Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Runtime Half of the Safety Case Aboard Autonomous Ships: Evaluating Provisos with Claim Graphs and Warrant Records under the MASS Code

Fredrik Olsson

fredrik.x.olsson@ri.se

RISE Research Institutes of Sweden

Ted Sjöblom

RISE Research Institutes of Sweden

Luis Sanchez-Heres

RISE Research Institutes of Sweden

Krister Blanch

RISE Research Institutes of Sweden

Research Article

Keywords: MASS Code, safety case, runtime assurance, operational design domain, non-compensatory aggregation, warrant record

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11458659/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

The Runtime Half of the Safety Case Aboard Autonomous Ships: Evaluating Provisos with Claim Graphs and Warrant Records under the MASS Code

Anonymous authors

Affiliation withheld for review.

Abstract

A safety case makes design-time claims about a vessel that hold only while certain conditions are met in operation, and we call checking these conditions, and responding when one fails, the runtime half of the safety case. Without crew or shore link, an autonomous vessel must perform that runtime half itself. Treating the runtime half as a conformity assessment, of the kind performed in class surveys and flag certification, we derive five functions that a mechanism performing it aboard must provide. We find that the International Code of Safety for Maritime Autonomous Surface Ships requires one of these functions under any interpretation, and the other four only under certain interpretations of three of its provisions. To find out whether the five functions can be implemented, and what an assessor can establish from its record, we propose an architecture that provides all five. The safety case is held aboard as a graph of claims. On each cycle, a non-compensatory evaluator assigns each claim a standing based on its conditions of application and writes a warrant record for it. We illustrate the architecture with a staged reproduction of the *Royal Majesty* receiver failure, beside two composite-score policies, whose records show which source failed but not which claims it withdrew. To the Code's experience-building phase we recommend interpreting the three provisions so that all five functions are required, and list open questions on approval, change control and record keeping to be settled.

Keywords: MASS Code, safety case, runtime assurance, operational design domain, non-compensatory aggregation, warrant record

Abbreviations

GNSS	global navigation satellite system
ISM	International Safety Management
MASS	Maritime Autonomous Surface Ship
ODD	operational design domain
OE	operational envelope
SOLAS	International Convention for the Safety of Life at Sea
SPI	safety performance indicator
STCW	Standards of Training, Certification and Watchkeeping for Seafarers

1 Introduction

Autonomous and remotely operated ships are moving from demonstration projects to applications for approval. The first approvals will rest, as approvals of novel ships do today, on a safety case: a structured argument, supported by a body of evidence, that the vessel is acceptably safe for its intended operation and operating environment (adapted from UK Ministry of Defence 2007). Notations and guidance for such arguments include Goal Structuring Notation, assurance-case building blocks and SACE (SCSC Assurance Case Working Group 2021; Bloomfield and Netkachova 2014; Hawkins et al. 2022), and Glomsrud et al. (2024) apply them to an autonomous ferry. The argument is written during design, assessed by an authority before the vessel enters service, and kept as documentation afterwards.

A safety case is conditional: it argues that the vessel behaves acceptably while its sensors, operating conditions, and traffic situation remain within the envelopes covered by verification and validation. We call these conditions the provisos of the case. Whether they hold can only be determined during operation. On a conventional ship, the master and the officer of the watch make this check throughout a voyage and respond when a proviso fails. The check and the response are in this sense the runtime half of the safety case, although the case does not name them as such. An autonomous vessel without crew or shore link must make the check itself or enter a predefined fallback state.

The 1995 grounding of the cruise ship *Royal Majesty* off Nantucket illustrates a failed runtime check (National Transportation Safety Board 1997; Lützhöft and Dekker 2002). Its satellite navigation receiver lost its fix and continued to emit dead-reckoned positions. The receiver flagged these positions as invalid, but the integrated bridge system did not read the flag and for about 33 hours treated them as measurements. Its display gave no sign that the position had lost its basis, the indication on the receiver in the chart room went unnoticed, and the watch did not cross-check the position against an independent source. No record held the receiver's mode or the validity flag. Investigators timed the failure from anomalies in the course and speed recorded in the bell log, and established its mechanism from the equipment design and tests after the accident (National Transportation Safety Board 1997, pp. 19, 30 and 36). The investigation documents the failure in enough detail to reproduce it, and we use the case as a running example throughout the paper.

A similar failure occurred in 2023 aboard the ro-pax ferry *Marco Polo* in Hanö Bay, on an electronic chart display and information system (ECDIS) installed two years earlier (Swedish Accident Investigation Authority 2025). A date fault in the receiver that supplied the navigation system made the ECDIS reject its position and switch to dead reckoning. The alarm was acknowledged at a change of watch. After that, only a small marking on the ECDIS and a note on one radar showed the switch, and for about three hours the ECDIS showed the vessel on its planned route until it ran aground. The second receiver supplied a correct position throughout, but the navigation system took its position only from the receiver chosen at a manual selector.

The instrument under which autonomous cargo ships will be approved is the International Code of Safety for Maritime Autonomous Surface Ships (the MASS Code), which took effect on 1 July 2026 (International Maritime Organization 2026a). Numbers in parentheses without a source refer to clauses of the Code. It applies to cargo ships with systems for autonomous or remote operation that are covered by chapter I of the International Convention for the Safety of Life at Sea (SOLAS), and to their remote operation centres, where the Administration deems compliance with other instruments impracticable or insufficient (2.1). The *Royal Majesty* and the *Marco Polo* are passenger ships and therefore outside its scope. Neither failure depended on the type of ship, and we use both cases as illustrations although the Code does not apply to them. The Code does not use the term safety case. Its approval basis nevertheless has the same parts: a concept of operations, a risk assessment, and verification and validation evidence. The Code is goal-based and technology-neutral, supplements the existing conventions, and is non-mandatory while a mandatory instrument is prepared (1.2). The Maritime Safety Committee is scheduled, at its 112th session in December 2026, to set the framework for an experience-building phase in which Member States and industry apply the Code and report their findings. Work on the mandatory instrument is scheduled to begin in 2028, with adoption in 2030 and entry into force in 2032 (International Maritime Organization 2026b). The mandatory instrument will draw on the current Code and on the results of the phase. Since no settled interpretive practice exists yet, the interpretations adopted during the phase may shape the mandatory text. We intend this paper as input to that phase.

We address three questions.

1. What must an autonomous vessel do aboard to establish, while it operates, that the conditions its safety case assumes still hold?
2. Does the MASS Code require it?
3. Can it be implemented, and what can an assessor establish from its record?

Section 2 answers the first question by deriving five functions an onboard mechanism must provide, and section 3 answers the second by mapping five clauses of the Code to them. Section 4 defines an architecture that provides the functions (formal model in appendix A), and section 5 answers the third by running a concept implementation of it against a staged reproduction of the *Royal Majesty* receiver failure. Section 6 discusses the implications for approval, recommendations to the experience-building phase, limitations and further work, and section 7 concludes.

2 The runtime half of the safety case

2.1 Where a safety case states its provisos

A safety case decomposes a top-level goal into sub-goals and items of evidence, and the notations in use record beside each step why the evidence supports the goal (SCSC Assurance Case Working Group 2021; Bloomfield and Netkachova 2014). Each such step has the parts that Toulmin (2003) identified in practical arguments, and we use his vocabulary throughout. A practical argument moves from *data*, later called *grounds* (Toulmin et al. 1984), to a *claim*. A *warrant* justifies this move by stating that data of this kind support a conclusion of that kind. The warrant rests on *backing*, such as experience, testing, or authority. Backing is kept apart from the warrant because it supports the warrant through empirical, institutional, or legal means and is not part of the inference itself. A *qualifier* states the strength of the claim. A *rebuttal* names a circumstance in which the warrant does not apply, so that the claim must be withdrawn. Figure 1 applies the six parts to the proposition on which the *Royal Majesty*'s bridge system relied: that the position was known to within a stated bound.

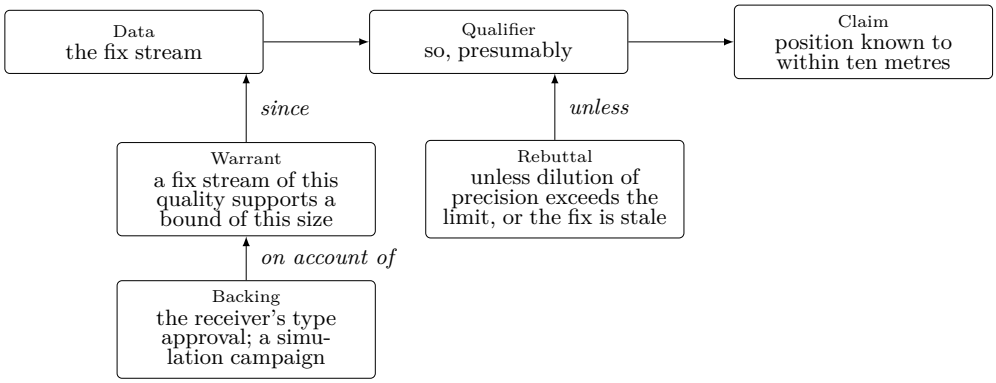


Fig. 1 Toulmin's six parts applied to the position argument on which the *Royal Majesty*'s bridge system relied.

In a safety case, the provisos appear as rebuttals. In Goal Structuring Notation they are usually written as assumption or context elements (SCSC Assurance Case Working Group 2021). A case for an acceptable navigation function, for example, depends on sensors operating within their validated envelope, on conditions remaining within the verified design domain, and on traffic situations represented in the test campaign. Before service, an assessor can judge whether the backing and the provisos are sufficient. The assessor cannot, however, determine whether the provisos hold at a given time during operation.

2.2 Design-time and runtime assurance

Safety assurance establishes and maintains justified confidence that a system is acceptably safe, and the safety case documents that confidence. *Design-time assurance* establishes warrants and backing before operation, and an authority assesses them through certification. *Runtime assurance* checks, while the system acts, that the rebuttals of its warrants remain absent. It also determines which actions remain permitted when a rebuttal is present. We use the term only for the obligations that a safety case leaves to operation.

Rushby (2008) introduced runtime certification, in which part of a certification argument is discharged only in operation. Denney et al. (2015) keep the safety case under revision as operational evidence arrives. Assurance 2.0 requires that the defeaters of an argument be identified and either refuted or accepted as residual doubts before deployment (Bloomfield and Rushby 2022; Bloomfield et al. 2024). UL 4600 conditions claims of a design-time case on safety performance indicators, metrics compared with thresholds whose violation falsifies the claim (Underwriters Laboratories 2023). In each approach, evidence from operation can falsify a proviso in the safety case.

Consider the claim in figure 1, that the position is known to within ten metres. The safety case supports the navigation function on that condition, which makes the condition a proviso of the navigation argument and an obligation on whoever operates the ship. The obligation is *discharged*, in the sense in which a duty is discharged, when the operator establishes that the condition holds at the relevant time and closes the obligation. On a conventional ship, the officer of the watch discharges it by checking the position at frequent intervals and keeping a record of the watch, as section A-VIII/2, part 4-1 of the Seafarers' Training, Certification and Watchkeeping (STCW) Code requires (paragraphs 25 and 31) (International Maritime Organization 2017). Where circumstances allow, the position is fixed by more than one method (paragraph 47), such as radar ranges to the shore against the satellite fix. The *Royal Majesty*'s watch did not make this check (section 1).

We treat discharging a proviso as a conformity assessment, of the same kind as class survey and flag certification. ISO/IEC 17000 divides any conformity assessment into three activities: selection, determination against specified requirements, and review with decision and attestation (International Organization for Standardization and International Electrotechnical Commission 2020). Applied to a single proviso in operation, determination and review give a discharge three elements, which we name as follows. Selection, the choice of what to assess, is made when the safety case states the proviso.

- **An obligation precise enough to check.** This is the specified requirement against which determination is performed. A bound of ten metres can be checked, whereas "the sensors perform within their envelope" cannot.
- **Evidence that applies at the relevant time.** This is the information that the checks gather and that determination tests against the obligation. Here it is the radar ranges taken on this watch. In general it is a measurement, a monitor output, or a signed confirmation.

- **An authority who judges the evidence sufficient and closes the obligation.** This is review, decision and attestation. Here it is the officer, who signs the log.

