A hierarchical tree-based decision-making approach for assessing the relative trustworthiness of risk assessment models

A hierarchical tree-based decision making approach for assessing the relative trustworthiness of risk assessment models

Tasneem Bani-Mustafa[(1)], Nicola Pedroni[(2)], Enrico Zio[(3), (4), (5)] , Dominique Vasseur[(6)] & Francois Beaudouin[(7)]

[(1)] *Chair on System Science and the Energetic Challenge, EDF Foundation*
Laboratoire Genie Industriel, CentraleSupélec/Université Paris-Saclay, 3 Rue Joliot Curie, 91190 Gif-sur-Yvette, France , *tasneem-adeeb.bani-mustafa@centralesupelec.fr*
[(2)] *NEMO Group, Energy Department, Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 Torino (Italy)*
[(3)] *MINES ParisTech, PSL Research University, CRC, Sophia Antipolis, France*

[(4)] *Energy Department, Politecnico di Milano, Via Giuseppe La Masa 34, Milan, 20156, Italy*

[(5)] *Eminent Scholar, Department of Nuclear Engineering, College of Engineering, Kyung Hee University, Republic of Korea*

[(6)] *EDF R&D, PERICLES (Performance et prévention des Risques Industriels du parc par la simulation et les Etudes)*
EDF Lab Paris Saclay - 7 Bd Gaspard Monge, 91120 Palaiseau, France
[(7)] *EDF R&D, PRISME (Performance, Risque Industriel, Surveillance pour la Maintenance et l'Exploitation)*
EDF Lab Chatou - 6 Quai Watier 78401 Chatou

Abstract:

Risk assessment provides information to support Decision Making (DM). Then, the confidence that can be put in its outcomes is fundamental, and this depends on the accuracy, representativeness and completeness of the models used in the risk assessment. A quantitative measure is needed to assess the credibility and trustworthiness of the outcomes obtained from such models, for DM purposes.

The present paper proposes a four-levels, top-down, hierarchical tree to identify the main attributes and criteria that affect the level of trustworthiness of models used in risk assessment. The level of trustworthiness (level 1) is broken down into two attributes (Level 2), three sub-attributes and one "leaf" attribute (Level 3), and seven basic "leaf" sub-attributes (Level 4). On the basis of this hierarchical decomposition, a bottom up, quantitative approach is employed for the assessment of model trustworthiness, using tangible information and data available for the basic "leaf" sub-attributes (Level 4). Analytical Hierarchical Process (AHP) is adopted for evaluating and aggregating the sub-attributes, and Dempster-Shafer Theory (DST) is adopted to consider the uncertainty and the inconsistency in the experts' judgments.

The approach is applied to a case study concerning the modeling of the Residual Heat Removal (RHR) system of a nuclear power plant (NPP), to compute its failure probability. The relative trustworthiness of two mathematical models of different complexity is evaluated: a Fault Tree (FT) and a Multi-States Physics-based Model (MSPM). The trustworthiness of the MSPM model is found to outweigh that of the FT model, which can be explained by the fact that MSPM takes into account the components failure dependency relations and degradation effects. The feasibility and reasonableness of the approach are, thus, demonstrated, paving the way for its potential applicability to inform DM on safety-critical systems.

Keywords:

Risk-Informed Decision Making (RIDM), Model Trustworthiness, Fault tree, Multi-States Physics-Based Model (MSPM), Analytical Hierarchical Process (AHP), Dempster-Shafer Theory (DST), Nuclear Power Plant (NPP).

1. Introduction

Risk assessments is based on complex *models* to represent functional life and physical behavior of (safety-critical) systems and processes of interest and provide predictions of safety performance metrics (Aven and Zio, 2013). These models are conceptual constructs (translated into mathematical forms), built on a set of assumptions (hypotheses) made on the basis of the available knowledge.

