that the mental model of the User is aligned with the system capabilities to ensure a proper assistance or even collaboration.

A framework for automation levels such as the one in SAE J3016 gives a view that is easy to grasp for end users and designers. It helps designers to identify easily where its new system may be positioned and what are the major consequences regarding human-system collaboration.

An incoming challenge posed by human-AI collaboration is regarding the new role of human operator: how do we enable a proper adjustment of the human operator without being over-reliant on AI execution?

G. Part 6 – Specific deadlocks for operational design of AI-based systems

Introduction to part 6

From the previous section related to shared commonalities between System Engineering and AI Engineering, we have identified the similarities as well as the specifics to consider for AI-based systems, at a process level. At a Specification level, however, there are specific deadlocks for the operational specification that must be anticipated and worked on to avoid typical pitfalls currently met in the domain: AI agents pursuing goals that are different from the user's goals, etc. These pitfalls are currently known in the AI field, but they still have to be properly formalized. We consider that the aforementioned issues can be avoided/mitigated by getting an overview of the Operational Specification of the AI-based system, rather than staying at the level of the ML Component. Indeed, working on the ML component alone is not enough to solve the specification issues of an AI-based system, as we may miss out on operational elements that could be captured by the Operational Specification of the AI-based system.

Overview of part 6

In this section:

- We revisit some incentives of implementing AI that also appear to be specific deadlocks for the operational specification of AI-based systems.
- We highlight how those deadlocks introduce two kinds of issue for AI specification: Specification Gaming and Goal Mis Generalization. Through these two issues, we analyze the best practices to implement and pitfalls to avoid in order to ensure that the Operational Specification properly reflect the expected Intended Purpose of the AI-based system.

G.1 Specific deadlocks to be encountered with AI operational specification

G.1.1 Dimensionality of inputs

Conventional programming addresses a combination of inputs and outputs that is limited and remains affordable for human designers. If the subject seems too complex, it can be split as many times as necessary in order to be manageable.

AI components allow to address issues where the inputs are more complex (like Computer Vision) and where combination of inputs and outputs becomes exponential. It is quite impossible to describe each one but it is necessary to define intended behaviors. Methodology has to be defined to solve this contradiction without learning each individual situation.

Thus, for AI (Machine Learning & Deep Learning), it is necessary to identify each typology of operational situation that has to be described to address the expected intended behaviors.

Example: at Operational level, an AI-based system for pedestrian detection has to consider a wide variety of entities that qualify as a pedestrian, due to several factors, such as the size of the pedestrian, the color of its clothes, its mean of locomotion, etc.

G.1.2 Multiple Intended behavior acceptable for a given operational situation

Conventional programming is generally considered bijective: for a single given situation corresponds a single given behavior ("bijective" behavior).

That is not the case for AI, where:

- several situations could lead to the same behavior ("surjective" behavior)
- Or a single situation could lead to several possible behaviors

This level of flexibility can help in reproducing human behaviors where multiple solutions are acceptable and possible.

Operational Specification aims to identify such situation and to specify the expected situations and the ones which must be avoided. Human designers may be able to answer to local situations taken individually. However, the results of the combination of those situations need a tradeoff at Operational Level, since the description of intended behavior becomes fuzzy.

The combination of each solution answering a local situation is null, when considering the intersection of some local solutions.

Example 1: in automated driving, an automated vehicle driving behind a truck at low speeds can either stays in the lane or proceeds with overtaking if safety conditions are met.

Example 2: an automated vehicle arrives at a crosswalk with an incoming pedestrian. It can either slow down to give time to the pedestrian to cross the road without stopping, or comes to a stop just before the crosswalk. Both behaviors are deemed acceptable in this situation.

G.1.3 Ability of generalization of AI: Ability of the system to treat situations never encountered before (and not explicitly specified)

In Conventional programming, programs have little resilience to new/unexpected situations. New situations encountered by the system are either within the expected design range or out of it. New situations out of design range can systematically lead to unexpected behaviors and they are taken into consideration when design and implementation are updated.

