EasyChair Preprint

№ 15861

Building Confidence: an Ontological Approach to Assurance of Safety-Critical Systems

Odd Ivar Haugen

February 23, 2025

Building Confidence: An Ontological Approach to Assurance of Safety-Critical Systems

Odd Ivar Haugen
Group Research and Development department, DNV AS
Trondheim, NORWAY
e-mail: odd.ivar.haugen@dnv.com.

***Abstract*—This paper explores the ontology of assurance in safety-critical systems, emphasising the importance of knowledge and confidence in system behaviour. Assurance is defined as providing grounds for justified confidence in system properties, such as safety and security. The paper discusses the main concepts of assurance, including system requirements, confidence, and justification. It discusses the CESM metamodel for understanding system behaviour and emergent properties. The paper also highlights the importance of objectivity in assessing the strength of knowledge and the role of verification in generating evidence as a part of the argumentation. The assurance case is presented as a systematic way to represent knowledge and support decision-making.**

***Index Terms*—Assurance, CESM metamodel, Confidence, Emergent properties, Knowledge, Objectivity, Risk, System behaviour, System safety**

A. Introduction

Engineering safety-critical systems require both rigour and precision in the entire system lifecycle, from the concept and design phase to the deployment and operation phase and finally to system retirement. The stakeholders, such as the system operator, the regulator, and society at large, need to be assured that the system behaves as intended, that is, is safe, reliable, secure, and, in general, behaves responsibly. What does it take to assure the different stakeholders so that they become confident that the system can be deployed? This paper explains the ontology of assurance, that is, the elements and their relationship so that the assurance effort becomes capable of providing the necessary confidence.

Assurance is defined as providing grounds for justified confidence that a claim has been or will be achieved [1]. The claim can be any proposition about the system's property, like safety, security, or, in the case of a system based on artificial intelligence (AI), is fair [2].

B. Main concepts of assurance

Assurance is about becoming confident that the system behaves in a way that is acceptable to the stakeholders. Here, stakeholders are seen as any person, group of persons, governmental regulator, society, or even the natural environment. In short, it is an entity that is affected by the behaviour of the system.

A claim is a property of interest about the system. The claims can be thought of as system requirements; that is, "this" is how the system should behave in order for the system to be accepted by the stakeholders. Analysing the previous statement reveals, as a first approach, the four principal criteria that must be in place to achieve acceptance:

1) the system requirements must reflect the interest of the stakeholders,
2) refining these requirements into technical specifications must maintain the essence of these requirements,
3) the system's adherence to these requirements must be secured and adequately substantiated,
4) 1, 2, and 3 must be communicated to the stakeholders or their representatives in such a way that they can make intelligible decisions.

It is clear from the above items that the key to system acceptance is *knowledge*. Indeed, knowledge may be said to be the "hub" of assurance. The stakeholders must know that the system behaves acceptable. Knowledge is a prerequisite for confidence, which reduces the uncertainty about the system.

Confidence is different from trust. Confidence is something that can be merited through demonstrating adequate capability; trust, however, has to be earned through time; that is, trust is closely connected to an agent's intention. This means that confidence can be merited through demonstrating adequate capability (technical system and responsible agent); trust must be earned through time by a responsible agent adhering to sound and recognised ethical principles.

As assurance is about providing grounds for justified confidence, this paper will therefore focus on how to demonstrate adequate system capability so that the stakeholders can make intelligible decisions based on their knowledge and, thereby, their level of confidence in the system.

It should be noted that assurance is an epistemic activity, while risk management encapsulates both epistemology and intervention in the real world [3].

The system capability is, in this context, equivalent to how the system behaves in normal operation and how it behaves in abnormal situations.

The system risk is defined as the "effect of uncertainty on objectives" [4] and reflects the consequence and uncertainty that the system causes a loss for stakeholders. The uncertainty is here divided into two types: epistemic and aleatory [5].

Item three in the above list requires that the system behaviour adherence to the requirements is substantiated; that

is, claims about the system must be substantiated through sound and relevant argumentation. For an argument to be sound, it must be generated in accordance with acknowledged methodologies using reliable tools and adequately skilled people.

To assess the soundness and the strength of arguments, an assessor not only needs to be a subject matter expert but also needs guidance about what can be regarded as acceptable methods and processes to develop arguments; that is, he needs guidance about the argument's *objectiveness*. A higher degree of objectivity increases the strength of the argument, which is necessary when the risk is high, such as for safety-critical systems.

C. Assurance and confidence - an overview

Confidence can be thought of, in statistical terms, as a quantitative measurement of uncertainty, e.g. an interval indicating the confidence that the value of a parameter is likely to fall within. However, confidence may also be thought of as a feeling that reflects the coherence of the information and the cognitive ease of processing it [6]. Assurance is defined as "grounds for justified confidence that a claim has been or will be achieved". The definition does not limit assurance to one or the other type of confidence; hence, assurance addresses both.

Both types of uncertainties pose challenges. The frequentist approach to quantifying uncertainty requires robust statistical data. Here lies a few major obstacles, some of which are: the inherent complexity of many safety-critical systems, the novelty of the technology, statistically significant data from rare events, and assigning probabilities to inherently social aspects.