Current evidence is valid only for a limited time. The range check made at the start of the watch does not establish the position an hour later. A runtime discharge is therefore a continuing obligation, renewed at a defined rate, which ISO/IEC 17000 calls surveillance. A discharge also requires a specified response when the evidence lapses, and the safety case names that response when it lists a degraded state. Table 1 compares the three elements on a conventional ship and on an autonomous vessel. On a conventional ship, a later reader of the log audits the officer's judgement. On an autonomous vessel, the assessor delegates the authority by approving the tests and responses in advance. The delegation covers applying the approved rule, not judging its adequacy, so a later reader must be able to confirm from the record that the approved rule was applied, and to which evidence.

Element	Conventional ship	Autonomous vessel
Obligation	a proviso in the safety case, which the officer has read and understood	the same proviso, written as a test the vessel can apply to its current state
Current evidence	a check the officer makes, such as radar ranges against the fix, valid for this watch	a measurement or monitor output, to which the vessel applies the test each time it evaluates the proviso
Authority	the officer, who judges the check sufficient, closes the obligation and enters it in the log	the vessel, applying tests and responses that the assessor approved in advance

Table 1 The three elements of a discharge on a conventional ship and on an autonomous vessel.

Figure 2 places the discharge aboard. The vessel's functions, such as track keeping and collision avoidance, read the sensors and command the actuators. In one mode the automation keeps the track and avoids collisions. In another a remote operator steers, and the onboard collision avoidance only advises. The mode of operation sets this division of work for every function. The safety case argues that this *control path* behaves acceptably while its provisos hold. We place beside it an *assurance path* whose stages are the elements of a discharge (table 1), together with a record: the checks supply the evidence, the determination tests the provisos against it, the response exercises the authority over the mode of operation, and the record keeps each determination with its basis. Aboard the *Royal Majesty*, the control path continued on dead-reckoned positions after the assurance path had lapsed: the evidence of the lost fix was shown only on the receiver in the chart room, as two small characters and a one-second chime, and the watch did not monitor it (National Transportation Safety Board 1997, pp. 37–38). Runtime assurance in the sense of ASTM F3269 (spelled run-time assurance there) separates two paths in the same way, with a monitor that switches control from a complex function to a recovery function (ASTM International 2021). Its monitor checks the state of the system against a safety boundary, whereas the assurance path of figure 2 determines whether the provisos of the safety case hold.

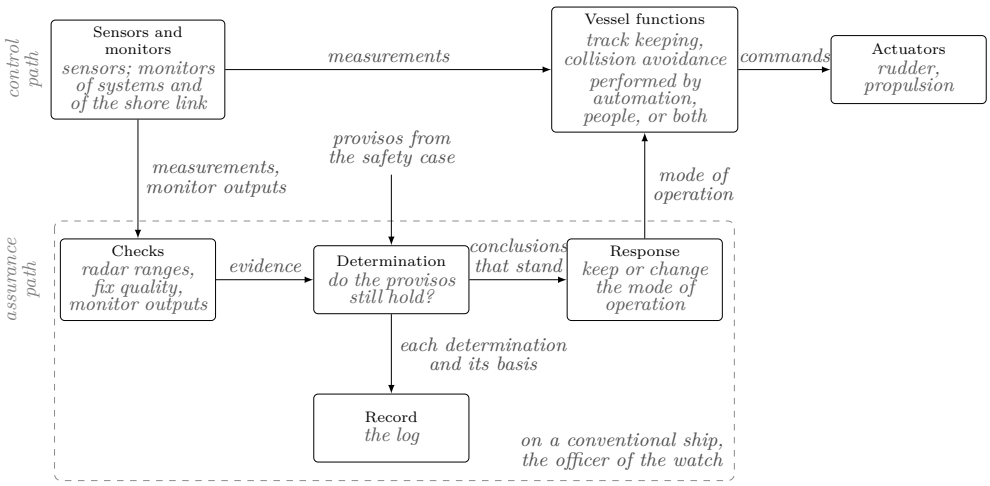


Fig. 2 The control path and the assurance path aboard.

A person, whether crew or a remote operator, can still hold the authority, as the officer of the watch does today, but only for a check that the safety case names and only if the vessel shows the evidence. A statement that an operator will intervene names no check, and people left to monitor automation detect its failures poorly (Bainbridge 1983; Parasuraman and Riley 1997). The safety case must therefore state the assurance path explicitly whenever the person who holds the authority no longer gathers the evidence.

A proviso attaches to a warrant, and the warrants of a safety case are chained: the conclusion of one step serves as data for the next. The position conclusion of figure 1 is used to keep the track, the track is used in the collision assessment, and which modes of operation the vessel may use depends on whether these conclusions hold. When the position is withdrawn, track keeping and collision avoidance lose their basis as well, unless they have another source of position. Aboard the *Royal Majesty*, the receiver marked its output invalid, but the autopilot did not read the mark and kept steering the ship on the dead-reckoned position (National Transportation Safety Board 1997, pp. 30 and 36). The failure was detected at the receiver and never reached track keeping. A vessel with two receivers can keep its track on the second when the first fails, with less margin. The *Marco Polo* had such a second receiver, but its navigation system read only the receiver chosen at a manual selector (Swedish Accident Investigation Authority 2025, pp. 12 and 33–34). A runtime discharge therefore requires that each withdrawal reach every conclusion that depends on it, and withdraw those that have no other support left. The record must then show which conclusions a failure withdrew. From the failed receiver alone, a later reader cannot tell whether the vessel was right to stop keeping its track.

2.3 Functions of onboard runtime assurance

We argue that any mechanism carrying the runtime half aboard must provide five functions. Table 2 lists them, each beside the property of a runtime discharge from section 2.2 from which it follows.

Property of the discharge	Function the mechanism must provide
Authority exercised aboard, since connectivity is itself a proviso and evaluation cannot depend on the shore	1. Hold the justification for the vessel's conduct aboard, in a form the runtime can evaluate.
Obligation	2. Represent the conditions for withdrawing a conclusion as predicates over current state.
Current evidence, renewed at the surveillance rate	3. Identify, at each determination, the evidence and reasoning supporting every conclusion.
Dependency between conclusions	4. Propagate each withdrawal to the dependent conclusions and withdraw those left without support.
Authority delegated in advance, so its exercise must be auditable afterwards	5. Record every evaluation and withdrawal, not only the latest, with enough detail for a later reader to reconstruct each decision and its basis.

Table 2 The five functions of onboard runtime assurance and the property of a runtime discharge from which each follows.

The officer of the watch performs all five functions today without naming them. The passage plan, the standing orders and the officer's training hold the case for the passage. The officer knows what would invalidate a fix (poor geometry, a stale receiver, a range that disagrees with the chart) and makes the check. The officer also knows what the check bears on: a bad position affects the track, the clearing bearings and the collision assessment. The check goes into the log, and a week later the officer can explain which check supported which decision. A vessel can give that explanation only from what it recorded at the time. For this reason function 5 covers every evaluation. In figure 2, functions 1 to 4 belong to the determination and function 5 to the record.

Under function 5, a later reader must be able to reconstruct each decision and its basis. Reconstruction can mean three operations, which require different inputs (table 3). *Inspection* reads from the record which conclusions stood at a given instant and on what grounds. *Recomputation* applies the approved rule to the recorded evidence and shows whether the vessel applied it correctly. *Verification* shows whether the justification that the vessel evaluated is the approved one, which requires the record to identify the version in force. A log of inputs and outcomes lets a reader rerun the software and reproduce each outcome, but it holds no basis to inspect.

These functions establish which conclusions stand, but not what the vessel does about a withdrawn one. That choice belongs to whoever holds the authority (section 2.2). When the remote operation centre can be reached, a remote operator makes it, and needs to see each withdrawn conclusion with its evidence and the conclusions that depend on it. When the centre cannot be reached, the vessel follows the rule approved in advance. Both choices rest on the same onboard evaluation.

Operation	What it answers	Inputs
Inspection	what stood at a given instant, and on what grounds	the record alone
Recomputation	whether the vessel applied the approved rule	the recorded evidence and the approved justification
Verification	whether the justification the vessel evaluated is the approved one	the record's identification of the version in force and the approved justification

Table 3 Three operations on a record and the inputs each requires. The approved justification lies outside the record.

The unit of evaluation is the individual conclusion. Each warrant in the safety case yields one conclusion, and what the vessel may do depends on which of them stand. A safety case written as a single warrant for the whole vessel could only say whether the vessel may continue, not which proviso had failed.

3 The MASS Code

Section 2.3 derived five functions that a vessel without crew or shore link must provide to check the provisos of its safety case during operation (table 2). We now examine whether the MASS Code (International Maritime Organization 2026a) requires them.

Approval under the Code rests on documents that the Administration assesses before the vessel enters service: a concept of operations, a chapter 7 risk assessment, and verification and validation evidence, judged against criteria agreed under clause 6.2 (annex 1). These documents form the design-time argument of section 2.2. When the assumptions, systems or equipment are changed during operation, the affected part of the risk assessment is to be repeated (6.4.2; annex 1, table 1, operation column). A receiver that loses its fix during a voyage is not such a change, and the vessel has to deal with it at sea. We therefore look for clauses that require the vessel itself to detect such a failure, determine what follows from it, and record the determination. For each clause we note whether its requirement follows from the text alone or depends on how the text is interpreted. Administrations will settle the interpretations during the experience-building phase (section 1), so we state each one we rely on.

The reading relies on four of the Code's terms, listed in table 4. The operational envelope (OE) and the operational design domain (ODD) bound what is evaluated during operation. The degraded state and the fallback state are the two operating states that can follow a deviation: a degraded state while the vessel can still operate within its OE, and a fallback state when it cannot. The vessel's OE includes the ODDs of all relevant systems and functions (8.3). Fjørtoft and Rødseth (2020) and Rødseth et al. (2022) developed the operational envelope for autonomous-ship approval, and guidance for unmanned naval craft uses a counterpart concept (European Defence Agency 2022).

Term	Clause	Definition, paraphrased	Attaches to
Operational envelope (OE)	4.27	the description of the MASS's operational capabilities and limitations	the vessel
Operational design domain (ODD)	4.26, 8.3.2.1	the design range in which a system can operate as intended, including internal events such as failure or degradation, and relevant external factors such as available resources, geographical or environmental situations	one function or system
Degraded state	4.9	a deviation from normal operation (operation within the OE, 4.24) that may lead to a fallback state	the vessel
Fallback state	4.12, 8.4	a predefined state entered when the MASS cannot remain within its OE	the vessel

Table 4 Terms of the MASS Code used in the argument.

3.1 Five clauses read from the runtime side

8.3.2.2: the ODD must be evaluable while the vessel runs.

Clause 8.3.2.1 requires the ODD of an individual function to be based on a risk assessment and lists the required design information, including safe operating conditions and foreseeable malfunctions with their degraded states. Clause 8.3.2.2 then requires that a system "should operate within its ODD and should be able to detect whether its current state of operation meets the ODD". Detection therefore requires an onboard representation of the ODD that the running system can evaluate. A monitor that compares measurements with configured thresholds is enough to satisfy the clause, and no interpretive choice is involved.

8.3.3: determining whether the vessel can remain within its OE.

Clause 8.3.3 reads in full:

> A single autonomous or remotely-operated system or function deviating from its ODD should not necessarily result in the MASS deviating from its OE. As long as the MASS, as an integrated system, can continue to be operated within its OE, the deviation of an autonomous or remotely-operated system or function from its ODD should be considered as a degraded state.

The clause lets a component-level deviation remain a degraded state only while the integrated vessel can operate within its OE. It does not say how that condition is evaluated. In particular, it supplies no aggregation rule. The wording permits averaging of component states, and it equally permits a rule under which healthy components cannot compensate for an unsupported capability. Nor does the clause require the evaluation: it permits a function's deviation to stop short of the vessel level but does not require the deviation to stop there. We call this permission non-propagation. Two interpretations are therefore consistent with the text. Under the weak interpretation, a design treats every ODD deviation as a deviation from the OE, enters a predefined fallback state (8.4.1) and alerts the remote operation centre, consistent with the human master responsible for the MASS in every mode (8.7.3). Under the strong interpretation, the vessel exercises non-propagation and must then determine, at the

moment of deviation, whether it can continue within the OE in force at that time. Since coverage and connectivity form part of the OE (8.3.1.2, 17.2.9), the determination under the strong interpretation must sometimes be made aboard without a person, and it requires more than the monitor of clause 8.3.2.2.

9.10.4: recording the basis of the determination.

Clause 9.10.1 requires data logging for the analysis of performance, failures and incidents, and clause 9.10.2 sets a retention floor of thirty days. Clause 9.10.4 sets the content requirement:

> Logged data should contain appropriate detail to restore a complete record of the MASS operations, including, where applicable, remote operations and automated decision-making processes.