Whereas in AI, we know the AI agent is at least likely to provide an answer, although not satisfactory. AI may struggle at first to learn from unexpected events, but we know that this can be improved through training. The advantage here is that we do not have to specify every variant of a given situation in order for the AI component to provide a solution based on its learning.

We expect AI to help in covering situations not encountered before, and thus, that are unspecified. Yet nowadays, while they can perform well on data statistically similar to training data, they can still be thrown off by completely unexpected events. The goal here is to distinguish the typology of answer that the

system will be able to guarantee: from an operational viewpoint, what are the correlations that appear to be relevant (as well as those that are not)? The same method shall be pursued at the system level.

G.1.4 Difficulty to specify automation issues

Functional Specification shall describe the whole functionalities of the system. It implies to decompose the Intended Purpose into subfunctions and subfunctions into elementary tasks. Tasks allocated to the system have to be described whereas tasks allocated to humans are not detailed, with only the interfaces being described.

The intrinsic philosophy of automation is to move the boundaries between the actions performed by the system and those performed by the user.

- For conventional systems, expected behavior is limited to capabilities offered by the system to the user;
- For autonomous systems, expected behavior of the system goes further by including functions usually performed by human users (functional capabilities) and the way to perform it (prestation).

On the Functional side, we need to decompose and translate into simple functions the actions intuitively performed by human operators, so that they could be achieved by a system.

On the Prestation side, we need to consider that there is variability in the expected behaviors for given situations. It is quite difficult to describe precisely this kind of elements.

The combination of these two topics that adds complexity to the operational design of an AI-based system

- On the positive side, the generalization capabilities of AI are tailored to provide answers regarding automation topics;
- However, at its current level of maturity, AI technology is exposed to undesired emergent behaviors.

G.1.5 Operational specification for AI-based systems as a tool to prevent misalignment

Incentives to use AI-based systems instead of Conventional Software Systems lead to specific deadlocks. In the following sections, we provide elements that give a better view on the related challenges.

In particular, we describe the potential consequences of underestimating the operational viewpoint through: **Specification Gaming** and **Goal Mis Generalization (GMG)**.

We also provide leads on how to avoid those misalignments in the Operational Specification for AI-based systems.

G.2 Misalignment between specifications and risks at operational levels

G.2.1 Misalignment between specifications

Gaps are to be expected between Intended Purpose and its associated implementation. When such gap appears, several issues may manifest.

There are two types of misalignment that illustrate specification issues for AI:

- **Outer misalignment** which consists in the use of loopholes to satisfy part of the Intended Purpose➔ This is **Specification Gaming;**
- **Inner Misalignment** which illustrates the discrepancy between Intended Purpose and Agent Objectives ➔ This is **Goal Mis Generalization (GMG).**

We detail in the following sections these two issues in specification.

G.2.2 Goal Mis Generalization (GMG)

G.2.2.1 Introduction to GMG

DeepMind defines Goal Mis Generalization (GMG) as "an **instance of mis generalization** in which **a system's capabilities generalize** but **its goal does not generalize** as desired. When this happens, the system competently pursues the wrong goal." (Shah et al., 2022).

An additional definition in this article refers to a representation of the AI-based system *"pathological behavior, in which a learned model behaves as though it is optimizing an unintended goal, despite receiving correct feedback during training (…)"*.

GMG is considered as part of Robustness problems: *"In contrast, in goal mis generalization, the behavior must remain coherent. Some of the system's competencies remain intact, allowing it to coherently pursue a mis generalized goal"*. We can consider that system's competencies remain coherent in a certain given context, but not at all time nor in all conditions.

From the information above, it can be understood that GMG**:**

- Is caused by an under-specification of the AI-based system;
- Represents a pitfall for designers that may be initially tricked by the behavior of the AI agent on its capabilities to perform actions competently despite aiming for an undesired goal;
- Could range from small inconveniences (a bug) to safety failures in critical systems, if not foreseen early on.