In general terms, risk models describe the future *consequences* (usually seen in negative, undesirable terms with respect to the planned objectives) potentially arising from the operation of given systems and activities, and the associated *uncertainty* (INSAG, 2011). The quantitative outcomes are, then, compared with predefined safety criteria, for Risk-Informed Decision Making (RIDM) (Dezfuli *et al.*, 2010); (NRC, 2010); (Eiser *et al.*, 2012).

In recent times, there has been a vivid discussion on the fundamental concept of risk and related foundational issues on its assessment: see, e.g., (Aven, 2013a), (Aven, 2016), (Cox and Lowrie, 2015). From a general perspective, it is understood that the outcomes of risk assessments are conditioned on the *background knowledge* and *information* available on the system and/or process under analysis (Bjerga, Aven and Zio, 2014), (Zeng *et al.*, 2016), including assumptions and presuppositions, phenomenological understanding, historical system performance data used and expert judgment made (Flage and Aven, 2009), (Aven, 2013b), (Veland and Aven, 2015), (Berner and Flage, 2016), (Bani-Mustafa *et al.*, 2018). Then, the risk indices may have a more or less solid foundation, depending on the validity of the hypotheses made, which in turn depends on the supporting knowledge: poor models, lack of data or simplistic assumptions are examples of potential sources of (model) uncertainty "hidden in the background knowledge" of a risk assessment (Berner and Flage, 2016). The modeling of a system or process needs to balance between two conflicting concerns: (i) *accurate representation* of the phenomena and mechanisms in the system or process and (ii) definition of the proper *level of detail* of the description of the phenomena and mechanisms, so as to allow the timely and efficient use of the model. Differences between the real world quantities and the model outputs inevitably arise from the conflict of these two concerns (Paté-Cornell, 1996); (Bjerga, Aven and Zio, 2014); (Danielsson *et al.*, 2016). Since (i) the importance placed on modeling and simulation is increasingly high within safety-critical system engineering contexts and (ii) the fundamental value of a risk assessment lies in providing informative support to (high-consequence) decision making (DM) (Simola and Pulkkinen, 2004); (EPRI, 2012); (Eiser *et al.*, 2012); (Zweibaum & Sursock, 2014), the *confidence* that can be put in the accuracy, representativeness and completeness of the models is fundamental. Also, a satisfactory level of assurance must be provided that the results obtained from such models are *credible* and *trustworthy* for the decision-making purposes for which they are employed. Moreover, in some contexts where the system of interest is subject to multiple hazards (e.g., a Nuclear Power Plant (NPP) exposed to flooding and earthquakes risks), a Multi-Hazards Risk Aggregation (MHRA) process is required to obtain a final risk metric that can inform decision making. However, risk estimates for different (risk) contributors are typically obtained using different models (i.e., in practice, different PRAs), each one having its own level of maturity and relying on its particular background knowledge. This inconsistency might be problematic, as MHRA is often carried out by a simple arithmetic summation of the risk estimates from different contributors, ignoring the possibly different levels of knowledge, which the risk estimates are based on (EPRI, 2015). Another situation, where the use of risk models with different credibility might be problematic, is that of choosing between the implementation of two different sets of risk reduction measures. For example, in a pure RIDM, a decision maker would always choose the option leading to the lower level of risk; however, his/her decision could change if he/she considered the level of trustworthiness, which the corresponding risk estimates are based on. For all these reasons, the *confidence*, *credibility* and *trustworthiness* (resp., *model uncertainty*) that is associated with model predictions (and that reflects the *amount* and the *strength* of the *knowledge* available on the problem of interest), must be accurately and quantitatively assessed (Aven and Zio, 2013); (Bjerga, Aven and Zio, 2014); (Flage and Aven, 2015).