The determination under clause 8.3.3 is an automated decision-making process. Under a weak interpretation of "restore", it is enough to log the inputs, retain the software version, and rerun the software on the inputs, which reproduces the process. Under a strong interpretation, restoring the process also recovers the basis for its result: the evidence and the rule that produced it at the relevant instant. The qualifiers "appropriate detail" and "where applicable" govern how much of the basis is retained under the strong interpretation. They do not by themselves select between the two interpretations.

10.4 and 10.5: availability throughout the operational life.

Clause 10.4 requires software to be "transparent and explainable at all stages of its operational life, and for all decision-making processes". Clause 10.5 requires entities that develop, deploy, or operate software to implement mechanisms for its accountability and proper operation. It also requires software to be auditable and traceable to those entities, and requires governance mechanisms for oversight, impact assessment, audit, and due diligence throughout its operational life. Clauses 10.2 and 10.7.2 apply the same temporal scope to testing and to operator verification. Clause 10.4 continues:

> The transparency and explainability should ensure relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to autonomous capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.

Under a governance interpretation, clauses 10.4 and 10.5 concern this understanding and the organisational mechanisms listed in the third sentence of 10.5. Documentation, training, and governance frameworks then satisfy both, with no change to onboard software after certification. The second sentence of 10.4 reads naturally in this way, and the opening paragraph of chapter 10 supports it, since it places the principles in the approval process. Under an operational interpretation, the first sentence of 10.4 extends further than the second: "all decision-making processes" at "all stages of its operational life" include a decision made at sea, and personnel can understand such a decision only from its basis. The first sentence of 10.5 then requires mechanisms that ensure proper operation at runtime, and clause 10.4 requires an autonomous decision at sea to remain explainable at the time it is made.

3.2 What the Code requires and what it leaves to interpretation

Under either interpretation of each clause, the Code evaluates individual functions. The ODD attaches to one function and the OE to the vessel (table 4), and clause 8.3.3 makes whether the vessel can remain within its OE depend on whether each function is within its ODD. Section 2.3 reached the individual conclusion as the unit from the warrants of the safety case. Whether a function is within its ODD is one such conclusion, so the two routes agree at the level of the function. The Code leaves open the aggregation rule, that is, how the states of the individual functions combine into the determination of whether the vessel can remain within its OE.

For each clause, table 5 gives what the vessel must do in operation, which functions of section 2.3 the clause requires, and the interpretation the requirement depends on.

Clause	What the vessel must do in operation	Functions from section 2.3	Interpretation on which the row depends
8.3.2.2	Detect whether each function is within its ODD.	2: the conditions for withdrawing a conclusion are predicates the vessel can evaluate.	None.
8.3.3	When a function leaves its ODD, determine whether the vessel can continue within its OE. The determination must be possible aboard, since the link may be down.	1: the justification is held aboard. 3: at each determination, the evidence is tied to the conclusion it supports. 4: a withdrawal reaches every dependent conclusion whose requirements it leaves unmet.	That the vessel exercises the non-propagation the clause permits, rather than entering fallback after every deviation.
9.10.4	Keep a record from which the basis of each determination can be recovered.	5: each evaluation and withdrawal is recorded with its basis.	That restoring the record includes recovering the basis, not only rerunning the software on the inputs.
10.4, 10.5	Keep decisions explainable and auditable for as long as the vessel is in service.	1 to 5, renewed throughout the operational life.	The operational interpretation, rather than the governance interpretation.

Table 5 Mapping from MASS Code clauses to the five runtime functions.

The last three rows each admit a weak interpretation. Under the weak interpretations, the vessel enters fallback after every deviation (8.3.3), the log holds the inputs for rerunning the software (9.10.4), and documentation and training provide explainability (chapter 10). A vessel built to them needs only a monitor and a fallback trigger aboard. The design documents cover everything else the Code requires, and functions 1, 3, 4 and 5 are not needed. The choice between the two sets of interpretations therefore depends on whether the design documents can stand in for a determination made at sea.

4 The claim-graph architecture

We define one mechanism that provides the five functions of table 2. It fills the determination and record stages of the assurance path (figure 2): the safety case is held aboard as a graph of claims, an *evaluator* determines on each cycle which of the claims stand, and each evaluation is kept in the data log of clause 9.10. The construction follows Verheij's account of Toulmin's warrant as a conditional (Verheij 2005, 2003): if the grounds hold, the claim follows unless a rebuttal applies. The claim of one conditional can serve as grounds for another, so a rebuttal of one warrant changes every conclusion that depends on it. Appendix A states the model formally and proves the properties used here.

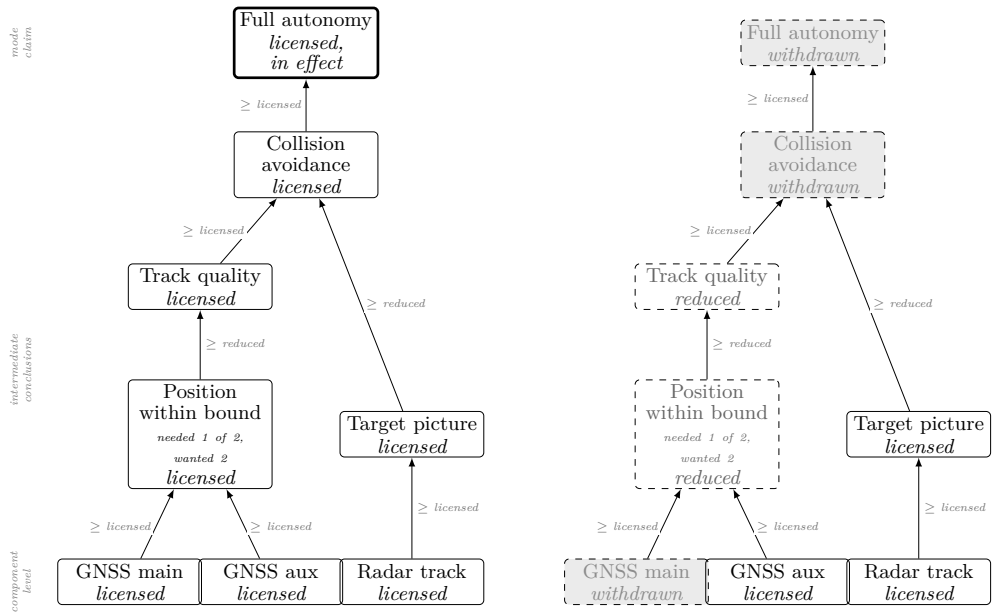
The other two stages of the assurance path lie outside the mechanism. The *evidence layer* performs the checks of the assurance path: on each cycle it assembles the measurements of the sensors and the outputs of the system and shore-link monitors into one snapshot. The *response stage* puts into effect one of the modes of operation (4.22) that the evaluation supports, including the fallback state (8.4), under the criteria of the concept of operations (8.6.1), and alerts the crew or the remote operator. The properties proved in appendix A hold only where these two stages and the evaluator behave as the model assumes, and appendix B states the assumptions as obligations on an installation.

4.1 The claim graph

A *claim* is one proposition together with the argument for it. It is the unit that sections 2.3 and 3.2 arrived at: one conclusion per warrant in the safety case. Most claims stand on other claims, and few are argued from evidence alone. A collision avoidance claim, for example, stands on a track quality claim and a target picture claim. Track quality in turn stands on position, and position stands on the claims of the receivers that contribute fixes to it. The claims therefore form a directed graph that runs from the vessel's evidence to its conclusions. This graph, which we require to be acyclic, is the *claim graph*. Panel (a) of figure 3 draws the fragment just described. Each arrow runs from a claim to a claim that stands on it. The claims a given claim stands on are its *grounds*: the grounds of collision avoidance are track quality and the target picture.

Claims without grounds, such as the claims of the receivers, are *source claims*. They stand on evidence alone: measurements and status reports from the vessel's equipment, such as the age of a receiver's latest fix and its dilution of precision.

At the top of the graph stand the modes of operation that the concept of operations describes (4.22, 8.6.1). Each mode is represented as a claim like any other, grounded in the capabilities that the mode requires. We call these *mode claims*. The mode claims are declared in an *order of activation*, and the final mode claim is a fallback state (8.4) that has no grounds and always stands.



(a) every condition established

(b) after the main GNSS source falls silent

Fig. 3 Claim-graph fragment before (a) and after (b) the main global navigation satellite system (GNSS) source falls silent. Arrows run from each ground to its dependent claim, and labels give the minimum accepted standing. Position within bound has a needed count of one and a wanted count of two. Full autonomy is a mode claim. The later mode claims in the order of activation are omitted.

4.2 The claim

A single claim holds Toulmin's argument in five fields. Three are settled when the claim is authored, and the other two are evaluated on every cycle. The example below is the position argument of figure 1, the node *Position within bound* in figure 3.

Statement. What the claim asserts, and at what strength. The position claim asserts that the vessel's position is known to within ten metres. Toulmin's qualifier becomes part of the statement, so a second bound requires a second claim.

Warrant. The general rule by which grounds of this kind support a statement of this kind. For the position claim, the warrant states that a fix stream supports a ten-metre bound while its coverage is full and its dilution of precision is below a set limit. Throughout this paper, *warrant* means this field of a claim.

Backing. Why the warrant holds, with the provenance of that reason, such as the type approval or simulation campaign that established the ten-metre bound. Backing is kept apart from the warrant because a warrant can remain in force after the backing for it has lapsed.

Conditions of application. The circumstances under which the warrant applies. Each condition is a predicate that the evaluator can check. It names one item of evidence and the values that establish it. For the position claim, the conditions are a dilution of precision inside the limit that the campaign covered and a fix younger than a stated age. In Verheij's reading of Toulmin's form, a claim follows from its grounds unless a rebuttal applies (Verheij 2005), so a rebuttal that goes unobserved leaves it standing, as the position did aboard the *Royal Majesty*. A condition of application reverses that default: the claim stands only while its evidence is present and in band. A condition whose evidence never arrived therefore fails in the same way as one whose value is rejected.

Grounds. The claims directly beneath this one. The position claim has two grounds, the claims of its two receivers. In Toulmin's scheme grounds are data (Toulmin et al. 1984). Here grounds name claims only, and evidence enters through the conditions of application. A measurement therefore supports a claim through the claim of the system that produced it.

4.3 Evaluation

On every cycle, the evaluator assigns each claim a *standing*. A claim is *licensed* when its argument holds in full and *withdrawn* when its argument does not hold. Between the two, a claim is *reduced* when it holds on fewer licensed grounds than full strength requires, for instance on one of two redundant receivers. The standings are ordered from licensed through reduced to withdrawn, and the label on each arrow in figure 3 states the lowest standing that the grounds must hold.

How many grounds must hold, and at what standing, is written in a *requirement*. A requirement names a set of grounds, a standing, and two counts. The *needed* count is the number of grounds that must hold that standing or better for the claim to stand at all. The *wanted* count is the number that must be licensed for the claim to stand at full strength. The position claim, for example, names its two receivers with a needed count of one and a wanted count of two. One licensed receiver is thus enough for the claim to stand, but the claim then stands reduced. The gap between the two counts is the *allowance*, and it is approved with the graph, so every allowance is visible to the assessor before the voyage. The counts also separate two kinds of grounds. For redundant grounds, such as the two receivers, the wanted count equals the number of members and the needed count is lower. For alternative grounds, where a set number of members suffices and further members add nothing, the two counts are equal, so losing a surplus member leaves the claim licensed. A claim is reduced exactly when it stands on an allowance (proposition 1), so its standing distinguishes normal operation from operation on a declared allowance.

The evaluator applies the rule below to one snapshot of the evidence, in one pass over the graph.

```
a requirement names its grounds, a standing s, and two counts, needed and wanted:
    met:              at least needed of the grounds hold s or above
    at full strength: at least wanted of the grounds are licensed

for each claim c, grounds before dependents:
    if some condition of application of c is not established by the evidence:
        standing(c) = withdrawn
```

```
    else if some requirement of c is not met:
        standing(c) = withdrawn
    else if every requirement of c is at full strength:
        standing(c) = licensed
    else:
        standing(c) = reduced

standing modes = the mode claims whose standing is not withdrawn
```

Figure 3 shows the rule applied to the fragment of section 4.1. In panel (a), every condition is established and every requirement is at full strength. All claims are therefore licensed, and full autonomy is in effect. In panel (b), the main global navigation satellite system (GNSS) receiver has fallen silent. Its fix-age condition is no longer established, and its claim is withdrawn. The auxiliary receiver still meets the position claim's needed count of one, but the wanted count of two is no longer met. Position therefore becomes reduced. Track quality requires position at reduced or better, so it becomes reduced as well. Collision avoidance requires licensed track quality, so it is withdrawn, and full autonomy is withdrawn with it. The response stage then activates the next standing mode in the order of activation, which the figure omits.