G.2.2.2 Examples of GMG

GMG definition is not only restricted to Reinforcement Learning and could be extended to AI as a whole.

GMG illustrates that while the definition of a proper Intended Purpose for the AI-based system is essential, it is not sufficient to guarantee that the implemented AI system pursues this Purpose exactly.

We can find some examples of GMG as presented by DeepMind (DeepMind Safety Research, 2023) in Figure 18 below:

Example	Intended goal	Misgeneralised goal	Capabilities
Spheres	Traverse spheres in the correct order	Follow the red bot	Traversing the environment. Following an agent
Tree Gridworld	Chop trees sustainably	Chop trees as fast as possible	Chopping trees at a given speed
Evaluating Expressions	Compute expressions with minimal user interaction	Always ask questions before computing expression	Querying the user. Performing arithmetic

Goal misgeneralisation ingredient	Example: Spheres
1. Train a system with a correct specification.	Run deep reinforcement learning (RL), rewarding the agent for visiting spheres in the correct order.
2. The system only sees specification values on the training data.	The agent only sees that Trajectory 1 is +3 reward and Trajectory 2 is -2 reward.
3. The system learns a policy...	The agent learns to follow the red blob...
4. ...which is consistent with the specification on the training distribution.	...which indeed produces high-reward Trajectory 1 instead of low-reward Trajectory 2.
5. Under a distribution shift...	When you replace the expert bot with an anti-expert bot...
6. ...the policy pursues an undesired goal.	...the agent follows the anti-expert and accumulates negative reward.

Figure 18 Examples of GMG from DeepMind

G.2.2.3 Definition of Objective function and GMG

Depending of the type of ML context, an Objective function is rather defined as a reward. In this case, we aim to maximize the reward rather than reducing the error rate.

The Objective function is considered a "proxy" for System Designers used to translate their expectations on what the AI-based system should do. This Objective function can be seen as a proxy to translate the notion of Intended Purpose.

The CSET Issue Brief (Tim G. J. Rudner & Helen Toner, 2021) showcases examples of proxy for ML applications (in regards to their intention) that we synthesized below.

- AI agent for image classification:
  - **Intention**: *"find a model that classifies any image correctly";*
  - **Objective function (the proxy):** *"find a model that misclassifies the smallest number of images in a given set of training image–label pairs".*
- AI agent for Natural Language Processing (NLP):
  - **Intention:** *"find a model that gives a sensible response to any text prompt";*
  - **Objective function (the proxy):** *"find a model that predicts which word comes next in a text".*

We remark that the gap between the Intention and the proxy that is the Objective function leads to gaps that can then result in specification issues.

However, if we take this work from CSET in the light of the Intended Purpose of the AI Act, it reveals the need to recontextualize the AI application at system level in order to have the operational context and allocate adequately requirements to the AI module.

G.2.3 Specification Gaming

G.2.3.1 Definition of Specification Gaming

Specification Gaming is defined by (Tim G. J. Rudner & Helen Toner, 2021) as *"a particular failure mode in specification that can occur after an objective function has been specified by a human designer. It refers to a phenomenon where machine learning algorithms "game" whatever specification they were given,*

finding ways to achieve the specified objective with techniques that are totally disconnected from what the operator wanted".

- We consider it is plausible that Specification Gaming can result from Functional insufficiencies (insufficiency of specification or performance limitation) from the SOTIF ISO 21448:2021 standard (*ISO 21448*, 2022).

The team at DeepMind goes further by describing Specification Gaming as *"a behavior that **satisfies the literal specification** of an objective **without achieving the intended outcome**."* (Victoria Krakovna et al., 2020)

- This can be translated in the following situation where the objective function is met while the Intended Purpose (designer intent) is not;
- The AI agent should rather follow correlations that are more easily accessible in situations when the specification is insufficient to properly translate the intended purpose;
- Ex: An AI agent that classifies photos between male and female workers only by using the presence or absence of ties on the pictures.