The second type of confidence also poses challenges. We cannot base decisions concerning the safety and well-being of stakeholders and society on pure feelings but on strong knowledge based on facts and trustworthy evidence.

Therefore, assurance may provide grounds for justified confidence through uncertainty quantification only if based on robust statistics, that is, knowledge about properties of the statistical distribution of the parameter in question, and/or judgemental assessments only if based on sound argument substantiating the truthfulness of the claim.

Therefore, the immediate goal, or primary effect of assurance, is to generate knowledge, knowledge to decrease or establish the uncertainty about a claim, addressing both types of uncertainty when appropriate. A Functional Analysis System Technique diagram (FAST diagram) illustrates the relation between assurance, knowledge and confidence (Figure 1). As

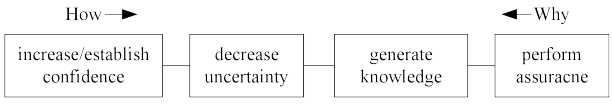

Fig. 1. FAST diagram connecting knowledge to confidence

knowledge is the "hub" of assurance, knowledge must be treated systematically and expressed explicitly to enable it to be rigorously scrutinised. This is to avoid that confidence being based on unsubstantiated feelings and pure guesswork. The assurance case is a systematic and explicit way of representing and treating knowledge.

As safety is an emergent property [7], the knowledge about the truthfulness of the claim must address all system aspects that affect emergence. Elements necessary in analysing emergent behaviour in engineered socio-technical systems are encapsulated in the systems approach.

Intuitively, the higher the risk that the system poses to stakeholders, the higher confidence we need that it will indeed behave as expected. As knowledge decreases uncertainty and increases confidence, we need a way to assess the strength of knowledge. Assessing the strength of knowledge is key to adjusting the assurance effort to risk level. Figure 2 depicts how the different items discussed above are connected.

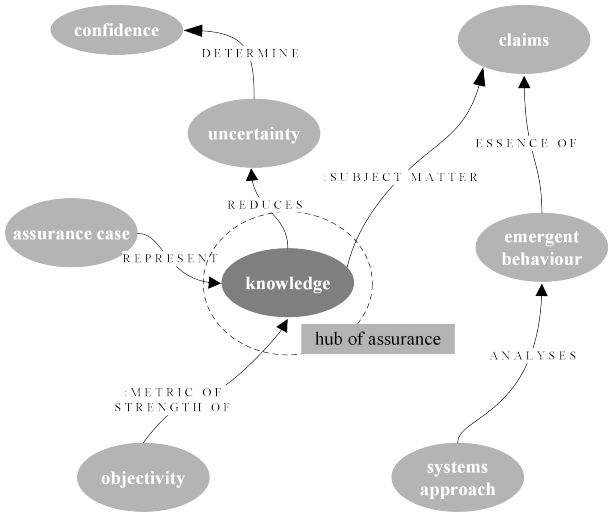

Fig. 2. Overview of the ontology of assurance

D. Assurance and system risk

Figures 1 and 2 showed how knowledge generated in the assurance effort reduces uncertainty, and that uncertainty determines confidence. Moreover, as earlier established, uncertainty is one part of the risk concept. Hence, assurance and risk are connected through uncertainty (Figure 3). There is, however, another connection in addition to the one mentioned above. In the top right corner of Figure 2, it is indicated that the subject matter of the knowledge is the claim. Claims are statements about system properties that address the system requirements elicited by stakeholders and their concerns and objectives (Figure 4). Stakeholders are generally risk avert [6] and are concerned about the consequences of losses. They need adequate confidence that potential losses are acceptable. Assurance addresses these concerns by generating knowledge about the truthfulness of the claims made about the system properties.

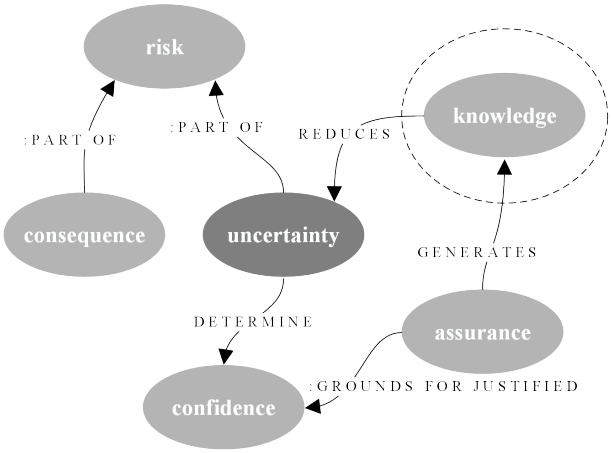

Fig. 3. Assurance is connected to risk through uncertainty

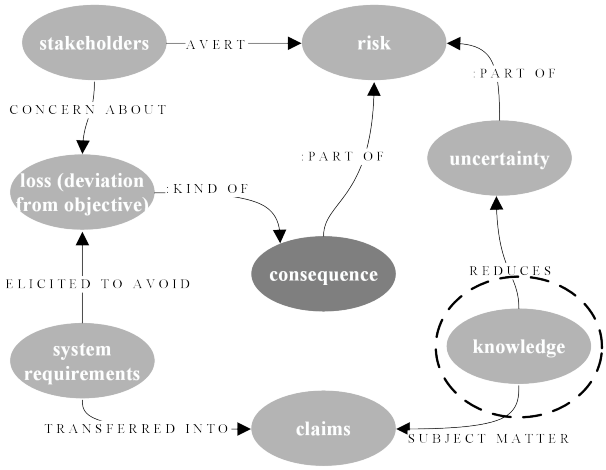

Fig. 4. Assurance is connected to risk through claims

E. Assurance and the Systems Appoach

A way to understand and analyse complex systems and emergence, is to model the system behaviour in terms of its composition, structure, mechanisms and the environment in which it operates. These system aspects are termed the CESM metamodel [8]:

- **Composition** (C): Collection of all the parts or objects in the system.
- **Environment** (E): Systems outside (excluded from) the target system, but act upon, or are acted upon by, the target system.
- **Structure** (S): The relationships and bonds among the system agents and between the system agents and the environment.
- **Mechanisms** (M): The processes that make the system behave in the way that it does.

The emergent behaviour becomes a function of the above elements; that is, any system s may be modelled, at any given instance, as the quadruple: μ(s) = <C(s), E(s), S(s), M(s)>. As μ(s) is an emergent property, and emergent properties exist on different levels of abstraction (LoA) [9], the CESM must also be instantiated at these LoAs.

This can be visualised by the system triangle (Figure 5) where the corner of the triangle illustrates "CSM" encapsulated by "E". The "system" in the middle represents μ(s). μ(s) emerges, therefore, as a result of the conceptual interaction between the corners of the triangle, but also between the triangle and the environment (E). To move the analysis between the LoAs, a rule-based gradient is used, termed the gradient of abstraction (GoA).

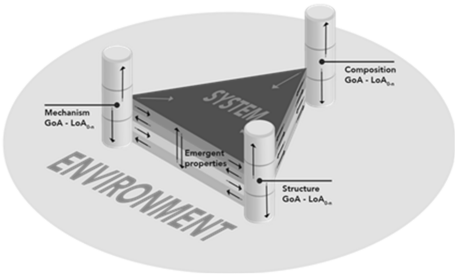

Fig. 5. The CESM triangle showing

For each element in the CESM metamodel, we can assign different system model categories [10]:

- Composition: **Object model** representing the system elements and components and their ontological relationship to each other.
- Environment: Also modelled as a system containing all aspects of the CESM metamodel, which means that the environment must be represented by models representing the composition, structure and mechanisms (our target system is part of the environment of its environment).
- Structure: **Agent model** includes entities such as controllers, actuators, sensors, humans, and subsystems. The agent concept includes authority, responsibility, goals, concerns, motivation, and wishes (humans).
- Mechanisms: **Function model** represents the operations that must be performed (by the agents) to achieve goals.

Examples of system model instantiation of the agent model is the control structure known from Systems-Theoretic Process Analysis (STPA) [7]. Another agent model may focus more on the agent's goals, motivation, concerns and wishes, like a model used in a stakeholder analysis where social and business aspects are emphasised.

A function model may focus on the preconditions, resources, and timing for achieving it, like the model used in the Functional Resonance Analysis Method (FRAM) [11].

The functional dependencies between functions, like in the Functional Analysis System Technique (FAST) [12] may be used as GoA to move the analysis between abstraction levels, that is, to represent the system at different LoAs [13].

The systems approach described above used in assurance, can be summarised by the following statements [3]:

- The conceptual interaction between the system composition (C), environment (E), structure (S), and mechanisms (M) models the system behaviour.

- The kind of, and number of levels of abstractions (LOAs) used in the modelling is determined by the knowledge sought through the assurance effort.
- The systems approach is used in every aspect of the assurance effort, such as: system description, describing the system boundary, describing the environment in which the system is operating, system analysis, verification and validation, and elicitation of system requirements.

Figure 6 depicts the relationship between the systems approach and assurance. The system safety requirements are formulated at safety claims. As safety is an emergent property that emerges through the interaction between the system entities, it can be modelled through the CESM metamodel.

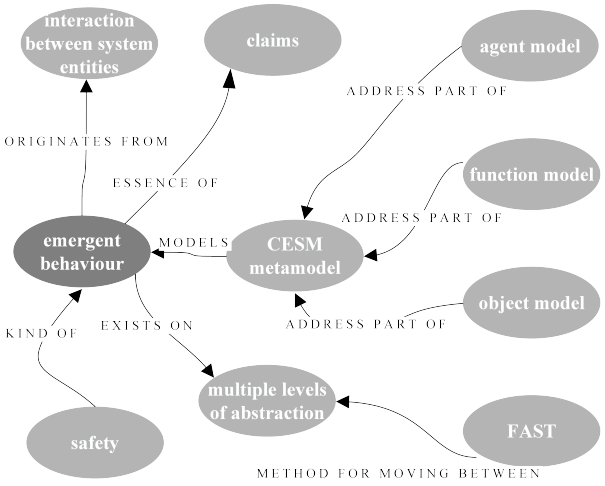

Fig. 6. The systems approach is connected to assurance through the claims or requirements

F. Assurance, epistemology and justification

Recall that the concept of risk incorporates, in addition to the consequence, two kinds of uncertainties: epistemic and aleatory. Knowledge reduces epistemic uncertainty. If the risk is high, like for safety-critical systems, the argument supporting the claim must be strong. The strength of the argument and, thereby, the strength of the knowledge reduces the epistemic uncertainty and, thereby, the risk[1].