The rule is non-compensatory (Einhorn 1970). Collision avoidance in panel (b) is withdrawn whatever the standing of the target picture. A lost ground can be made up only inside a requirement, by the other members that the requirement declares interchangeable. A ground below the standing that its dependent requires withdraws that dependent, whatever the standings of unrelated grounds, and standings are never pooled across the graph into one figure of merit.

Because the graph is acyclic, an order that places grounds before dependents exists, and one pass in any such order settles every standing (proposition 3). The graph has no attack relation. In an abstract argumentation framework, attacks can leave several stable extensions or none, and a well-founded attack relation ensures exactly one (Dung 1995). Defeat enters only through conditions of application. A failed condition withdraws a claim by denying that its warrant applies, as an undercutting defeater does (Pollock 1987), and no claim in the graph asserts the opposite of another. Uniqueness then follows from the acyclicity of the graph, and recomputation from the record depends on it. A manual override therefore cannot set the standing of a claim directly. It enters as evidence that a condition reads.

The graph does not change during the voyage, and the rule reads one snapshot and keeps no state between cycles, so a standing changes only when the evidence changes. Quantities that depend on duration, such as the age of a fix or the publication rate over a window, must therefore enter as evidence, and a condition reads them like any other value. Staleness limits and plausibility gates likewise sit outside the graph, in the evidence layer (figure 2). As long as the evidence layer only removes values (obligation E2), it can only lower a standing (proposition 6) and shrink the set of standing modes (corollary 7).

Holds and smoothing also sit outside the graph, in the response stage. The graph leaves the data of the control path unchanged. A function may still receive a position whose basis has lapsed, but the mode whose claim stands on that position is no longer permitted.

4.4 Modes of operation

A mode claim stands on the capabilities that the mode requires and on the envelope in force. The envelope, which changes with area, traffic and voyage phase, enters the graph as evidence and determines which mode claims may stand. Since absent evidence withdraws a claim, a mode claim comes to stand only on affirmative support, while it can be withdrawn on missing evidence alone. Every supported envelope must therefore be represented when the graph is approved. An envelope that is not represented supports no mode other than the final fallback state.

The response stage puts one of the standing modes into effect in one of two ways. When the claim of the mode in effect is withdrawn, the response stage at once *activates* the next standing mode in the order. Otherwise a mode is *selected*, by the crew, the remote operator or the master, or by a default rule stated under the criteria of the concept of operations. Any standing mode may be selected, including one later in the order than the earliest standing mode, and a return to an earlier mode is always a selection. Since only standing modes are put into effect, the vessel operates only on allowances that the approved graph contains. Holds in the response stage act after evaluation, so the narrowing results of section 4.3 do not cover them, and a hold may delay a return to an earlier mode but not the activation that a withdrawal requires (obligation L1). A vessel is in a degraded state (4.9) when the mode in effect stands reduced, or when an activation has put into effect a mode other than a fallback state.

4.5 The warrant record

Each cycle produces one *warrant record* per claim. The record copies the claim's three settled fields unchanged. It adds what the evaluator read (the failed conditions with their evidence and the standing that each ground held) and the assigned standing. Figure 4 shows how the evaluation of one claim produces its record. The record alone therefore supports inspection (table 3): it shows what the claim asserted, on what warrant, and why it received its standing. Beside the warrant records, the response stage is expected to report each mode it puts into effect, and whether the mode was activated or selected and by whom (obligation L2). The record thus holds both the modes that the evidence permitted and the mode in which the vessel operated.

Since the standings depend only on the graph and the evidence (proposition 3), a reader who holds both can recompute any decision. A log that contains only outcomes supports none of inspection, recomputation and verification, since it holds neither the structure nor the evidence that they require. A non-compensatory rule whose log holds only its outcome supports none of them either, so the aggregation rule and the content of the record are separate design choices.

The record shows the basis of a decision only if the vessel evaluated the graph that the assessor approved. Verification establishes this by comparing the digest named in the record with the digest of the approved graph, so the record identifies the graph in force without storing its conditions and requirements. Verification is independent of recomputation, because an evaluator may apply an unapproved graph correctly.

Verification covers the graph but not the processing that produces the evidence. Each condition of application reads one item of evidence through one predicate

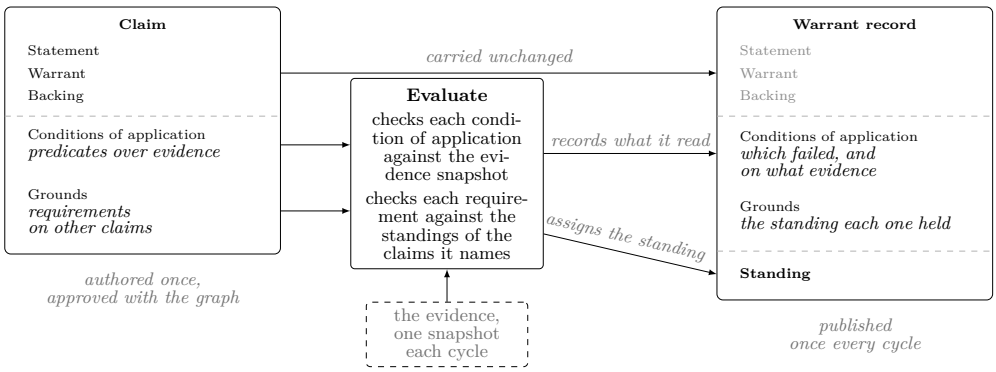


Fig. 4 How the evaluation of one claim produces its warrant record. The three settled fields are above the dashed line and the two conditional fields below it.

(appendix A), so a relation between two items cannot be written as a condition. The cross-check that the *Royal Majesty*'s watch did not make is such a relation: it compares the position from one source with the position from another. The evidence layer must compute it and publish the result as a further item, and the graph then holds only a predicate on that result. The record names each value that the evaluator read, but neither the record nor the digest identifies how the value was computed. An installation can extend the digest to the configuration of the evidence layer, so that verification also covers its limits and acceptance bands, which the implementation of section 5 does not do.

4.6 The five functions and existing mechanisms

The architecture provides the five functions of section 2.3. The claim graph is the justification held aboard, evaluated without anything from ashore, and its conditions of application are the withdrawal conditions, written as predicates over current evidence (functions 1 and 2). The warrant records of each cycle tie the evidence and the standings of the grounds to the conclusion they support, and in the same pass a withdrawal reaches every dependent claim whose requirements it leaves unmet (functions 3 and 4). The sequence of warrant records and the digest of the graph in force support inspection, recomputation, and verification (function 5). Putting a standing mode into effect is left to the response stage, as section 2.3 required.

Table 6 sets the architecture beside seven mechanisms. We selected those that evaluate an element of an assurance argument, or a risk model in place of one, while the system operates: conditional safety certificates (ConSerts) (Schneider and Trapp 2013), digital dependability identities (Schneider et al. 2015), dynamic safety cases (Denney et al. 2015), dynamic assurance cases (Asaadi et al. 2020), ENTRUST (Calinescu et al. 2018), supervisory risk control (Utne et al. 2020; Johansen and Utne 2022), and the monitoring clauses of UL 4600 (Underwriters Laboratories 2023). ConSerts, dynamic assurance cases, ENTRUST and supervisory risk control reach their determination aboard, and dependability identities may be evaluated aboard or ashore. Dynamic safety cases update the argument automatically or after human

review, depending on the change, and UL 4600 does not state where its safety performance indicators (SPIs) are evaluated. The last column of table 6 lists the record that each mechanism leaves. ENTRUST keeps its assurance argument and the verification evidence for it after each reconfiguration, and UL 4600 requires SPI violations and metric data to be collected and tied to the claim that each SPI conditions. The guidance on Safety Assurance of autonomous systems in Complex Environments (SACE) (Hawkins et al. 2022) requires a documented approach for determining, during operation, whether the system is leaving its defined operating context (stage 7), which the conditions of application provide here, but it sets no runtime evaluation of the case itself.

The contract-based frameworks used for maritime autonomy (Glomsrud et al. 2024) address a different question. A contract specifies how a component behaves, whereas a warrant specifies whether the justification for relying on a claim still holds. A component may therefore satisfy its contract after the evidence that it does so has become unavailable.

The components of the architecture are inherited: Toulmin's layout, Verheij's conditional reading of the warrant, and a non-compensatory evaluation of the kind that ConSerts apply to guarantees. The contribution of the architecture is to combine them into a claim graph whose evaluator produces, for every claim on every cycle, the conditions that failed and the standing that each ground held.

5 Mechanistic illustration

We ran the architecture of section 4 against a staged reproduction of the receiver failure aboard the *Royal Majesty* (section 1) to establish what information an assessor can recover from its record, and whether that information depends on the aggregation rule. The illustration assesses neither whether the published mode was appropriate nor any safety outcome.

Since the graph, its conditions and the staged evidence are all authored inputs, the illustration shows that the implementation follows the rule of section 4.3. It does not show that the graph is complete. Section 5.4 shows what the logs of two composite-score policies, run on the same evidence, contain at the same instant. One of them is non-compensatory, like the claim graph, so any difference between its log and the warrant record comes from what each records, not from how each combines the sources.

5.1 Setup

We implemented the architecture of section 4 on an open maritime middleware stack (reference withheld for review). The emulated vessel has twelve sources, including two GNSS receivers and a radar, which stand in for the sensors and monitors of figure 2. A health monitor forms the evidence layer and rates each channel against its expected publication rate and the bands of its values. It reports a channel as inactive once the channel has been silent for three seconds. The evaluator is the middleware's warrant aggregator, and a recorder writes the data log. The illustration has no vessel functions, actuators or response stage. In place of a response stage, the aggregator publishes the first standing mode in the order of activation, and the recorder records

	What is evaluated in operation	What it permits or changes	Aggregation	Record left behind
ConSerts (Schneider and Trapp 2013)	guarantee–demand match	the guarantee variant that holds	Boolean, non-compensatory	none specified
Dependability identities (Schneider et al. 2015)	dependability model	integration and degradation level	depends on the model chosen	none specified
Dynamic safety cases (Denney et al. 2015)	confidence in claims, from monitored assurance variables	update of the argument or the system, automated or after review	Bayesian confidence, update rules partly automated	updated argument
Dynamic assurance cases (Asaadi et al. 2020)	confidence in assurance properties of the system	risk mitigation during the mission	probabilistic measure	none specified
ENTRUST (Calinescu et al. 2018)	probabilistic requirements, by runtime quantitative verification	configuration of the system, or failsafe	thresholds on requirements, then weighted cost	argument and verification evidence after each reconfiguration
Supervisory risk control (Johansen and Utne 2022)	online risk metric	control mode, speed, machinery	probabilistic, compensatory	none specified
UL 4600 SPIs (Underwriters Laboratories 2023)	indicator values	response to violations, including revision of the case	threshold on each indicator	SPI violations and metric data, per claim
This architecture	claim standing	the standing modes, from the surviving claims	non-compensatory, propagating	transitions and periodic snapshots containing the basis of each determination

Table 6 Comparison with seven runtime assurance mechanisms. ConSerts, dependability identities and dynamic safety cases cite runtime certification (Rushby 2008), which proposed that part of a certification argument be discharged by monitoring in operation.

it. No component acts on the published mode. The aggregator also provides a hold on returns to an earlier mode, which section 4.4 places in the response stage. The illustration runs with the hold at zero. A return to an earlier mode is a selection in the architecture, and the aggregator makes it by default as soon as the claim of that mode stands again.

We staged the failure's *signature*, that is, the evidence as the integrated bridge system received it. After the receiver lost its fix, its dead-reckoned positions arrived on the same interface and in the same sentences as before, and the bridge system did not read the flag that marked them invalid (National Transportation Safety Board 1997, pp. 17 and 36). On the main receiver of the emulated vessel, from 120 seconds into the session, the position stream therefore continues at its nominal rate with plausible positions, and its quality channel continues to report a full three-dimensional fix. The staging places the loss in the satellite-count and dilution-of-precision channels, which fall silent until they resume at 480 seconds, so the fix stream itself does not show it.