G.2.3.2 Examples and nature of Specification Gaming

Specification Gaming can be simplified as an error in the specification of the objective in which the AI system will rather favor solutions easier to implement. The poorer the specification, the simpler it is for an AI system to "game" the specification, i.e. to find a loophole in this specification.

Specification Gaming can also manifest through bugs that use defects in the software.

Specification Gaming is not AI-specific and can even be found in living creatures (examples of dolphin wanting more fish reward, or even mice). Besides, it is not necessarily negative: it becomes problematic when an AI agent benefits from weaknesses in the specification to perform actions that are antagonistic with the Intended Purpose.

G.2.3.3 Risks related to specification of AI systems

(Amodei et al., 2016) written by some employees of Google Brain and OpenAI highlights several misbehaviors related to an inadequate specification of the objective function or of the expected behavior of the AI-based systems. These misbehaviors can be separated in three categories:

- **Specification of the wrong formal objective function:** *"the designer may have specified the wrong formal objective function, such that maximizing that objective function leads to harmful result"* (no matter the quality of the learning or data)
- **Specification of an objective function too expensive to evaluate frequently:** *"the designer may know the correct objective function, or at least have a method of evaluating it (for example explicitly consulting a human on a given situation), but it is too expensive to do so frequently, leading to possible harmful behavior caused by bad extrapolations from limited samples"*
- **Adoption of an undesirable behavior during the learning process:** *"the designer may have specified the correct formal objective, such that we would get the correct behavior were the system to have perfect beliefs, but something bad occurs due to making decisions from insufficient or poorly curated training data or an insufficiently expressive model."*

We can notice that these misbehaviors (called “safety problems” in the article) can be categorized as either Specification Gaming or Goal Mis Generalization.

For each of those safety problems, several causes (“research problems”) can be identified:

- **Safety Problem 1 - “Specification of the wrong formal objective function”:** can result from Negative side effects (the specified objective function ignores aspects of the environment that have a potential of harm) or Reward hacking (the specified objective function can be maximized in a way not intended by the designer, hence “Specification Gaming”)
- **Safety Problem 2 - “Specification of an objective function too expensive to evaluate frequently”:** is caused by a phenomenon called Scalable Oversight ➔ This is one is interesting because it could be Welding situation?
- **Safety Problem 3 - “Adoption of an undesirable behavior during the learning process”:** can result from unsafe and unredeemable actions from Reinforcement Learning actions (covered by the “Safe Exploration” research issue) or from a lack of robustness to a shift in distribution leading to unpredictable and overlooked wrong decisions (covered by the “Robustness to distributional shift” research issue)

The article gives interesting suggestions on how to deal with each of these “research problems” by using examples. While we suggest to keep in mind those proposals, they may not be treated at the same stages:

- Safety Problem 1 could be treated through the Operational Specification;
- However, Safety Problems 2 & 3 should be treated through the System Specification, as they are closer to the implementation of AI at component level.

G.2.3.4 Links with Emergent Behaviors

These misalignment issues for AI could be linked to a topic that becomes more widespread in the filed: Emergent Behaviors. These Emergent Behaviors are described as *“complex patterns or behaviors that arise spontaneously from the interactions of simpler elements or systems”* (Noura Elgendi, 2023) and do not result from an intended action of programming. They are not strictly limited to programming as they can also be observed in animal species.

The article showcases an example of AI-based systems with a Hide-and-Seek games where two teams start displaying Emergent Behaviors as the numbers of iterations and games increases. It shows how the AI agents are able to adapt and display complex strategies from a simple use case. AI-based systems displaying Emergent Behaviors can still achieve a solution through unexpected means, but these elements of uncertainty could hazardous depending on the use case.

We consider that Emergent Behaviors could be prevented as early as the start of Operational Specification, with the appropriate observation of the behaviors of an AI agent. As far as expected operational correlations could be described, unwanted correlations could be prevented.