Although knowledge is not easily defined[2], it must be linked to accessible facts about the subject matter. Moreover, building confidence through knowledge requires, not only apparently truthful propositions (claims), but also that the reasoning is sound, relevant and adequate; that the proposition is justified: *"Someone who is very confident but for the wrong reasons would also fail to have knowledge"* [14]. The reason for believing that a proposition represents the truth must be justified.

Justification may be thought of as an argument for why we hold certain beliefs or why we think those beliefs are reasonable and true. These justifications may be under the law or before God. However, in the context of assurance, justifying beliefs must be based on knowledge, or in other words, be based on epistemic justification [15]. (A safety-critical system needs, of course, to conform to laws and regulations; however, the point is that the justification must be based on knowledge).

Assurance seeks epistemic justification to establish if a proposition can be turned into a belief, that is, belief through warranted propositions.

Belief revision is the process of changing beliefs based on new data [16]. It is important to emphasise that good reasoning is no guarantee of truth. Seeking the truth and believing to have found it using sound methods and reasoning is no guarantee to actually have found it.

Justifying a proposition may, in principle, entail an infinite chain of justifications (infinitism): The justification of the justification of the justification... This is, of course, unacceptable. The question, then, is when to stop this chain of justifications.

One strategy is to continue until the supporting justifications become self-evident, that is, propositions that do not need further justification (foundationalism). This kind of justification results in a hierarchy of propositions, and the "bottom" of this hierarchy consists of fundamental propositions, that is, self-justified propositions.

Alternatively, we may ensure that the propositions support each other, that is, the propositions are coherent (coherentism). With this strategy, there are no fundamental propositions. Critiques claim that this strategy can lead to circular argumentation [15].

A reasonable approach is to combine the two strategies, that is, ensuring coherence within the set of propositions and justification, and stopping the chain of justification when reaching a self-justified proposition.

In practice, one may not reach a self-evident fundamental level for several reasons. One reason may be that there is a dispute about whether such a level is actually reached[3]; another reason may be that continuing the chain of justification requires disproportionate resources. Therefore, there may be residual uncertainty as to whether a proposition represents the truth.

Other sources of uncertainty are that there may exist evidence that weakens the proposition, or there may be a lack of available evidence. Moreover, other obstacles may

[1]As risk consists of three parts, risk can, of course, be reduced by altering the system design or the operational condition. These risk reducing strategies affect the consequence and/or the aleatory uncertainty. However, as this paper is concerned with assurance, which is an epistemic endeavour, these two strategies are not further discussed, but are left for a future paper on the relationship between assurance and risk management.

[2]The classic definition is: "Justified true belief".

[3]Showing compliance towards an international industry standard is often regarded as such a self-justified belief, that is, providing evidence that a system complies with such a standard is often regarded as adequate believing a proposition that e.g. a system is reliable, fair, safe and secure. An international standard should reflect good industry practice. However, e.g. artificial intelligence (AI) is a novel technology that even if there exist a relevant international standard, it may not be regarded as self-justified because the standard itself does not necessary reflect any industry practice (because there do not exist any such practice), or at least the practice may be inadequate. This means that it might be necessary to continue the justification chain further when assuring novel complex systems, e.g. based on AI.

hinder the generation of additional evidence, such as technical limitations, ethical concerns, lack of statistical data, or other practical causes.

There is no universal uncertainty threshold for when an agent will accept a proposition and when he rejects it. Moreover, given a justification of a proposition, there is no universal law governing the level of uncertainty an agent will feel about its truthfulness.

Belief revision depends not only on the properties of the justification of the proposition but also on the agent's epistemic state, that is, the agent's required rationality to turn a proposition into a belief, prior belief and any other properties important for the agent to represent facts about the world.

The uncertainty threshold for an agent's belief revision also depends on aspects such as the risk (perceived and/or actual) of accepting or rejecting a proposition (including being indifferent). Moreover, an agent's level of uncertainty, given a justification of a proposition, depends not only on the strength of the justification, but also on aspects such as the degree of being susceptible to cognitive biases[4] [6] and rhetoric. Obviously, we should strive to minimise aspects of belief revision that are unrelated to the properties of the justification.

An agent's prior beliefs cannot, and should not, be controlled and cannot be totally known. Nevertheless, prior belief is central to belief revision. Data-oriented Belief Revision (DBR) [17] (simplified illustration in Figure 7) is a model of belief revision that can illustrate the role of prior belief in belief revision.