Figure 5 draws the complete claim graph. It extends the fragment of figure 3 with a track-keeping claim and declares five mode claims in their order of activation. The conditions of the main receiver's claim read the two channels in which the staging places the loss. The graph holds one envelope throughout and assumes a working shore link.

The five modes are a subset chosen for the illustration. A real concept of operations defines its own modes for each voyage phase (8.6.1), and they may be more numerous or different. Under full autonomy, the vessel performs collision avoidance and track keeping. Under assisted autonomous, it still avoids collisions on its own, while a remote operator monitors a track kept on a reduced position. Under remote controlled, the remote operator navigates, with the onboard collision avoidance and position as aids. Under supervised remote, the vessel keeps to a restricted manoeuvre, such as holding position or proceeding at reduced speed, on its position or its target picture, while the remote operator monitors it. Minimal safe mode is the only fallback state.

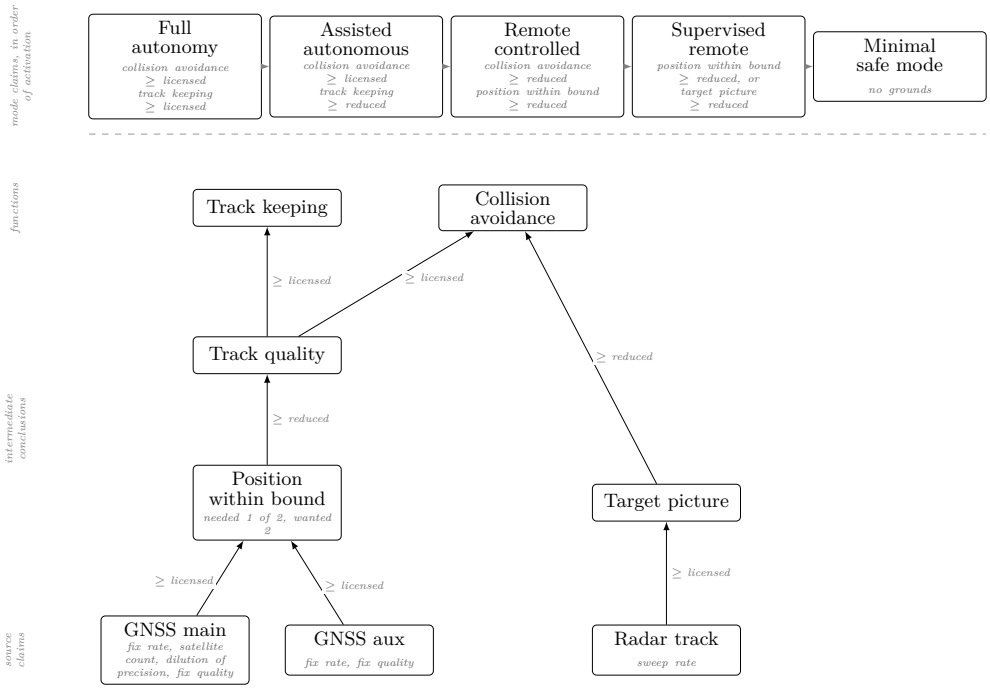


Fig. 5 The complete graph used in the illustration. Arrows run from each ground to its dependent claim and labels give the minimum accepted standing; source claims list the evidence their conditions of application read. The mode claims above the dashed line are shown in their order of activation. Their arrows are omitted, and each states its requirements in text.

5.2 The evaluation

During the staged degradation, the claim graph withdrew the main receiver's claim because two of its conditions of application could not be established: the satellite-count and dilution-of-precision evidence was absent. The withdrawal propagated as in figure 3(b), and track keeping was withdrawn together with collision avoidance. Of the mode claims, only supervised remote and minimal safe mode still stood, supervised remote on either of its grounds, the reduced position claim and the licensed target picture. With full autonomy in effect, the next standing mode in the order of activation was also the first, and the aggregator therefore published supervised remote, as an activation would. The withdrawal took effect at 124 seconds, once the channels had been silent for longer than the evidence layer's limit. The aggregator returned to full autonomy at 487 seconds, once the channels had resumed long enough for the evidence layer to report them active again. This return was the default selection of section 5.1.

5.3 The record

A recording of the session contains the evidence, the determinations, every standing transition, periodic snapshots, and a marker for the staged degradation. Each snapshot names the digest of the graph in force. The recording does not store a warrant record for every cycle. Standings change only at transitions, and the recorded evidence and the graph determine the record of any cycle by recomputation (proposition 3). The transitions and periodic snapshots are therefore enough to recover every evaluation that function 5 requires.

Figure 6 shows the warrant record in force at $t = 300$ seconds. The implementation predates the terms of this paper. It prints the activated mode as the `Licensed level` and a failed condition of application as a `rebuttal fired`. Since no transition occurred after the snapshot at 295.4 seconds, this record is that snapshot. It contains the chain of standings from track keeping down to the receivers. For the main receiver, it names the two failed conditions and the evidence on which each failed. At the snapshot, both channels had been silent for 175 seconds, against a limit of three.

The record supports the three operations of table 3. Inspection gives figure 6 from the snapshot at 295.4 seconds, the latest before $t = 300$ seconds, and the transitions after it. Recomputation from the graph and the recorded evidence reproduces every transition and activated mode in the session. Verification finds that the digest named in each snapshot equals the digest of the authored graph. Since the author and the assessor are the same party here, the match shows only that the record supports the comparison. It does not show that an approval process was followed.

5.4 Record contents under two composite-score policies

The recording also holds the determinations of two composite-score policies from the same middleware release. Both read the same evidence and publish one of the same five modes. The *compensatory composite* averages the health scores of the twelve declared sources and compares the mean with a threshold for each mode. The *essential-flagged composite* caps that score at the health score of any source marked essential, which

```
Licensed level: SUPERVISED_REMOTE   (t = 295.4 s)

[WITHDRAWN] track_keeping: the vessel holds its planned track
  warrant: Track keeping requires track quality at full strength: [...]
  backing: staged demonstration configuration; NTSB/MAR-97/01 as the
           negative case
  ground: track_quality [REDUCED]

[REDUCED] track_quality: the vessel's track over ground is maintained
          within corridor
  warrant: [...]
  ground: position_within_bound [REDUCED]

[REDUCED] position_within_bound: own position is known to within ten metres
  warrant: Two independently licensed receivers support the bound at full
           strength; a single licensed receiver supports it at reduced
           strength.
  backing: redundancy argument over the staged sensor fit
  ground: gnss_main_fix [WITHDRAWN]
  ground: gnss_aux_fix [LICENSED]

[WITHDRAWN] gnss_main_fix: gnss_main's fixes support a ten-metre position
            bound
  warrant: A fix stream at nominal rate, with current satellite-tracking
           evidence [...] supports a ten-metre bound.
  backing: emulated receiver in the staged session; bands in
           config/sensor_fit.yaml
  rebuttal fired: satellite_evidence_absent, satellite-tracking evidence
                  absent or out of band
    evidence: gnss_main/location_fix_satellites_used INACTIVE: rate 0.0 Hz;
              activity:INACTIVE (silent for 175.3s (limit 3.0s))
  rebuttal fired: dop_evidence_absent, dilution-of-precision evidence
                  absent or out of band
    evidence: gnss_main/location_fix_hdop INACTIVE: rate 0.0 Hz;
              activity:INACTIVE (silent for 175.3s (limit 3.0s))

[LICENSED] gnss_aux_fix: gnss_aux's fixes support a ten-metre position bound
  warrant: A fix stream at nominal rate with self-reported quality at a full
           3D fix supports a ten-metre bound; the auxiliary receiver exposes
           no satellite-tracking side channels, so its warrant rests on fewer
           grounds and is licensed accordingly.
  backing: emulated receiver in the staged session; bands in
           config/sensor_fit.yaml
```

Fig. 6 The warrant record in force at $t = 300$ seconds in the staged case, read from the recording. Long lines are wrapped, and elisions are marked [...]. The claims for collision avoidance, the target picture and the radar are omitted. The two failed conditions that withdraw the main receiver, and through it full autonomy, are set in bold.

makes it non-compensatory. Their determinations follow from these authored rules (table 7).

With one of the twelve sources at zero, the compensatory composite scores 0.92 against a full-autonomy threshold of 0.85 and keeps full autonomy, while the essential-flagged composite publishes minimal safe mode. The aggregation rule thus changes the determination but not what an assessor can recover from the log. Inspection of either log shows the failed source, but not the vessel capabilities that the failure withdrew or their grounds. Rerunning a composite on the recorded source assessments reproduces its published modes, and each determination names the digest of its policy, so both logs support recomputation and verification. Since the composites evaluate no conclusions, neither operation recovers a basis. The log of the essential-flagged composite therefore

Policy	Selection rule	Published mode	Recorded basis
Compensatory composite	score against mode thresholds	full autonomy, score 0.92	score and twelve source assessments
Essential-flagged composite	same score, capped by the essential source	minimal safe mode	failed source and cause of cap
Claim graph	first standing mode in the order of activation	supervised remote	claim standings, warrants, grounds, failed conditions and evidence

Table 7 Determinations and record contents of the claim graph and the two composite-score policies during the staged degradation.

holds no more of the basis than that of the compensatory composite, although its rule is non-compensatory (section 4.5).

6 Discussion

6.1 The claim graph in the Code's terms

Three parts of the architecture have counterparts in the Code. A function's ODD is expressed as the conditions of application on the claim that the function grounds, so checking those conditions is the ODD detection that clause 8.3.2.2 requires. The set of standing modes represents the determination under clause 8.3.3 of whether the vessel can remain within its OE. The warrant record holds the basis that clause 9.10.4 requires under its strong interpretation, as far as the evidence that the evaluator read (section 4.5). It also holds what clauses 13.3.3 and 13.3.6 require of an alert: its consequences are the withdrawn claim and the conclusions that depend on it, and its cause is the conditions and evidence that withdrew it. Acknowledging the alert leaves the standing unchanged. Aboard the *Marco Polo*, the alarm text disappeared on acknowledgement while the loss of position persisted (Swedish Accident Investigation Authority 2025, p. 36).

The Code defines a mode of operation by whether the functions are performed autonomously or remotely (4.22), leaves the modes and the criteria for changing between them to the concept of operations (8.6.1), and requires the vessel to enter a predefined fallback state when it deviates from its OE (8.4.1). The MASS Safety Certificate is supplemented by a Record of MASS, a form in the appendix to the Code. It lists the modes of operation for each voyage phase and each fallback state with its conditions of entry, and where several fallback states are available it states the order in which they are activated (1.4 and 1.5). Representing each mode as a claim and extending the order to all modes is our construction. In the graph, each mode of the concept of operations is a mode claim, and so is each fallback state. A fallback state is grounded in the capabilities that its fallback response requires, and only the final mode claim has no grounds. The non-compensatory rule is also our choice among the rules that clause 8.3.3 permits (section 3.1).

The Code permits more than one fallback state (8.4.1) and, depending on the risk assessment, requires more than one independent fallback state to be available at any time during normal operation (8.4.4). In the graph, the availability of each is its

standing, which the vessel evaluates on every cycle. The order of activation expresses the hierarchy among the fallback states, but not conditions of entry that depend on which failure occurred. Such conditions are left to the response stage as default rules of selection, together with the fallback response (8.4.2, 8.4.3) and the notification of the crew and the remote operators (8.4.5). Since the final mode claim always stands, the graph cannot show a deviation beyond the final fallback state, and the determination that contingency plans are needed (8.5) lies outside it.

The Code does not prescribe a graph. A lookup table, a machine-readable safety case or annotated control flow can provide the five functions, provided the mechanism records the reasoning state with each result. Guidance for unmanned naval craft recommends precomputed responses (European Defence Agency 2022). However, a table that records only the selected case does not meet the strong interpretation of clause 9.10.4, because its justification remains in the design documentation. We chose a graph because it composes claims and so avoids enumerating every combination of component states and contexts.

6.2 Approval, change and records

The graph belongs to the approval basis. Its conditions of application and requirements are evaluation criteria, which are agreed with the Administration (6.2). The certificate names the Record of MASS by document and version number. The mode claims state the content of the Record in a form that the vessel evaluates, and the digest in each warrant record identifies the graph version as the version number identifies the Record. At assessment, an Administration can select an instant from a trial or voyage and apply the three operations of table 3 to the decision taken then.