G.3 Conclusion on Part 6

The issues of Specification Gaming and Goal Mis Generalization illustrate the concrete consequences that result from improperly raising a challenge to solve. In this context, these challenges can lead to undesired behaviors of AI-based systems and reveal potential insufficiencies of specification and loopholes. Defining appropriately the needs and Intended Purpose of such system is key to mitigate those side effects, but we also need to consider the constraints represented by the deadlocks evoked in section G.1 above.

Achieving an Operational Specification that is consistent with the Intended Purpose of the AI-based system seems mandatory. However, some challenges remain:

- How to describe complex environments?
- How to describe properly expected behaviors in operational specification to prevent goal misspecification?
- How does the synthesis of stakeholders' needs explain where the expected behavior is precisely defined and where it is more informative?

In synthesis, the previous points can be summarized by the following question: How can the Operational Specification become more resilient to specification insufficiencies in AI-based systems context?

We suggest to go forward to apply all the methodological guidelines mentioned in this document on system level use cases involving AI-based systems to have a cost-benefits evaluation of the methodology, and evaluate the gaps with usual AI methodologies, regarding the ability to achieve the Intended Purpose.

H. Conclusion on Operational Approach

AI-based systems are developed and used in the context of automating tasks previously performed by human beings. In this context, we shall define as soon as possible the expected level of automation and the resulting interactions between the AI-based system and its related actors. It is useful to define and characterize intended behavior, either at operational level, or for architectural and implementation choices.

We hope that this document illustrates the necessity to realize an Operational Specification for AI-based systems, and the different elements to consider in such specification to achieve the level of trustworthiness required for critical applications.

In this conclusion, we would like to put the emphasis on the following elements as a methodological guideline to achieve the Operational Specification:

- We expect to collect all stakeholder needs for the whole life cycle of the system. It requires a first vision of the system to allow them to express more and more precise needs in an iterative process. During this process, a synthesis of the needs is done, resulting from a convergence between stakeholders and including trade-offs between needs and system constraints.
- As for conventional systems, AI-based system design relies on reference systems which allows to converge more efficiently to the final design. It gives relevant information on operational context and technical feasibility, regarding expected implementation.
- The Operational Approach shall guarantee the identification of proper responsibilities/cooperation between human operators and the system. It implies that the stakeholders shall focus at system level for a broader view of the system context instead of considering the AI component viewpoint. This approach allows to join System Engineering operational approaches and the trend of major AI actors to put forward contextual system approaches in the preliminary activities of their AI development.

These recommendations regarding Operational Specification gives key elements to describe the Intended Purpose of the AI-based systems required by the AI Act, and increases the ability of the implemented system to answer this Intended Purpose. This is one dimension of Trustworthiness.