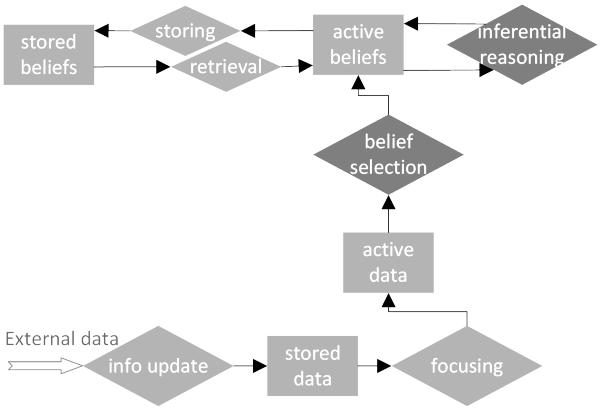

Fig. 7. Simplified epistemic processing in DBR [17]

After new data is available about a proposition (External data), the data is assessed to determine their relevance and strength, possibly forming a new or updated belief set, termed *belief selection* in Figure 7. This process regulates the interaction between data and beliefs, what to believe in, and with what strength.

As belief revision is tightly connected to the agent's prior beliefs and possible degrees of cognitive biases, we cannot assess the epistemic strength of the justification by appealing to the agent's prior beliefs, or what seems to be "very reasonable" and the like. What seems reasonable is an internal feeling in each agent and is largely based on his current epistemic state.

Instead, the agent needs to be nudged towards sound rationality of assessing uncertainty using a more comprehensive framework of thinking about the level of uncertainty (epistemic strength of the justification), without being forced into an epistemic strait jacket of predefined categories of epistemic levels.

We want to decrease uncertainty as the risk of accepting a false proposition increases. The opposite may not be so obvious, that we also want to decrease uncertainty when risk increases by rejecting a true proposition. Accepting a false proposition, on the one hand, or rejecting a true proposition, on the other, represents assurance risk[5].

Decreasing uncertainty to the point of accepting a proposition, or in other words, revising one's belief, can be achieved by both strengthening the justification that the proposition is true, and/or by increasing effort in seeking justification that the proposition is false without finding such justification. Sometimes, the only way to justify a proposition p is to find a strong justification that $\neg p$ is not the case[6].

A way to accommodate for proper assessment that knowledge is built on epistemic justification is through argumentation. While belief revision describes how we should update our beliefs, argumentation is a way to make belief revision occur. "The two concepts are two sides of the same epistemic coin" [17], [18].

G. Objectivity - a metric of strength of knowledge

By generating knowledge about the system, the epistemic uncertainty about deviation from objective changes, that is, knowledge about how an accident may occur or the potential consequence should it occur. High risk means severe potential consequences combined with a large degree of uncertainty (epistemic and/or aleatory). As knowledge decreases uncertainty, high risk requires strong knowledge, that is, knowledge substantiated with strong grounds for justification.

Justification, and thereby knowledge, is, among other things, based on artefacts representing the system and its properties, together with how these artefacts are interpreted, that is, the reasoning used to conclude based on these artefacts. Artefacts,

[4]Perhaps the most commonly known is the so-called confirmation bias, that is, our tendency to seek evidence that confirms our prior beliefs. However, most other cognitive biases are at work, like the *illusion of understanding* and *what you see is all there is (WYSIATI)*, that is, our tendency of believing that we understand complex topics by filling in the information gaps and the epistemic gaps so that the story becomes compelling and coherent, which leads to confidence in the truthfulness of the story (or proposition in this case).

[5]Risk is divided into system risk (undesired consequences of operating a system) and assurance risk (risk of making wrong decisions due to weak or inaccurate knowledge). Assurance risk is an epistemic risk and involves reducing knowledge-related errors.

[6]A famous statement from software testing illustrates this: Software testing cannot prove the absence of bugs, only their presence. A proposition that some software code is bug free (p) cannot be proven through testing alone. Software testing tries to find bugs, and when no bugs are found, one may start to believe p because one haven't found evidence that $\neg p$ is the case. (Of course, as most testing is non-exhaustive, not finding bugs does not mean the absence of bugs.)

such as training data, algorithms, source code, and system descriptions, may represent the system directly. Other kinds of artefacts may indirectly represent it, e.g., artefacts generated through verification, such as test cases, test results and results from inspections and reviews. The strength of knowledge is directly linked to these artefacts and the process of generating and collecting them.

Distinguishing weak from strong knowledge requires a metric by which the strength of knowledge can be assessed. By comparing the definitions of knowledge and assurance, we recognise the similarities. Both definitions contain the term ”justified”: The degree of justification for a true belief (knowledge) - the grounds for justified confidence (assurance). Degree of justification is central in assessing both strength of knowledge and degree of confidence (via uncertainty as shown in Figure 2). A high degree of confidence requires strong ground for justification.

The degree of objectivity encapsulates the aspects important for assessing the degree of justification, that is, the strength of knowledge. Hence, the strength of knowledge is measured through the degree of objectivity. The likelihood that the result of an enquiry represents the truth increases if it is conducted in an objective manner, including the artefacts produced and used in that enquiry.

Ensuring consistency and repeatability in our enquiries requires that the concept of objectivity is described. Objectivity in this context is a multi-dimensional, non-orthogonal and non-binary concept [19]. Hence, objectivity cannot be treated based on a reductionist approach.

There are three categories (i.e. dimensions) that lay out the space of objectivity [19], [13] (Figure 8):

1) properties and processes by which the artefacts are generated
2) reasoning, or the thinking about those artefacts
3) social processes concerning items 1 and 2.