Within this context, the objective of the present paper is to propose a decision-making approach based on a combination of hierarchical trees, the analytical hierarchal Process (AHP) (Saaty, 1980) and Dempster-Shafer Theory (DST) (Beynon, Curry and Morgan, 2000), (Beynon, Cosker and Marshall, 2001) to assess the relative trustworthiness of different models used in risk assessment. The main contribution of the work lies in the original structured *integration* of the techniques mentioned above in a systematic framework, to pragmatically and quantitatively address the problem of evaluating model trustworthiness, and accounting for the inevitable issue of *uncertainty* and *inconsistency* in the experts' judgments.

The proposed approach has been applied to assess the relative trustworthiness of two models (of different complexity and level of detail) used to estimate the failure probability of a Residual Heat Removal (RHR) System of a NPP: a classical Boolean logic-based Fault Tree (FT) and a Multi-State Physics-based Model (MSPM) (Unwin *et al.*, 2011), (Lin *et al.*, 2013), (Lin, Li and Zio, 2015), (Lin *et al.*, 2016).

A review of the approaches proposed in the literature to assess the trustworthiness and credibility of a model is presented in Section 2. In Section 3, a hierarchical tree-based decision making framework for assessing model trustworthiness is presented. In Section 4, the proposed framework is applied to a case study concerning the RHR system of a NPP. Finally, in Section 5, we discuss the results and provide some conclusions.

2. Assessing the trustworthiness and credibility of risk assessment models: a critical review of literature

In this section, we survey some approaches proposed in the open literature to assess the trustworthiness and credibility of mathematical models.

Few methods have been proposed to assess the confidence (i.e., the credibility and trustworthiness) that is associated with engineering model predictions, and that reflects the amount and the strength of the knowledge available on a generic system, or process of interest. In the literature, the trustworthiness of a method or a process is often measured in terms of its maturity. The concept of a model maturity was first used to assess the maturity of a function of an information system (Oberkampf *et al.*, 2007); (Paulk *et al.*, 1993); (Zeng *et al.*, 2016). Later, a framework called Capability Maturity Model (CMM) has been developed to assess the maturity of a software development process, in the light of its quality, reliability and trustworthiness (Herbsleb *et al.*, 1997). Recently, the CMM model has been extended and a Prediction Capability Maturity Model (PCMM) has been developed to evaluate and assess the maturity of modeling and simulation efforts (Oberkampf, Pilch and Trucano, 2007). Other examples of maturity assessment approaches have been developed in different domains, such as master data maturity assessment, enterprise risk management and hospital information system (Zeng *et al.*, 2016). In (Di Maio *et al.*, 2015) and (Zeng *et al.*, 2016) a hierarchical framework based on the analytical hierarchical process (AHP) has been developed to assess the maturity and prediction capability of a prognostic method for maintenance DM purposes. Finally, a framework for assessing the credibility of models and simulation (M&S) is proposed by (Nasa, 2013). In this framework, eight factors are used to assess M&S credibility and are categorized in three groups: (i) M&S development, including verification and validation; (ii) M&S operations, including input pedigree (a record of traceability from the input data source), results uncertainty and results robustness; (iii) supporting evidence, including the use history, M&S management and people qualifications. This framework seems plausible and covers important elements. However, three main issues should be considered. First, the approach is abstractly presented, leading to omit some important elements that fall under the main attributes of this framework. For example, while the model focuses on the "input pedigree" represented by the input data, it ignores a very important element, i.e., model assumptions, that can be also a part of M&S development. Second, while the authors claim that there is no need for weighting the different elements, as there is no numerical aggregation required, this would lead to a misconception, since the elements are not equally important in practice.

In the more specific field of "strength of knowledge" assessment in risk assessment models, both qualitative and semi-quantitative approaches have been proposed. In (Flage and Aven, 2009), a "crude" qualitative, direct grading of the strength of knowledge that supports risk assessment based on (mathematical) models is introduced. The authors try to classify the strength of knowledge to {minor, moderate, significant}, with respect to four criteria including: (i) phenomenological understanding of the problem; (ii) availability of reliable data; (iii) reasonability of assumptions made; (iv) agreement (consensus) among experts (i.e., low value-ladenness) (Flage and Aven, 2009); (Berner and Flage, 2016); (Aven, 2013b); (Veland and Aven, 2015); (Bani-Mustafa *et al.*, 2018).