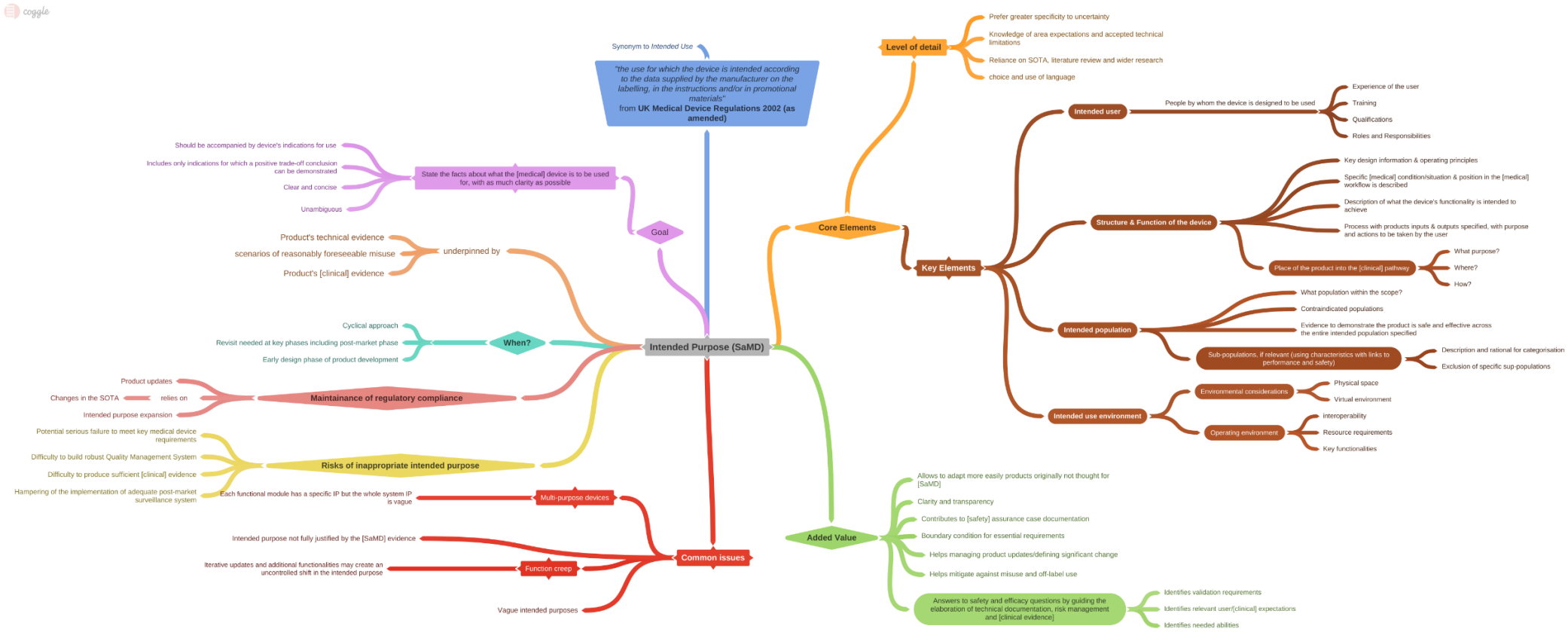
Regarding future works, we have identified two axes of improvement.

- First, this approach must be consolidated by working on relevant Use Cases at system level, that we did not get the opportunity to work upon.
- Secondly, we hope that this approach can evolve into a dedicated operational process once matured.

We covered in this deliverable the importance of developing an Operational Approach in AI-based systems. This approach must be completed by a System Approach, that will translate operational design into technical elements that can be used for implementation activities. This will be developed in the deliverable 218B.