In service, a change to the graph falls under management of change. The Code requires the affected part of the risk assessment to be repeated when the concept of operations changes (6.4.2) and leaves the procedure to others. For computer-based systems it refers to Unified Requirement E22 of the International Association of Classification Societies, which requires documented management of change, including changes to configurable parameters (International Association of Classification Societies 2023, 6.2 and 6.11). The International Safety Management (ISM) Code requires changes to the documents of the safety management system to be approved by authorised personnel (International Maritime Organization 1993, 11.1 and 11.2). A warrant record that names a digest other than the approved one shows that an unapproved graph was in force.

The master, and those responsible for managing the operation, can intervene in or override the system (8.7.1.3, 8.7.3). An intervention leaves the graph unchanged. It either selects a standing mode, which the response stage reports to the record (obligation L2), or enters as evidence that a condition of application reads (section 4.3).

The voyage data recorder, which clause 9.10.6 applies to every MASS, records the status and output of the ship's equipment. None of its data items records the basis of an automated decision (International Maritime Organization 2012, 5.1.1 and 5.5). Under the Casualty Investigation Code, investigators should have ready access to all recorded data (International Maritime Organization 2008, 16.4.2). Of the data recorded by the ship's equipment, its definition of a marine safety record names only

the information from the voyage data recorder (International Maritime Organization 2008, 2.15). Its recommended practices state that such records should not be disclosed in criminal, civil, disciplinary or administrative proceedings except under stated conditions (International Maritime Organization 2008, 23.2).

6.3 Recommendations to the experience-building phase

The experience-building phase collects the experience of Member States and industry in applying the Code (International Maritime Organization 2026b). We offer it a recommendation on the three interpretations that section 3.2 left open, three questions that apply to every vessel approved under the Code, and four matters that arise once a vessel carries a mechanism that provides the five functions of table 2.

Interpretations of clauses 8.3.3, 9.10.4 and chapter 10 of the Code.

We recommend the strong interpretations of clauses 8.3.3 and 9.10.4 and the operational interpretation of chapter 10 (table 5). Under them, the vessel determines aboard whether it can remain within its OE after a function leaves its ODD, records the basis of each determination, and keeps each decision made at sea explainable from that basis. The OE lists limits on area, coverage, traffic and weather (8.3.1.2). The design documents state these limits, but whether the vessel is within them at a given instant can be determined only during the voyage. A design that falls back after every deviation can still monitor the limits with thresholds. Under clause 8.3.3, a function that has left its ODD puts the vessel in a degraded state as long as the integrated vessel can still be operated within its OE. Such a design never evaluates that condition, so none of its deviations leads to a degraded state. Under the weak interpretation of clause 9.10.4, the log holds the inputs, and rerunning the software on them reproduces each result. The reason for the result must still be reconstructed from the design documents, as the investigators of the *Royal Majesty* had to do for the mechanism of the failure. Both composite logs in the illustration met that interpretation, and neither showed which capabilities the failed source withdrew. The warrant record held the basis of the determination (figure 6, table 7). Clause 9.10.5 gives the investigation authorities access to the record, so under the strong interpretation they get the basis with it. Documentation written before the voyage, which the second sentence of clause 10.4 names, cannot explain a decision whose basis arose during the voyage.

Open questions.

1. Which interpretation of "restore" in clause 9.10.4 Administrations apply. Investigations should record whether the basis of each automated decision could be recovered from the log or had to be inferred from design documents. The same record settles the chapter 10 question, since a decision whose basis can be recovered only from documentation is explained under the governance interpretation alone.
2. Whether approved vessels exercise the non-propagation that clause 8.3.3 permits or enter fallback after every deviation, and, where they exercise it, which aggregation rule the Administration accepted. Administrations should record the rule on each approved vessel and whether each component-level deviation propagated

to the vessel level. In the illustration, one evidence stream produced three modes under three rules that the clause permits (table 7), so data from several vessels can be compared only if the rule on each is recorded.

3. How often an unnamed proviso, one that the safety case does not state, fails during operation. Any runtime mechanism checks only the provisos represented in it. Reports should therefore identify cases in which a vessel continued after an unrepresented proviso failed. Where the justification is machine-readable, such a case is a claim that remained standing on evidence that should have withdrawn it. Where people discharge the provisos, it is an intervention prompted by a proviso absent from the safety case. These cases can inform later claim graphs and revisions of the instrument. The illustration contains an unnamed proviso, although it does not fail: the auxiliary receiver publishes no satellite count or dilution of precision, so a dead-reckoned fix on it would have kept its claim licensed, and its warrant text (figure 6) states the gap before any failure.

Matters for a mechanism that provides the five functions.

1. Whether a change to the conditions alone, with unchanged software, is a modification of the concept of operations or a software update that may call for an additional survey (5.2.2.4), and whether an approval can define a range within which the company may make such changes.
2. Whether the master may amend or suspend the justification held aboard, or only select among the modes it permits, and who approves a change when the master is at a remote operation centre (14.2.1.11).
3. How long the records of the evaluations, and the versions of the justification that they name, must be kept. Clause 9.10.2 sets at least thirty days, and no instrument states how long a version of the justification remains available.
4. Whether these records are marine safety records under the Casualty Investigation Code.

6.4 Limitations and further work

The derivation of the five functions and the reading of the Code are arguments from text.

- The five functions follow from the properties of a runtime discharge (section 2.2). The argument shows that a mechanism carrying the runtime half aboard needs each of them, but not that the five are sufficient, and we have not compared them with the safety case of an approved vessel.
- The reading of the Code (section 3) is our own and covers the clauses that bear on the five functions. No interpretive practice exists yet (section 1), and the mandatory instrument may word these clauses differently.

The mechanism has three limitations.

- The claim graph is acyclic and has no attack relation (section 4.3) and no quantitative qualifiers, so each claim states one bound and a second bound on the same quantity needs its own claim. The evaluated standing remains three-valued.
- The graph was authored by hand for the staged case. We demonstrate no method for translating a safety case, concept of operations or risk assessment into claims and predicates. The wellformedness conditions of appendix A do not assess coverage either, and the auxiliary receiver in section 6.3 is one instance of that gap. A circumstance that spans several items of evidence (a limit that depends on the mode of operation, for example) needs a claim of its own or an item computed by the evidence layer, and the number of claims needed for a whole vessel has not been measured. Nor have we examined how authoring and maintenance would be divided among suppliers, the integrator and the Administration.
- Record integrity and part of the record's provenance are outside scope. The illustration establishes neither tamper evidence nor that the recording is complete. The record lists failed conditions and their evidence but omits the conditions that were established, and the graph digest does not cover the evidence layer (section 4.5).

The illustration uses staged evidence. Its backing fields identify the emulated setup, and the recording includes the degradation profile and its schedule.

- The injected degradation has a known onset and shape. This makes the architecture's response observable but excludes the variability of service conditions. A single run cannot estimate detection performance, the rate of false withdrawals, variance or sensitivity to multipath.
- The architecture is instantiated once, on one middleware stack, and that stack's health vocabulary and subject granularity may have shaped the design. Appendix B states the obligations on the surrounding layers, but no second stack has been checked against them.
- The graph holds one envelope throughout and assumes an available shore link (section 5.1), although connectivity is a proviso in any safety case for remote operation (table 2). It includes no claim emitted by a learned component with statistical backing.

One line of further work extends the graph to the remote operation centre. A remotely operated vessel has provisos on the link, on the remote operator's situational awareness and on the transfer of control, and the Code sets requirements on each. If these provisos were represented as claims, the graph could state at a handover which conclusions the receiving party takes over and on what evidence. It would also leave a record of the transfer.

A further line concerns learned components. The design range of such a component is largely set by the data on which it was trained and validated, and the proviso that its current input resembles those data is not a threshold on a value the vessel publishes. A runtime mechanism can check this proviso only through an out-of-distribution monitor, which is itself a statistical estimate with provisos of its own.

The guidance on the assurance of machine learning in autonomous systems (AMLAS) covers the assurance argument for a learned component from its requirements to its deployment. Its deployment stage already requires the assumptions of the component to be monitored in operation, including whether its inputs remain close to the data on which it was trained, and its erroneous outputs to be logged (Hawkins et al. 2021, stage 6). How a condition of application should be written for such a component, and what the record should hold when the basis of a conclusion is a statistical estimate, apply to any mechanism that provides the five functions, and we leave them open.

A third line concerns the people who read the record. The record holds what clauses 13.3.3 and 13.3.6 require of an alert (section 6.1), but we designed no presentation of it and ran no operator evaluation. Shown to the crew aboard or to a remote operator, the record could give the basis of a recommendation or a decision in decision support systems and partially autonomous vessels as well as in autonomous ones. Human factors research could establish whether such a presentation supports the understanding of consequences and cause that the two clauses require, and whether it helps people judge when to rely on the system.

7 Conclusions

This paper named the runtime half of a safety case: the discharge in operation of the provisos that a design-time argument leaves open, by checking them and responding when one fails. The bridge team discharges them on a conventional ship, and an autonomous vessel without crew or shore link must discharge them itself. From the properties of a discharge in operation we derived five functions that an onboard mechanism must provide: hold the justification aboard, state the conditions for withdrawing a conclusion as predicates, tie each conclusion to its evidence at each determination, propagate each withdrawal to the conclusions that depend on it, and record every evaluation (table 2). We identified the individual conclusion as the unit of evaluation.

Five clauses of the MASS Code address these functions. Clause 8.3.2.2 requires on any reading that the vessel detect whether each function is within its ODD. Clauses 8.3.3, 9.10.4, 10.4 and 10.5 require the remaining functions only under three interpretations that the text admits but does not compel. The Code does not state how component deviations combine at the level of the vessel, nor how the justification is to be represented.

One mechanism can provide all five functions. In the proposed claim graph, each claim carries Toulmin's statement, warrant and backing, together with grounds read as in Verheij's conditional account and conditions of application that reverse its treatment of rebuttals. A non-compensatory evaluator assigns one of three ordered standings to each claim and in the same pass withdraws every dependent claim whose requirements the withdrawal leaves unmet. The modes whose claims remain standing are those the vessel may operate in, and the response stage puts one of them into effect. Every allowance on which a mode can stand is written into the graph and so is visible to the assessor before the voyage. Each cycle produces a warrant record per claim, holding the failed conditions with their evidence and the standing that each ground held.

We ran the implementation against the *Royal Majesty* signature on an emulated vessel. The graph withdrew the main receiver's claim, the withdrawal reached full autonomy, and the record supported inspection, recomputation and verification against the approved graph. An Administration can make these three checks for any instant of a trial or voyage. Verification covers the graph but not the evidence layer that computes its inputs. The logs of two composite-score policies run on the same evidence hold the failed source but not the conclusions it withdrew (table 7), although one of the two rules is non-compensatory. What an assessor can recover therefore depends on what the record holds, not on the aggregation rule.

For the experience-building phase, we recommend the three interpretations. Whether the vessel is within its operational envelope can be determined only during the voyage, and under the weaker interpretations the design documents stand in for that determination. We also offer the clause-to-function mapping, the implementation, three questions and four matters (section 6.3). The questions are which interpretation of "restore" Administrations apply, whether approved vessels continue in a degraded state after a component deviation and under which aggregation rule, and how often a proviso absent from the safety case fails in operation. The matters concern a vessel that carries a mechanism providing the five functions: how a change to its conditions is approved, who may amend the justification held aboard, how long its records and the versions they name are kept, and whether these records are marine safety records under the Casualty Investigation Code.

We have not compared the five functions with the safety case of an approved vessel, and the illustration used staged evidence (section 6.4). Further work is to extend the graph to the remote operation centre, so that at a handover it states which conclusions the receiving party takes over. It is also to write conditions of application for learned components, whose fit to the current input can be checked only by a statistical monitor. Trials with operators can show whether a display of the record lets the crew or a remote operator see which capabilities a failure withdrew and which conditions withdrew them.

Statements and Declarations

Use of AI tools. Claude (Anthropic) was used to draft and revise text, review the manuscript and develop parts of the illustration code. The authors reviewed all output and take full responsibility for the content.