In (Aven, 2013b) a more detailed, semi-quantitative approach (namely the "assumption deviation risk") has been introduced. This approach is based on the identification of all the main assumptions, on which the analysis is based. Then, the assumptions are converted into uncertainty factors and a rough evaluation of the deviation from the conditions defined by the assumptions is carried out. Finally, a score is assigned to each deviation that reflects the risk related to the deviation and its implications on the occurrence of given events and their consequences (Aven, 2013b). The approach has been generalized and systematized by Berner and Flage (2016), where guidelines to characterize the uncertainties associated to assumptions and deviations are also provided.

Also in (Bjerga, Aven and Zio, 2014) the effect and importance of "structural" assumptions, approximations and simplifications on risk assessment model outputs (Aven and Zio, 2013) is studied by means of different approaches, including subjective and imprecise probabilities and semi-quantitative scores (reflecting the degree of uncertainty associated to an assumption and the sensitivity of the model output to such assumption). The analysis serves as an input to the decision makers, to understand which assumptions are unacceptable and need "remodeling".

Finally, Lopez-Droguett and Mosleh discuss uncertainty in model predictions arising from model parameters and model structure. They argue that different evidence in evaluating model uncertainty can be considered, such as comparing the results of the model prediction to the actual measurements, qualitative or subjective evaluation of the model credibility and applicability (Droguett and Mosleh, 2008). In particular, for cases in which no model exists to address the particular problem of interest, and the analysis relies mainly on the subjective assumptions that the model is partially applicable to the problem, two main attributes define model uncertainty: model *Credibility* and model *Applicability* (Lopez Droguett & Mosleh, 2014). Model credibility refers to the quality of the model in estimating the unknown in its intended domain of application, and is defined by a set of attributes related to the model-building process and utilization procedure (*conceptualization and implementation*, which are in turn broken down into other sub-attributes). On the other hand, model applicability represents the degree to which the model is suitable for the specific situation and problem (represented by the conceptualization and intended use function attributes) (Lopez Droguett & Mosleh, 2014). A synthetic review is presented in Table A.1 in Appendix A.

As highlighted by the critical discussion above, different techniques can be found in the literature for assessing the strength of knowledge and the level of trustworthiness of risk models to inform DM. However, most of the aforementioned literature works treat the "factors" contributing to trustworthiness individually, without integrating them in a comprehensive framework for its assessment. In addition, the evaluation of the SoK and model trustworthiness is often carried out by directly scoring some "intangible" contributing factors (that cannot be easily translated into numbers). Finally, the evaluation is often carried out qualitatively or semi-quantitatively in the absence of rigorous evaluation protocols (scoring guidelines) to facilitate mapping of the verbal expressions into scores.

The present work tries to bridge these gaps by *originally integrating* concepts and attributes published in the open literature (see above) and available multi-attribute, multi-option MCDM techniques (e.g., decision trees and AHP) to produce a structured and systematic framework for the quantitative assessment of the trustworthiness of risk assessment models. The objective is twofold: (i) practically and quantitatively addressing the (relative) evaluation of model trustworthiness; (ii) treating the inevitable issue of uncertainty and inconsistency in the experts' judgments inherent in this type of analysis. Compared to the existing methods, the contributions of this paper include (see Section 3):

i. A conceptual hierarchical tree is developed to comprehensively represent the trustworthiness and to identify the main "tangible" (i.e., easily quantifiable) attributes and criteria that affect the level of trustworthiness of risk models;
ii. A top-down, bottom-up approach is developed for the practical evaluation of trustworthiness;
iii. Detailed scoring guidelines are provided to evaluate model trustworthiness in practice (i.e., the Saaty's linear scale and the balanced scale within the framework of the Analytical Hierarchical Process- AHP);
iv. A systematic procedure (based on Dempster Shafer Theory-DST) is outlined to take in due account the uncertainty and inconsistency in the experts' evaluations.