Item 1 is about interacting with the system and its stakeholders during its entire lifecycle. It is about the choice of methods, how they are applied, and how those decisions influence the properties of the outcomes, that is, the artefacts. This category also includes procedures, methods, techniques, first principles in physics, standardised equations, algorithms, etc.

Item 2, this category is about how people and organisations think and the reasons and positions they take based on their interests and roles. This includes the involved assurance agent's values and independence from the developer.

Item 3 is about the social processes that advocate different viewpoints, such as agreement among subject-matter experts about the suitability and the correct use of methods used to generate artefacts and how to think about those artefacts. This kind of objectivity can be thought of as a form of inter-subjectivity and is strengthened if the group consists of individuals with different but relevant competence. The content of standards is a result of such agreements.

An important activity in assurance is the generation and collection of evidence through verification and validation (V&V). V&V is described through two properties: 1) The level

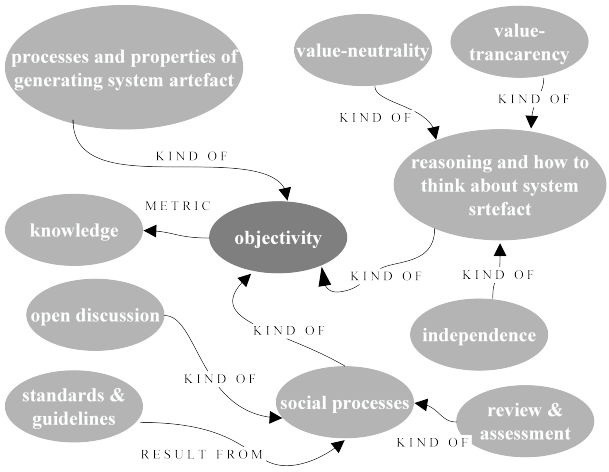

Fig. 8. Categories of objectivity

of intensity in the V&V effort, and 2) the level of rigour in the V&V effort [20]. V&V intensity is connected to the size of the scope, the number of system artefacts investigated, and the level of V&V involvement in each phase of the system lifecycle. V&V rigour is connected to comprehensiveness and thoroughness, leaving less room for logical inconsistencies and contradictions in the results, that is, performed with different levels of formality concerning techniques and documentation. One useful metaphor describing the relationship and difference between the two properties may be that increased V&V intensity makes the mesh width smaller and smaller while increasing the V&V rigour means that each mesh is investigated closer and closer.

The output from the V&V effort is the evidence representing the system properties of interest, such as safety, reliability, robustness and security. V&V intensity and rigour affect the evidence properties [20] such as quality, capability, and coverage.

Confidence is a result of the assessment of the strength of justification and knowledge through the degree of objectivity. Furthermore, through the V&V intensity and rigour, and the resulting evidence properties. The assessment cannot be a simple checklist which results in a numerical score aggregated as a simple sum or a single-dimensional category. The strength of knowledge must be assessed in each particular project in the context of a totality. That is, the strength (of knowledge) is not a resultant property of the degree of objectivity (and V&V), but emergent. Assessing the truthfulness (strength of justification and knowledge) of claims made about emergent properties in novel, complex safety-critical systems depends on the judgement of experts in the relevant disciplines. It is guided by the objectivity criteria described here.

H. Assurance case - a systematic way to represent knowledge

The assurance case is a way to represent knowledge (Figure 9 and Figure 2). At its core, an assurance case consists of a hierarchy of claims and arguments, including evidence that substantiates those claims. The claims are equivalent to the

before-mentioned propositions, and the argument is equivalent to the before-mentioned justifications. Moreover, claims can be understood as a reformulation of system requirements. A question may be how to lay out and organise arguments, which is the topic of this section.

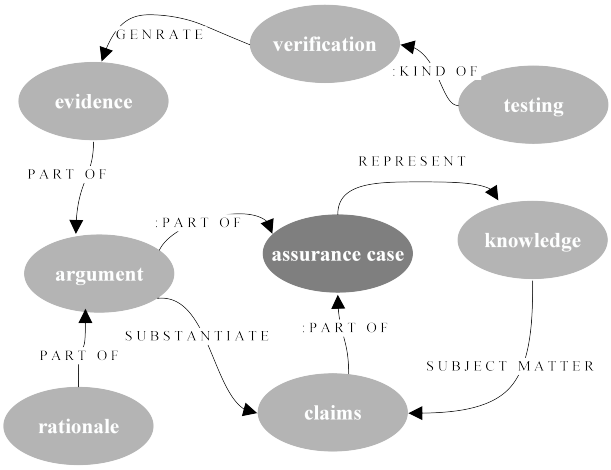

Fig. 9. The assurance case represents the knowledge in an assurance effort

One of the most recognised and influential layouts of arguments is the schema described by Stephen Toulmin in his 1958 book "The Uses of Arguments" [21]. Toulmin's motivation was to create a richer format that better reflected how people argued in reality instead of the more formal and traditional format consisting of premise and conclusion.