References

Asaadi E, Denney E, Menzies J, et al (2020) Dynamic assurance cases: A pathway to trusted autonomy. Computer 53(12):35–46. https://doi.org/10.1109/MC.2020.3022030

ASTM International (2021) Standard practice for methods to safely bound behavior of aircraft systems containing complex functions using run-time assurance. Tech. Rep. F3269-21, ASTM International

Bainbridge L (1983) Ironies of automation. Automatica 19(6):775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Bloomfield R, Rushby J (2022) Assessing confidence with Assurance 2.0. Technical Report SRI-CSL-2022-02, Computer Science Laboratory, SRI International, https://doi.org/10.48550/arXiv.2205.04522, also available as arXiv:2205.04522, revised May 2024

Bloomfield R, Netkachova K, Rushby J (2024) Defeaters and eliminative argumentation in Assurance 2.0. Technical Report SRI-CSL-2024-01, Computer Science Laboratory, SRI International, also available as arXiv:2405.15800

Bloomfield RE, Netkachova K (2014) Building blocks for assurance cases. In: IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), Naples, Italy, pp 186–191, https://doi.org/10.1109/ISSREW.2014.72

Calinescu R, Weyns D, Gerasimou S, et al (2018) Engineering trustworthy self-adaptive software with dynamic assurance cases. IEEE Transactions on Software Engineering 44(11):1039–1069. https://doi.org/10.1109/TSE.2017.2738640

Denney E, Pai G, Habli I (2015) Dynamic safety cases for through-life safety assurance. In: IEEE/ACM 37th International Conference on Software Engineering (ICSE), New Ideas and Emerging Results track, pp 587–590, https://doi.org/10.1109/ICSE.2015.199

Dung PM (1995) On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence 77(2):321–357. https://doi.org/10.1016/0004-3702(94)00041-X

Einhorn HJ (1970) The use of nonlinear, noncompensatory models in decision making. Psychological Bulletin 73(3):221–230. https://doi.org/10.1037/h0028695

European Defence Agency (2022) Best practice guide for unmanned maritime systems: Handling, operations, design and regulations. Tech. rep., European Defence Agency, ad hoc working group Safety and Regulations for Unmanned Maritime Systems (SARUMS), SARUMS BPG 2022, public version

Fjørtoft KE, Rødseth ØJ (2020) Using the operational envelope to make autonomous ships safer. In: Proceedings of the 30th European Safety and Reliability Conference (ESREL2020) and 15th Probabilistic Safety Assessment and Management Conference (PSAM15). Research Publishing, Singapore, https://doi.org/10.3850/978-981-14-8593-0_3496-cd

Glomsrud JA, Kemna S, Vasanthan C, et al (2024) Modular assurance of an autonomous ferry using contract-based design and simulation-based verification principles. Journal of Physics: Conference Series 2867(1):012043. https://doi.org/10.1088/1742-6596/2867/1/012043, preprint arXiv:2408.03244

Hawkins R, Paterson C, Picardi C, et al (2021) Guidance on the assurance of machine learning in autonomous systems (AMLAS). Tech. rep., Assuring Autonomy International Programme, University of York, https://doi.org/10.48550/arXiv.2102.01564, also available as arXiv:2102.01564

Hawkins R, Osborne M, Parsons M, et al (2022) Guidance on the safety assurance of autonomous systems in complex environments (SACE). Tech. rep., Assuring Autonomy International Programme, University of York, version 1, July 2022. Also available as arXiv:2208.00853v1

IEEE (2019) IEEE standard for floating-point arithmetic. Tech. Rep. IEEE Std 754-2019, IEEE, https://doi.org/10.1109/IEEESTD.2019.8766229

International Association of Classification Societies (2023) Computer-based systems. Unified Requirement E22 Rev.3, International Association of Classification Societies, June 2023; applies to ships contracted for construction on or after 1 July 2024

International Maritime Organization (1993) International management code for the safe operation of ships and for pollution prevention (ISM code). Resolution A.741(18), as amended, International Maritime Organization, consolidated text in force from 1 January 2015

International Maritime Organization (2008) Code of the international standards and recommended practices for a safety investigation into a marine casualty or marine incident (Casualty Investigation Code). Resolution MSC.255(84), International Maritime Organization, Parts I and II mandatory under SOLAS regulation XI-1/6 from 1 January 2010

International Maritime Organization (2012) Revised performance standards for shipborne voyage data recorders (VDRs). Resolution MSC.333(90), International Maritime Organization

International Maritime Organization (2017) STCW: International Convention on Standards of Training, Certification and Watchkeeping for Seafarers, Including 2010 Manila Amendments. International Maritime Organization, London, 2017 edition

International Maritime Organization (2026a) International code of safety for maritime autonomous surface ships (MASS code). Resolution MSC.595(111), International Maritime Organization, adopted May 2026, in effect 1 July 2026. Issued as MSC 111/22 annex 16

International Maritime Organization (2026b) Maritime safety committee (MSC 111), 13–22 May 2026: highlights. URL https://www.imo.org/en/mediacentre/meetingsummaries/pages/temporary-msc-111.aspx, Meeting summary, including the revised road map for the goal-based MASS instrument; accessed 11 September 2026

International Organization for Standardization and International Electrotechnical Commission (2020) Conformity assessment: Vocabulary and general principles. Tech. Rep. ISO/IEC 17000:2020, ISO/IEC, second edition

Johansen T, Utne IB (2022) Supervisory risk control of autonomous surface ships. Ocean Engineering 251:111045. https://doi.org/10.1016/j.oceaneng.2022.111045

Lützhöft MH, Dekker SWA (2002) On your watch: Automation on the bridge. Journal of Navigation 55(1):83–96. https://doi.org/10.1017/S0373463301001588

National Transportation Safety Board (1997) Grounding of the panamanian passenger ship Royal Majesty on Rose and Crown Shoal near Nantucket, Massachusetts, June 10, 1995. Marine Accident Report NTSB/MAR-97/01, National Transportation Safety Board, Washington, DC, URL https://www.ntsb.gov/investigations/AccidentReports/Reports/MAR9701.pdf

Parasuraman R, Riley V (1997) Humans and automation: Use, misuse, disuse, abuse. Human Factors 39(2):230–253. https://doi.org/10.1518/001872097778543886

Pollock JL (1987) Defeasible reasoning. Cognitive Science 11(4):481–518. https://doi.org/10.1207/s15516709cog1104_4

Rødseth ØJ, Lien Wennersberg LA, Nordahl H (2022) Towards approval of autonomous ship systems by their operational envelope. Journal of Marine Science and Technology 27(1):67–76. https://doi.org/10.1007/s00773-021-00815-z, published online 2021

Rushby J (2008) Runtime certification. In: Runtime Verification (RV 2008), Lecture Notes in Computer Science, vol 5289. Springer, Berlin, Heidelberg, pp 21–35, https://doi.org/10.1007/978-3-540-89247-2_2

Schneider D, Trapp M (2013) Conditional safety certification of open adaptive systems. ACM Transactions on Autonomous and Adaptive Systems 8(2):1–20. https://doi.org/10.1145/2491465.2491467

Schneider D, Trapp M, Papadopoulos Y, et al (2015) WAP: Digital dependability identities. In: IEEE 26th International Symposium on Software Reliability Engineering (ISSRE), pp 324–329, https://doi.org/10.1109/ISSRE.2015.7381825

SCSC Assurance Case Working Group (2021) Goal Structuring Notation community standard version 3. GSN Community Standard SCSC-141C, Safety-Critical Systems Club

Swedish Accident Investigation Authority (2025) MARCO POLO – grounding in northern Hanö Bay. Final report SHK 2025:03e, Swedish Accident Investigation Authority (SHK), Stockholm

Toulmin SE (2003) The Uses of Argument, updated edn. Cambridge University Press, Cambridge, https://doi.org/10.1017/CBO9780511840005, first published 1958

Toulmin SE, Rieke RD, Janik A (1984) An Introduction to Reasoning, 2nd edn. Macmillan, New York

UK Ministry of Defence (2007) Safety management requirements for defence systems, part 1: Requirements. Defence Standard 00-56 Issue 4, UK Ministry of Defence

Underwriters Laboratories (2023) Standard for safety for the evaluation of autonomous products. Tech. Rep. ANSI/UL 4600, Underwriters Laboratories, third edition

Utne IB, Rokseth B, Sørensen AJ, et al (2020) Towards supervisory risk control of autonomous ships. Reliability Engineering & System Safety 196:106757. https://doi.org/10.1016/j.ress.2019.106757

Verheij B (2003) DefLog: On the logical interpretation of prima facie justified assumptions. Journal of Logic and Computation 13(3):319–346. https://doi.org/10.1093/logcom/13.3.319

Verheij B (2005) Evaluating arguments based on Toulmin's scheme. Argumentation 19(3):347–371. https://doi.org/10.1007/s10503-005-4421-z

Appendix A Claim graphs

This appendix states the claim-graph model of section 4 formally and proves the properties that the main text uses.

A.1 Objects

Definition 1 (Standing)**.** The standings are the three-element chain

$$\text{withdrawn} \;<\; \text{reduced} \;<\; \text{licensed},$$

written $\mathcal{L}$.

Definition 2 (Evidence)**.** An evidence snapshot E is a partial map from a set of keys to values. How a snapshot is assembled lies outside the model.

Whether a key is present depends on the snapshot alone, so the model has no notion of time. An installation that discards values by age does so while it builds the snapshot.

Definition 3 (Condition of application)**.** A condition of application is a pair (κ, π) of an evidence key and a partial predicate on values. It is *established* under E when $\kappa \in \operatorname{dom} E$ and $\pi(E(\kappa))$ is defined and holds. A claim is *applicable* under E when every one of its conditions of application is established, and *inapplicable* otherwise.

A condition is not established when its key is absent, when the value lies outside the predicate's domain (as when text reaches a numeric predicate), or when the predicate rejects the value. By naming a key, the graph's author declares that the claim depends on it, so a claim whose evidence never arrives is withdrawn.

Each condition reads one key through one predicate, and the conditions of a claim combine by conjunction. A circumstance that is a disjunction over several keys is therefore expressed as alternative grounds: one claim per disjunct, each with its own conditions, and a requirement that any one of them suffices. A relation between keys, such as the distance between two positions, cannot be expressed in the graph and must enter as a key computed by the evidence layer.

Definition 4 (Requirement)**.** A requirement is a tuple $\rho = (S, s, k, m)$ of a finite set S of claims, a standing $s \in \mathcal{L}$, and integers k and m. Given an assignment σ of standings to claims, write

$$M_\sigma(\rho) = \big|\{\, c \in S : \sigma(c) \geq s \,\}\big|, \qquad L_\sigma(\rho) = \big|\{\, c \in S : \sigma(c) = \text{licensed} \,\}\big|.$$

The requirement is *met* when $M_\sigma(\rho) \geq k$, and *at full strength* when $L_\sigma(\rho) \geq m$.

The two thresholds distinguish redundant grounds from alternative grounds. Setting $m = |S|$ declares that all members of S are wanted, although k of them suffice,

so losing any member reduces the standing. The position claim of section 4.3, with two receivers, $k = 1$ and $m = 2$, is of this kind. Setting $m = k$ declares that any k members give full strength, so losing surplus members leaves the standing unchanged.

Definition 5 (Claim and claim graph)**.** A claim declares a statement, a warrant, a backing, a finite set A_c of conditions of application and a finite set P_c of requirements. Only A_c and P_c are evaluated; the other three are carried for a reader. A claim graph is a finite set of claims together with a finite sequence Λ of them, the mode claims in their order of activation. The support relation is $d \to c$ whenever $d \in S$ for some $(S, s, k, m) \in P_c$.

Whether a warrant is adequate for the statement it supports is left to the reader of the graph.

A.2 The rule

The rule assigns each claim a standing from the snapshot and from the standings of the claims it rests on.

Definition 6 (Standing of a claim)**.** Given a snapshot E and an assignment σ to the claims immediately preceding c in the support relation, $\sigma(c)$ is given by the first case below that applies:

$$\sigma(c) = \begin{cases} \text{withdrawn} & \text{if } c \text{ is inapplicable under } E, \\ \text{withdrawn} & \text{if some } \rho \in P_c \text{ is not met}, \\ \text{licensed} & \text{if every } \rho \in P_c \text{ is at full strength}, \\ \text{reduced} & \text{otherwise.} \end{cases}$$

A claim with $A_c = \emptyset$ skips the first case. A claim with $P_c = \emptyset$, such as a source claim, has no requirement that can be unmet, so the second case does not apply and the third holds vacuously. A source claim is therefore licensed when it is applicable and is never reduced. Evidence affects only whether a claim is applicable, and a reduced standing arises only from requirements.

Definition 7 (Allowance)**.** A requirement $\rho = (S, s, k, m)$ *offers an allowance* when $s <$ licensed or $k < m$, and is *taken* when it is met and not at full strength.