Moreover, the work in this paper is an attempt to support RIDM by giving indices on the trustworthiness, which can be pivotal in MHRA problems or in choosing among different alternatives for risk reduction measures (see the examples in the introduction).

3. Hierarchical tree-based decision-making approach for assessing the trustworthiness of risk assessment models

In section 3.1 below, we present the four levels, top-down tree used to characterize the trustworthiness (of a risk assessment model) by decomposing it into sub-attributes (e.g., number of model's assumptions, quantity of relevant data available, etc.) that can be quantified by the analysts; in Section 3.2, we describe a bottom-up procedure, based on the analytical hierarchy process (AHP), to assess the model trustworthiness by evaluating and aggregating the sub-attributes (identified as "leaf" attributes).

3.1. A hierarchical tree for model trustworthiness characterization: abstraction and decomposition

Many factors (attributes) affect the trustworthiness and credibility of analyses and models (for risk assessment in particular), and several studies and literature reviews have been made in order to identify them. Some of these are summarized as follows: (i) phenomenological understanding of the problem; (ii) availability of reliable data; (iii) reasonability of the assumptions; (iv) agreement among the experts; (v) level of detail in the description of the phenomena and processes of interest; (vi) accuracy and precision in the estimation of the values of the model parameters; (vii) level of conservatism; (viii) amount of uncertainty and others (see e.g., (Flage and Aven, 2009); (Berner and Flage, 2016); (Aven, 2013a); (Veland and Aven, 2015); (IAEA, 2006); (Bjerga, Aven and Zio, 2014); (Zeng *et al.*, 2016); (Oberkampf, Pilch and Trucano, 2007); (EPRI, 2012); (EPRI, 2015); (Bani-Mustafa *et al.*, 2018)). Some of these attributes (criteria), are not tangible and cannot be measured directly: as a consequence, other sub-attributes must be identified, which can be measured and/or subjectively evaluated. To this aim, on the basis of the critical literature survey presented in Section 2, we propose a method for model trustworthiness characterization and decomposition, which is based on the hierarchy tree shown in Figure 1.