The argument layout consists of six elements [22]: Claim (or Conclusion) (C), Data (D) (or Datum, Toulmin uses both terms), Warrant (W), Qualifier (Q), Backing (B), Rebuttal (R) (Figure 10). In the simplest form, (D) may be some evidence

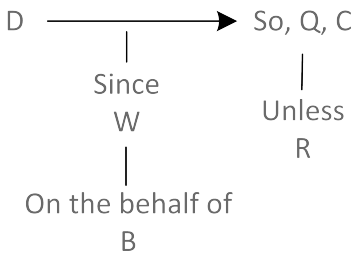

Fig. 10. General layout of an argument [22, p. 97]

that proves that (C) is the case. The transition between (D) and (C) may not be trivial, so a warrant needs to act as an inference licence between (D) and (C); that is, (W) acts as a bridge between (D) and (C). (W) may also be challenged, so a backing (B) may be needed to support (W), that is, why (W) holds. (Q) indicates the strength of the step (i.e. strength of the "bridge") from (D) to (C). (R) indicates circumstances in which (W) may not hold.

Although the elements of an argument described by Toulmin are necessary aspects of an epistemic justification substantiating a proposition or assertion, the schema, in its simplest form, is insufficient for assurance of complex systems. The schema needs to be expanded.

Firstly, in the assurance of complex systems, there are many claims. System claims represent statements about the system properties and its use. These requirements address many systems properties, including safety. Moreover, the claims must be refined at several levels of abstraction (LoAs). The LoAs link back to the LoAs connected to the systems approach and foundationalism.

Secondly, although one of Toulmin's key motivations was to enable "practical assessment of arguments" [22], he did not discuss aspects of argument assessment in detail. Clearly, when, e.g., a (top) claim is refined into two or more subclaims with accompanying justification, assessing the strength of each argument needs to be aggregated in some way to reflect the confidence in the top claim. Moreover, each element in the argumentation schema should be assessed, leading to a network of assessments on different elements of an argument on different LoAs.

Several expanded argument schemas based on Toulmin have been developed, such as Goal Structuring notation (GSN) [23] and Trust-IT [24].

An assurance case organises these arguments systematically and structured and represents the knowledge generated in the assurance (Figure 2). Different ways are possible based on the various argument schemas, such as [23] or [25]; both are compatible with [1]. A metamodel of an assurance case may also be found in [26].

I. Stakeholder's objectives and system requirements

Stakeholders hold objectives and pursue goals through utilising the AI system; that is, they use the system for a reason. A system's mission is expressed as system requirements are elicited and understood from these objectives.

The stakeholders may be users, developers, and bystanders who have nothing to gain from the system but may be affected by it. Through its legislation and standards, the government represents stakeholders that cannot be consulted directly, such as the natural environment, future generations, the general public, children, etc. In such cases, conformance to standards means meeting stakeholders' objectives and interests.

Stakeholders need confidence that their objectives are, or will be, fulfilled or that a deviation from those objectives is acceptable. Implicitly, stakeholders also hold the objectives of being safe, secure, and treated fairly. These objectives may not be directly linked to the reason for developing and using the system in the first place. The system requirements must incorporate such implicit objectives. These kinds of system requirements can be termed mission-supporting requirements, or non-functional requirements [27], or even system constraints[7] [7] (Figure 11).

[7]Prof. Nancy Leveson terms this "safety constraints", however, expanding the scope of such requirements to other system quality characteristics they can be termed as "system constraints".

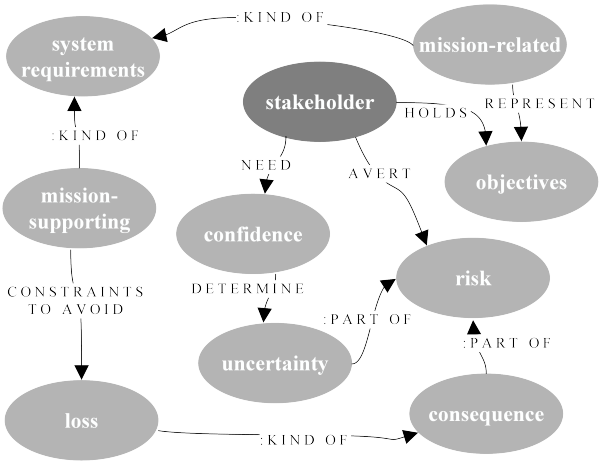

Fig. 11. Stakeholders hold objectives that determine the system requirements

In the context of assuring AI systems, mission-supporting system requirements should be based on a set of ethical principles such as: [28].

Conflicts often arise between requirements directly related to the mission of the system and the mission-supporting requirements. Moreover, similar conflicts may also arise between the objectives and goals of different stakeholders, and even between different objectives of the same stakeholder (e.g. long-term vs. short-term goals). One understanding of ethics is: "the identification, study, and resolution or mitigation of conflicts among competing values or goals" [29]. The assurance effort should document[8] the trade-offs made between competing goals.

J. Conclusion

Assurance is a critical epistemic activity that provides justified confidence in system properties such as safety. By systematically generating and assessing knowledge, assurance efforts reduce uncertainties and support informed decision-making. The paper underscores the importance of objectivity in evaluating the strength of knowledge and the role of verification in producing evidence with adequate properties. The CESM metamodel offers a robust framework for understanding and analysing system behaviour. As technologies evolve, particularly with the advent of AI, assurance methodologies must adapt to address new challenges and ensure stakeholder confidence. Ultimately, effective assurance contributes to the safe, reliable, and responsible deployment of complex systems, benefiting stakeholders and society at large.