I. Annex

I.1 Annex 1 Mind Map of Intended Purpose inspired from SaMD


Intended Purpose (SaMD)
Synonym to Intended Use
"the use for which the device is intended according to the data supplied by the manufacturer on the labelling, in the instructions and/or in promotional materials"
from UK Medical Device Regulations 2002 (as amended)
Goal
State the facts about what the [medical] device is to be used for, with as much clarity as possible
Should be accompanied by device's indications for use
Includes only indications for which a positive trade-off conclusion can be demonstrated
Clear and concise
Unambiguous
underpinned by
Product's technical evidence
scenarios of reasonably foreseeable misuse
Product's [clinical] evidence
When?
Cyclical approach
Revisit needed at key phases including post-market phase
Early design phase of product development
Maintainance of regulatory compliance
relies on
Changes in the SOTA
Product updates
Intended purpose expansion
Risks of inappropriate intended purpose
Potential serious failure to meet key medical device requirements
Difficulty to build robust Quality Management System
Difficulty to produce sufficient [clinical] evidence
Hampering of the implementation of adequate post-market surveillance system
Common issues
Multi-purpose devices
Each functional module has a specific IP but the whole system IP is vague
Function creep
Intended purpose not fully justified by the [SaMD] evidence
Iterative updates and additional functionalities may create an uncontrolled shift in the intended purpose
Vague intended purposes
Core Elements
Level of detail
Prefer greater specificity to uncertainty
Knowledge of area expectations and accepted technical limitations
Reliance on SOTA, literature review and wider research
choice and use of language
Key Elements
Intended user
People by whom the device is designed to be used
Experience of the user
Training
Qualifications
Roles and Responsibilities
Structure & Function of the device
Key design information & operating principles
Specific [medical] condition/situation & position in the [medical] workflow is described
Description of what the device's functionality is intended to achieve
Process with products inputs & outputs specified, with purpose and actions to be taken by the user
Place of the product into the [clinical] pathway
What purpose?
Where?
How?
Intended population
What population within the scope?
Contraindicated populations
Evidence to demonstrate the product is safe and effective across the entire intended population specified
Sub-populations, if relevant (using characteristics with links to performance and safety)
Description and rational for categorisation
Exclusion of specific sup-populations
Intended use environment
Environmental considerations
Physical space
Virtual environment
Operating environment
Interoperability
Resource requirements
Key functionalities
Added Value
Allows to adapt more easily products originally not thought for [SaMD]
Clarity and transparency
Contributes to [safety] assurance case documentation
Boundary condition for essential requirements
Helps managing product updates/defining significant change
Helps mitigate against misuse and off-label use
Answers to safety and efficacy questions by guiding the elaboration of technical documentation, risk management and [clinical evidence]
Identifies validation requirements
Identifies relevant user/[clinical] expectations
Identifies needed abilities

J. Bibliography

Alan Faisandier, Garry Roedler, & Rick Adcock. (2023, November 20). *SEBoK - Stakeholder Needs Definition* [Wiki]. https://sebokwiki.org/wiki/Stakeholder_Needs_Definition

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety* (arXiv:1606.06565). arXiv. https://doi.org/10.48550/arXiv.1606.06565

ARP6983 (WIP) Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI - SAE International. (2023). [WIP Standard]. https://www.sae.org/standards/content/arp6983/

REPORT on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, Report-A9-0188/2023, European Parliament (2023). https://www.europarl.europa.eu/doceo/document/A-9-2023-0188_EN.html

Confiance.AI EC6. (2021). *Use Case Specification for Renault Welding* (EC6 > Deliverables > Batch 1). EC6.

Confiance.AI EC6. (2022). *EC6.8_Methods_and_Tools_for_ODD* (EC6 Deliverables). Confiance.AI.

DeepMind Safety Research. (2023, March 24). Goal Misgeneralisation: Why Correct Specifications Aren't Enough For Correct Goals. *Medium*. https://deepmindsafetyresearch.medium.com/goal-misgeneralisation-why-correct-specifications-arent-enough-for-correct-goals-cf96ebc60924

EASA Concept Paper: First usable guidance for Level 1 & 2 machine learning applications (Concept Paper proposed iIssue 2; p. 242). (2023). EASA. https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-concept-paper-proposed-issue-2

Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE

ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, 52021PC0206, European Commission, COM/2021/206 final (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Guidance: Crafting an intended purpose in the context of software as a medical device (SaMD). (2023, March 22). [Government website]. GOV.UK. https://www.gov.uk/government/publications/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd

IEC 62267:2009 Railway applications—Automated urban guided transport (AUGT)—Safety requirements (62267:2009; 1.0). (2009). https://webstore.iec.ch/publication/6681

IEC 62290-1:2014 Railway applications—Urban guided transport management and command/control systems—Part 1: System principles and fundamental concepts (62290-1:2014; 2.0). (2014). https://webstore.iec.ch/publication/6777

Introducing ChatGPT. (2022, November 30). Openai.Com. https://openai.com/blog/chatgpt

ISO 21448:2022 Road vehicles Safety of the intended functionality (International Standard Published 21448:2022; Version 1). (2022). https://www.iso.org/fr/standard/77490.html

ISO 26262-1:2018—Road vehicles—Functional safety—Part 1: Vocabulary (Version 2). (2018). https://www.iso.org/standard/68383.html