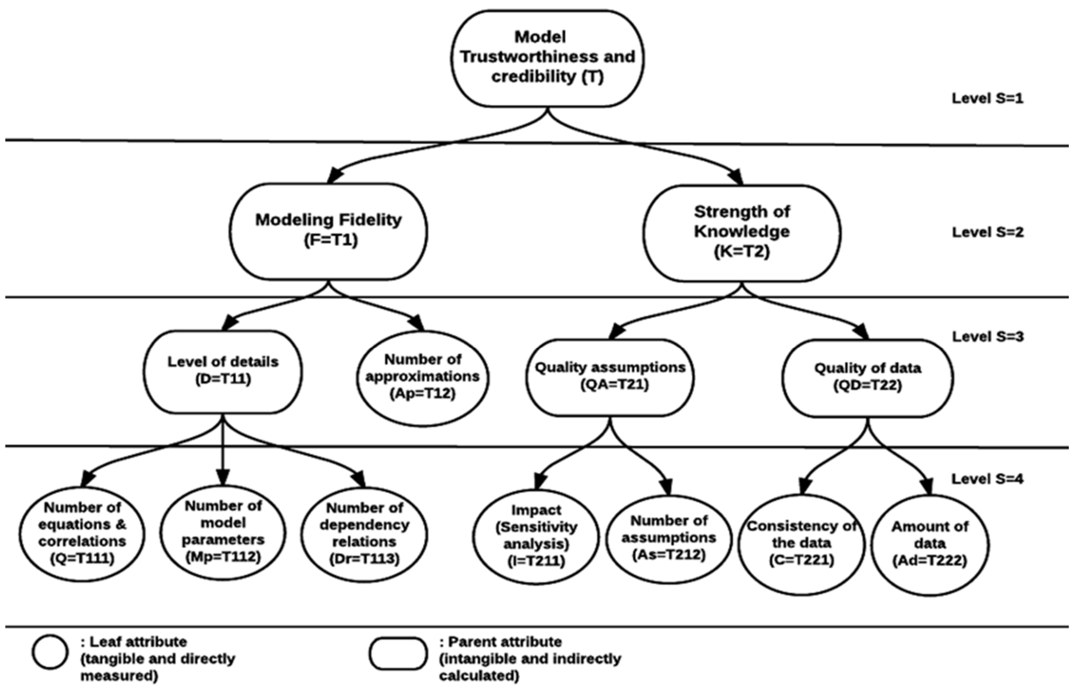


Figure 1 A hierarchical tree-based "decomposition" of the level of trustworthiness of a mathematical model

As mentioned above, many factors can be found in the literature that characterize the level of trustworthiness. In this paper, the model trustworthiness T (Level 1), is characterized by two attributes (Level 2): (i) strength of knowledge ($K = T_2$), which measures how solid the assumptions, data and information (which the model relies on) are (Flage and Aven, 2009); (ii) modeling fidelity ($F = T_1$), which embodies the ability of the model in representing the reality and the degree of implementing correctly the model. These two attributes are, in turn, decomposed into sub-attributes (Level 3). In particular, for the strength of knowledge, among the four sub-elements proposed in (Flage and Aven, 2009), two were found to be most relevant to the context of interest: i.e., quality of data ($QD = T_{22}$) and quality of assumptions ($QA = T_{21}$). The modeling fidelity ($F = T_1$) is defined by level of details ($D = T_{11}$) and number of approximations ($Ap = T_{12}$). With respect to $D = T_{11}$, it is argued that including more details about a problem is more representative and realistic, and hence more trustworthy. Also, note that the number of approximations ($Ap = T_{12}$) is considered as a basic attribute, since it can be measured directly: thus, it is not further broken down into other sub-attributes. The other three attributes of Level 3 are instead, broken down into more basic "leaf" attributes, as illustrated in Figure 1, that can be measured directly by "inspection" of the model whose trustworthiness need to be assessed. In particular, the level of detail ($D = T_{11}$) is characterized in terms of the number of equations and correlations ($Q = T_{111}$), the number of model parameters ($Mp = T_{112}$), and the number of dependency relations included ($Dr = T_{113}$). The overall quality of the assumptions ($QA = T_{21}$) is measured by the number of assumptions made ($As = T_{212}$) and by their impact ($I = T_{212}$) (which can be assessed, e.g., by sensitivity analysis). Finally, the quality of the data ($QD = T_{22}$) is described in terms of the amount of data available ($Ad = T_{221}$) and by the consistency of the data itself ($C = T_{222}$). The definitions of the attributes are given in Table 1, for the sake of clarity

Table 1 Definition of the attributes used to characterize the model trustworthiness