REFERENCES

[1] "ISO/IEC/IEEE International Standard - Systems and software engineering–Systems and software assurance –Part 1: Concepts and vocabulary," Mar. 2019.

[2] European Parliament, Council of the European Union, "Artificial Intelligence Act," Jun. 2024.

[3] "DNV-RP-0671 Assurance of AI-enabled systems," Sep. 2023.

[4] "ISO/IEC/IEEE International Standard - Risk management — Guidelines," Feb. 2018.

[5] C. R. Fox and G. Ülkümen, "Distinguishing Two Dimensions of Uncertainty," in *Perspectives on Thinking, Judging, and Decision Making: A Tribute to Karl Halvor Teigen*. Universitetsforlaget, 2011, p. Chapter 1.

[6] D. Kahneman, *Thinking, Fast and Slow*, 1st ed. New York: Farrar, Straus and Giroux, 2011.

[7] N. G. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety*. Cambridge, Massachusetts: MIT Press, Jan. 2012.

[8] M. Bunge, *Emergence and Convergence: Qualitative Novelty and the Unity of Knowledge*, ser. Toronto Studies in Philosophy. Toronto ; Buffalo: University of Toronto Press, 2003.

[9] O. Haugen, "Safety assurance of complex systems Part 2: Assurance and analysis," DNV AS, Høvik, Norway, Whitepaper, 2019.

[10] O. I. Haugen, "A Systems Approach to Modelling Emergent Behaviour in Maritime Control Systems Using the Composition, Environment, Structure, and Mechanisms (CESM) Metamodel," 2025.

[11] E. Hollnagel, *FRAM: The Functional Resonance Analysis Method, Modelling Complex Socio-Technical Systems*. Ashgate Publishing Limited, 2012.

[12] C. W. Bytheway, *FAST Creativity & Innovation: Rapidly Improving Processes, Product Development and Solving Complex Problems*. Fort Lauderdale, Fla: J. Ross Pub, 2007.

[13] O. I. Haugen, "The Systems Approach," in *Demonstrating Safety of Software-Dependent Systems; With Examples from Subsea Electric Technology*, T. Myhrvold and M. van der Meulen, Eds. DNV AS, 2022, pp. 145–163.

[14] J. Nagel, *Knowledge: A Very Short Introduction*, first edition ed., ser. Very Short Introductions. Oxford: Oxford University Press, 2014, no. 400.

[15] J. C. Watson, "Epistemic justification," *Internet Encyclopedia of Philosophy*.

[16] M. A. Falappa, G. Kern-Isberner, and G. R. Simari, "Belief Revision and Argumentation Theory," in *Argumentation in Artificial Intelligence*, 1st ed., I. Rahwan and G. R. Simari, Eds. Boston, MA: Springer Dordrecht Heidelberg, Jul. 2009.

[17] F. Paglieri and C. Castelfranchi, "The Toulmin Test: Framing Argumentation within Belief Revision Theories," in *Arguing on the Toulmin Model: New Essays in Argument Analysis and Evaluation*, D. Hitchcock and B. Verheij, Eds. Dordrecht: Springer Netherlands, 2006, pp. 359–377.

[18] ——, "Argumentation and Data-oriented Belief Revision: On the Two-Sided Nature of Epistemic Change," in *CMNA IV: 4th Workshop on Computational*, Jan. 2004.

[19] H. E. Douglas, *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press, 2009.

[20] O. I. Haugen, "Safety assurance of complex systems Part 3: Verification and evidence," DNV, Høvik, Norway, Whitepaper, 2019.

[21] B. Verheij, "The Toulmin Argument Model in Artificial Intelligence – Or: How semi-formal, defeasible argumentation schemes creep into logic," in *Argumentation in Artificial Intelligence*, 1st ed. Springer New York, NY, Jan. 2009, pp. 219–238.

[22] S. E. Toulmin, *The Uses of Argument*, 2nd ed. Cambridge University Press, 2002.

[23] "Goal Structuring Notation Community Standard," May 2021.

[24] J. Górski, Ł. Cyra, A. Jarzębowicz, and J. Miler, "Argument Strategies and Patterns of the Trust-IT Framework," *Polish Journal of Environmental Studies*, vol. 17, Jan. 2008.

[25] "Argevide - System assurance management tools - Assurance cases," https://www.argevide.com/home/, Oct. 2023.

[26] "Structured Assurance Case Metamodel (SACM)," Oct. 2023.

[27] A. van Lamsweerde, *Requirements Engineering: From System Goals to UML Models to Software Specifications*. Chichester, England ; Hoboken, NJ: John Wiley, 2009.

[28] High-Level Expert Group on AI, "Ethics Guidelines for Trustworthy AI," European Commission, B-1049 Brussels, Tech. Rep., Apr. 2019.

[29] L. McDaniel, "What Is Bioethics?" https://bioethics.msu.edu/what-is-bioethics.

[8]Decisions about these trade-off lays more to the responsibility of design and risk management.