ISO/IEC DIS 5338: Information technology - Artificial intelligence - AI system life cycle processes (Cobaz 5338; Version 1). (2023). https://cobaz.afnor.org/notice/NORME/XS142213/XS142213

ISO/IEC/IEEE 15288:2023 Systems and software engineering—System life cycle processes (15288). (2023). https://www.iso.org/standard/81702.html

ISO/IEC/IEEE 42020:2019 Software, systems and enterprise—Architecture processes (International Standard Published 42020; Version 1). (2019). https://www.iso.org/standard/68982.html

Kevin Forsberg, Richard Turner, & Rick Adcock. (2023, November 20). *SEBoK - Generic Life Cycle Model*. https://sebokwiki.org/wiki/Generic_Life_Cycle_Model

Mantissa, K., & Bohn, C. (2024). *Methodological Guideline for AI-based System Design at Operational and System Level: System Approach* [Methodological Guideline]. Confiance.AI.

Mercedes Benz. (n.d.). *DRIVE PILOT First Responder Interaction Plan*. https://www.mbusa.com/content/dam/mb-nafta/us/owners/drive-pilot/DRIVE_PILOT_First_Responder_Interaction_Plan.pdf

Noura Elgendi. (2023, March 27). *The Power and Perils of Emergent Behaviors in AI: What You Need to Know*. LinkedIn. https://www.linkedin.com/pulse/power-perils-emergent-behaviors-ai-what-you-need-know-elgendi-mba-1f

OpenAI Platform. (n.d.). [Documentation]. ChatGPT - Introduction. Retrieved 10 November 2023, from https://platform.openai.com

Pedro A. Ortega, Vishal Maini, & DeepMind Safety Research. (2018, September 27). Building safe artificial intelligence: Specification, robustness, and assurance. *Medium*. https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1

People + AI Guidebook. (2021). https://pair.withgoogle.com/guidebook

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 119 OJ L (2016). http://data.europa.eu/eli/reg/2016/679/oj/eng

Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.), 117 OJ L (2017). http://data.europa.eu/eli/reg/2017/745/oj/eng

Schweiger, A., Annighoefer, B., Reich, M., Regli, C., Moy, Y., Soodt, T., de Cacqueray, A., & Redon, R. (2021). Classification for Avionics Capabilities Enabled by Artificial Intelligence. *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, 1–10. https://doi.org/10.1109/DASC52595.2021.9594364

SEBoK. (2023, May 31). [Wiki]. Sebokwiki.Org. https://sebokwiki.org/wiki/Guide_to_the_Systems_Engineering_Body_of_Knowledge_(SEBoK)

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., & Kenton, Z. (2022). *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals* (arXiv:2210.01790). arXiv. https://doi.org/10.48550/arXiv.2210.01790

The HAX Toolkit Project. (n.d.). *Microsoft Research*. Retrieved 13 October 2023, from https://www.microsoft.com/en-us/research/project/hax-toolkit/

The Medical Devices Regulations 2002, Pub. L. No. 2002 No. 618 (2002). https://www.legislation.gov.uk/uksi/2002/618

Tim G. J. Rudner & Helen Toner. (2021). *Key Concepts in AI Safety: Specification in Machine Learning*. https://doi.org/10.51593/20210031

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, Shane Legg, & DeepMind Safety Research. (2020, April 21). *Specification gaming: The flip side of AI ingenuity* [Medium]. https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

Waymo. (2023). *Waymo Emergency Response Guide and Law Enforcement Interaction Protocol (October 2023).pdf*. https://storage.googleapis.com/waymo-uploads/files/documents/first-responders/Waymo%20Emergency%20Response%20Guide%20and%20Law%20Enforcement%20Interaction%20Protocol%20(October%202023).pdf

Title: Methodological Guideline for AI-based System Design at Operational and System Level: Operational Approach – 218A

Keywords: Operational Approach, Intended Purpose, Design Intent, Reference Systems, Operational Specification, Automation, Shared responsibilities

Our partners