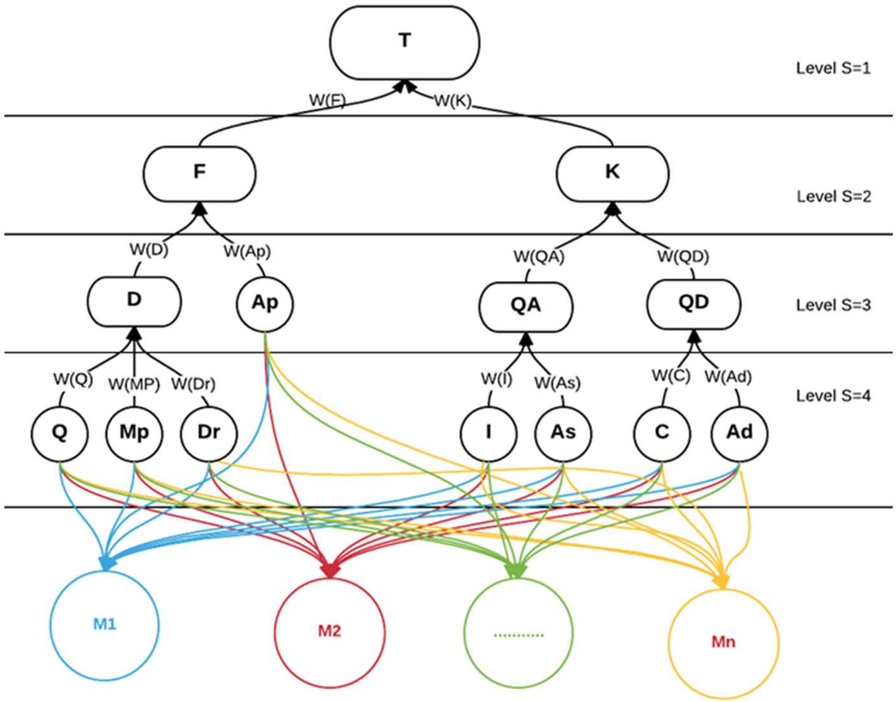
Attributes	Description
Modeling fidelity $F = T_1$ (Level S = 2)	Measures how close the model is to reality, i.e., the adequacy of the representation of the phenomena and processes of interest: the higher the modeling fidelity, the higher the trustworthiness of the model.
Strength of knowledge $K = T_2$ (Level S = 2)	Represents the level of understanding of the phenomena and the solidity of the assumptions, data, and information, which the model relies on: the higher the strength of knowledge, the higher the trustworthiness of the model.
Level of detail $D = T_{11}$ (Level S = 3)	Measures the level of sophistication of the analysis by quantifying to which level the "elements" and aspects of the phenomenon, process or system of interest are taken into account in the model: the higher the level of detail, the higher the trustworthiness of the model.
Number of approximations $Ap = T_{12}$ (Level S = 3)	Measures the number of approximations that the analyst introduces in order to facilitate the analysis: it affects the modeling fidelity. The lower the number of model approximations the higher the modeling fidelity.
Quality of assumptions $QA = T_{21}$ (Level S = 3)	In some studies, experts are obliged to formulate some assumptions, which might be due to the lack of data and information, to the complexity of the problem or lack of phenomenological understanding. The quality of those assumptions is an indication of the strength of knowledge: the higher the quality of the assumptions, the higher the trustworthiness of the model.

Quality of data $QD = T_{22}$ (Level S = 3)	Represents the availability of sufficient, accurate and consistent background data with respect to the purposes of the analysis: the higher the quality of the data, the higher the trustworthiness of the model.
Number of equations and correlations $Q = T_{111}$ (Level S = 4)	The number of equations and correlations used in modeling is an indication of the level of detail, hence of the modeling fidelity: the higher the number of equations and correlations, the higher the trustworthiness of the model.
Number of model parameters $Mp = T_{112}$ (Level S = 4)	The number of parameters introduced in the model is a measure of the level of detail (e.g., the number of components transition rates represents the level of discretization adopted to describe the failure process of a component or a system): the higher the number of model parameters, the higher the trustworthiness of the model.
Number of dependency relations $Dr = T_{113}$ (Level S = 4)	The larger the number of dependency relations that are taken into account, the more detailed and trustworthy the model.
Number of assumptions $As = T_{211}$ (Level S = 4)	The larger the number of simplifying assumptions, the lower the trustworthiness of the model.
Impact of assumptions $I = T_{212}$ (Level S = 4)	It quantifies how much assumptions can affect the model results (and it can be assessed by sensitivity analysis). The higher the impact of the assumptions, the lower the trustworthiness of the model.
Consistency of data $C = T_{221}$ (Level S = 4)	It is an indication of how suitable and representative the data are for a specific process or system. The consistency of data relies on the sources of the data. For example, if we are collecting data about the failure of a safety system's pump from different power plants, we should first understand whether the power plants are of the same type, whether the plants work at the same power level and whether the pumps have the same work function and capacity. The consistency of the data used is an indication of the quality of data, hence of the strength of knowledge: the higher the consistency, the higher the strength of knowledge and the trustworthiness of the model.
Amount of data $Ad = T_{222}$ (Level S = 4)	The higher the amount of data available, the stronger the knowledge. For example, the number of years of experience of a particular component in a plant can be sometimes considered an indication of the amount of data available. In any domain, a higher number of years' experience means a higher number of scenarios covered and hence a larger amount of data. The higher the amount of data, the higher the trustworthiness of the model.