Proposition 1 (Reduced is an allowance taken)**.** $\sigma(c) =$ reduced *if and only if c is applicable, every requirement of c is met, and at least one is taken. A requirement can be taken only if it offers an allowance.*

Proof The first sentence restates the fourth case of definition 6 together with the three before it. For the second, let ρ be taken, so $M_\sigma(\rho) \geq k$ and $L_\sigma(\rho) < m$. If $s =$ licensed then $M_\sigma(\rho) = L_\sigma(\rho)$, so $k \leq L_\sigma(\rho) < m$ and $k < m$; otherwise $s <$ licensed. Either way ρ offers

an allowance. Conversely let ρ have $s =$ licensed and $k = m$. Then $M_\sigma(\rho) = L_\sigma(\rho)$, so ρ is met exactly when it is at full strength, and it can never be taken. □

An allowance is thus either a tolerated standing below licensed ($s <$ licensed) or fewer licensed members than were wanted ($k < m$).

Definition 8 (Standing modes and activation)**.** The standing modes under E are the members of Λ whose standing is not withdrawn. Let $\lambda_i \in \Lambda$ be the mode in effect, which is given from outside the model. If the standing of λ_i is withdrawn, the activated mode is the first member of Λ after λ_i whose standing is not withdrawn.

A reduced standing mode stands on an allowance (proposition 1). Except through activation, the model does not determine which standing mode is in effect. Nor does it require a mode claim's requirements to imply those of the mode claims after it, so a reader assesses the order from the warrants and backing of the mode claims.

A.3 Wellformedness

The implementation checks six conditions when it loads a graph and rejects a graph that fails any of them. A rejected graph is never evaluated.

Definition 9 (Wellformed claim graph)**.** A claim graph is wellformed when

1. every claim named in the set S of a requirement is a member of the graph, and the support relation is acyclic;
2. every requirement has $1 \le k \le m \le |S|$;
3. every requirement has $s \in \{\text{reduced}, \text{licensed}\}$;
4. Λ is non-empty and its last member has no conditions of application and no requirements;
5. no member of Λ appears in the set S of any requirement; and
6. every predicate is built by conjunction and disjunction from at least one comparison or membership in a finite set given literally, with negation and disequality excluded.

Only condition (1) is needed for definition 6 to determine the standings. Without it, two claims that each require the other at licensed satisfy the definition both when both are withdrawn and when both are licensed, so the result would depend on how the evaluator breaks the cycle. The other five conditions exclude graphs that can be evaluated but whose results contradict the intended meaning of their fields. Condition (2) excludes a requirement that could be at full strength while unmet ($k > m$) or could never reach full strength ($m > |S|$). Condition (3) prevents a claim from standing on a withdrawn ground. Condition (4) guarantees that a standing mode follows every mode in Λ other than the last, since the last member is licensed under every snapshot. Condition (5) is needed because a mode claim can stand licensed without being in effect, and which mode is in effect is not monotone in the grounds, whereas requirements are (proposition 4).

Condition (6) excludes negation because a negated comparison accepts the floating-point non-numbers. For example, $\neg(\text{hdop} > 2)$ accepts a non-number and $\text{hdop} \leq 2$ does not, although the two predicates agree on every real value. Disequality is excluded for the same reason. Membership is limited to literal finite sets, which can still list the complement of an enumerated value, and De Morgan's laws give the complement of an interval without a negation node.

Besides the six conditions, the propositions below rest on the behaviour of the evaluator and of the installation. Appendix B states these assumptions as obligations and lists the results that depend on each (table B1).

A.4 Properties

Section 4 relies on the following properties. The first states that the two thresholds of a wellformed requirement are consistent.

Proposition 2 (Full strength implies met)**.** *For a wellformed requirement ρ and any σ, if ρ is at full strength then ρ is met.*

Proof Since licensed is the greatest element of $\mathcal{L}$, every member of S standing at licensed also stands at or above s, giving $M_\sigma(\rho) \geq L_\sigma(\rho)$. If ρ is at full strength then $L_\sigma(\rho) \geq m$, and $m \geq k$ by (2), so $M_\sigma(\rho) \geq k$. □

The order of the second and third cases of definition 6 therefore makes no difference to the result, because a claim whose requirements are all at full strength meets them all.

Proposition 3 (Existence and uniqueness)**.** *For a wellformed claim graph and any snapshot E there is exactly one assignment σ satisfying definition 6, obtained by one pass over the claims in any topological order of the support relation.*

Proof By (1), a topological order exists. Induct along that order. The value of $\sigma(c)$ depends only on E and on the already fixed standings of claims that precede c. The cases in definition 6 are ordered and end in an unconditional case. Exactly one value is therefore assigned at each step. □

Recomputing the standings from the same graph and snapshot therefore returns the original decision. The proposition concerns the graph after parsing, so one graph file yields the same graph in two evaluators only under obligation V2.

Proposition 4 (Monotonicity in grounds)**.** *Fix E and a claim c, and let $\sigma \leq \sigma'$ pointwise on the claims named in P_c. Then the value definition 6 assigns to c under σ is at most the value it assigns under σ'.*

Proof $M_\sigma(\rho)$ and $L_\sigma(\rho)$ are non-decreasing in σ, so met and at-full-strength are upward closed, and the conditions of c read E alone and are unchanged. If c is licensed under σ then every requirement is at full strength under σ, hence under σ'. If c is reduced under σ then

every requirement is met under σ, hence under σ', so c is licensed or reduced under σ'. If c is withdrawn under σ there is nothing to show. □

The proposition concerns the standings of grounds and does not extend to the evidence. A condition may fail on an implausibly good value, and the model permits such conditions, so improving a measurement can lower a standing. When one graph publishes a standing as evidence for another, the standing enters the second graph through a condition. Monotonicity then carries across the two graphs only if that condition is upward closed in $\mathcal{L}$, that is, if it accepts every standing above one it accepts (obligation E6).

Condition (6) secures the following property.

Proposition 5 (An uninformative value establishes nothing)**.** *Let v be a value of its type on which every comparison and every equality is false. Then no condition (κ, π) admitted by condition (6) is established when $E(\kappa) = v$.*

Proof Every atom of π is a comparison or a membership in a literal finite set tested by equality, so every atom is false at v. By (6) π is built from its atoms by conjunction and disjunction alone, so π is false at v. □

Under IEEE 754 (IEEE 2019) the floating-point non-numbers are such values. The evaluator meets the hypothesis only if it compares values with the type's own ordering and equality (obligation V1). The proposition does not apply to types without such a value, among them the integers and the booleans. For them, only the author's choice of predicate keeps a message format's default value from establishing a condition.

Filters upstream of the graph, such as staleness limits, change the snapshot before it is evaluated. Narrowing describes the changes that cannot add evidence.

Definition 10 (Narrowing)**.** A snapshot E' *narrows* E when for every key κ either $E'(\kappa) = E(\kappa)$, or $\kappa \notin \operatorname{dom} E'$, or no condition on κ admitted by condition (6) is established under E'.

In words, E' keeps each value of E, removes it, or replaces it with a value that establishes no admitted condition.

Proposition 6 (Narrowing cannot raise a standing)**.** *For a wellformed claim graph, let E' narrow E, and let σ and σ' be the assignments of proposition 3 under E and under E'. Then $\sigma'(c) \leq \sigma(c)$ for every claim c.*

Proof We first show that a condition established under E' is established under E. Take a condition on key κ. In the first case of definition 10, κ carries the same value in both snapshots. In the other two cases no condition on κ is established under E', since every condition of a wellformed graph is admitted by (6). Contrapositively, a condition unestablished under E is unestablished under E', so a claim inapplicable under E is inapplicable under E'.

Now induct along a topological order. If c is inapplicable under E' then $\sigma'(c) =$ withdrawn. Otherwise c is applicable under both snapshots, and both assignments reach the

remaining cases of definition 6, which do not read E. By the induction hypothesis, $\sigma' \leq \sigma$ at the claims preceding c, so proposition 4 gives $\sigma'(c) \leq \sigma(c)$. □

A filter that narrows the snapshot can therefore only lower standings, whatever its parameters, so staleness limits and plausibility gates need not be represented in the graph. The third case of definition 10 ranges over every predicate that condition (6) admits, whether or not the graph contains it. A value v that is equal to itself establishes the admitted condition $(\kappa, x \in \{v\})$, so the case applies to a present key only when its value is unequal to itself. A floating-point non-number is such a value and establishes no admitted condition (proposition 5). For types without such a value, narrowing can only remove the key or leave it unchanged. The proposition places no lower bound on what a filter removes. A filter that removes too little can leave a claim standing on invalid evidence, and no standing in the model reflects that failure.

Corollary 7 (Narrowing cannot add a standing mode)**.** *Under the hypotheses of proposition 6, the standing modes under E' are among the standing modes under E. If the mode in effect is withdrawn under E, it is withdrawn under E', and the activated mode under E' is the activated mode under E or a later member of Λ.*

Proof Every member of Λ withdrawn under E is withdrawn under E' by proposition 6, which gives the first statement and, applied to the mode in effect, the first part of the second. The first member after the mode in effect that is not withdrawn under E' is therefore not earlier than the first such member under E. □

Appendix B Obligations on the surrounding layers

The results of appendix A rest on behaviour of the evidence layer, the evaluator and the response stage that the checks at loading cannot cover. A violation therefore produces a wrong standing, or a record that misstates the mode in effect, without an error. Each obligation below is followed by the failure it prevents, and table B1 lists the results and the obligations on which each depends.

Result	Obligations
Propositions 1 and 2	none
Proposition 3 (existence and uniqueness)	E3, E4, E5; V2 for recomputation on another evaluator
Proposition 4 (monotonicity), across graphs	E6
Proposition 5 (uninformative value)	V1, on which E1 relies
Proposition 6 and corollary 7 (narrowing)	E2
Absence of a measurement withdraws its claim (definition 3)	E1, E7
The record shows the mode in effect	L1, L2

Table B1 Results of appendix A and the obligations on which each depends.

B.1 The evidence layer

E1. A producer with no measurement must omit the key or publish a value that no predicate accepts. Most message formats give an unset field a default value, typically zero, which is a valid speed, so a condition such as speed < 12 accepts a field that no one set. Every admitted predicate rejects a floating-point non-number (proposition 5), provided the evaluator meets obligation V1. Until tests confirm that it does, the producer should omit the key. An enumeration needs a reserved member for a missing measurement, kept out of every accepted set together with the format's default.

E2. A filter that discards evidence must remove the key or, for a floating-point key, replace its value with a non-number. Every other value of the types considered establishes some admitted condition (appendix A, after proposition 6). Such a filter is a narrowing, so it can only lower standings (proposition 6). A filter that substitutes a plausible value, as the *Royal Majesty*'s receiver substituted a dead-reckoned position, has no such guarantee.

E3. The snapshot must be assembled before the evaluation pass, from values observed close enough together to represent one state of the vessel. Otherwise two claims can be evaluated against different values of one key, and the standings are not the assignment of proposition 3. The obligation falls on the evidence layer because the model has no notion of time (definition 2).

E4. A producer of a key must not read the standings of the graph that consumes it. Such a key would let one claim withdraw another outside the declared requirements, and the standings would depend on the previous pass, which the record of one pass does not hold.

E5. Where graphs read each other's standings, the reads must form no cycle. The checks at loading see only one graph, and proposition 3 fails across a ring of graphs.

E6. A condition that reads another graph's standing must accept either licensed alone, or licensed and reduced. A condition that accepts *withdrawn* is wellformed, but an improvement in the other graph can then lower the reading claim, against proposition 4.

E7. Each source whose loss the graph is to represent must publish its own key. A loss reaches the graph only as a missing key (definition 3). If two receivers are merged into one position key and one drops, the key remains present and no claim is withdrawn.

B.2 The evaluator

V1. A comparison must use the value type's own ordering, and membership its own equality. Proposition 5 rests on every comparison and equality being false at a non-number. A range helper that clamps before comparing, or a set test by object identity, can accept a non-number, so the evaluator's conformance must be tested directly.

V2. The approved graph must store comparison literals in the representation the evaluator compares with. Two evaluators can round the decimal text 0.1 to different values, so recomputation on them can reach different decisions from the same file.

B.3 The response stage

L1. The response stage must keep a standing mode in effect. When the claim of the mode in effect is withdrawn, the activation follows in the same cycle. A hold or smoother that delays an activation keeps the vessel in a mode whose claim is withdrawn, and the record then shows a mode that the evidence did not support. A hold on a return to an earlier mode does not violate the obligation.

L2. The response stage must report each mode it puts into effect to the record. The report states whether the mode was activated or selected, and whether a person or a default rule made the selection. Without it, a later reader can establish which modes the evidence permitted but not in which mode the vessel operated.