3.2. Analytical hierarchical process (AHP) for model trustworthiness quantification

Given the hierarchical tree in Figure 1, the assessment of (relative) model trustworthiness is carried out within a multi-criteria decision analysis (MCDA) framework (Xu and Yang, 2001); (Triantaphyllou and Shu, 1998). In this setting, we suppose, in all generality, that a system, process or phenomenon of interest for a risk assessment can be represented by different mathematical models of possibly different complexity and level of detail, $\mathrm{M1}, \mathrm{M2}, .. \mathrm{M_l}, .., \mathrm{M_n}$. The task (i.e., the MCDA problem at hand) is to rank these alternative models with respect to their trustworthiness, in relation to the particular risk assessment problem of interest to support MCDA. In the present paper, the Analytical Hierarchy Process (AHP) proposed by (Saaty and Vargas, 2012) is adopted to this aim. Other MCDA approaches could be obviously used, as well. For example, a dual index approach is proposed in (Salehpour-Oskouei and Pourgol-Mohammad, 2018) based on Shannon entropy theory.

In this approach, the top goal, i.e., the decision problem considered (in this case, the model trustworthiness), is placed at the first level of the hierarchy and, then, decomposed into several sub-attributes distributed over different levels according to their degree of tangibility (see the detailed description in the previous Section 3.1). Finally, the bottom level of the hierarchal tree-based AHP model contains the different alternatives (i.e., the models M1, M2, …, Mn) that need to be evaluated with respect to the top goal (Saaty, 2008). Through pairwise comparisons among the elements and the attributes of the same level S, the alternative solutions (i.e., the models), can be ranked with respect to the decision problem in the top level (i.e., the model trustworthiness) (Saaty, 2008), (Zio et al., 2003).The A HP model for model trustworthiness assessment is represented in Figure 2. The first step required to assess the model trustworthiness by AHP is the determination of inter-level priorities (weights) for each attribute, sub-attribute, basic "leaf" sub-attribute and alternative solution i.e., $W(T_i)$, $W(T_{ij})$, $W(T_{ijk})$, and $W(M_l, T_{ijk})$, respectively. Notice that in practice, each weight represents the relative contribution of an attribute of a given level to the corresponding "parent" attribute of the upper level: for example, weight $\mathrm{W}\left(T_{ijk}\right)$ quantifies the contribution (i.e., the importance) of basic "leaf" sub-attribute T_{ijk} (Level 4) in the representation and definition of sub-attribute T_{ij} (Level 3); instead, weight $W(M_l, T_{ijk})$ is the weight of the $l-th$ model with respect to the basic "leaf" sub-attribute T_{ijk} .The weights $W(T_i)$, $W\left(T_{ij}\right)$ and $W\left(T_{ijk}\right)$ are calculated using pairwise comparison matrices filled by experts. Typically, experts use a linear scale to evaluate the relative importance (i.e., the contribution) of each criterion (of a given level *S*) with respect to the other. For example, the linear scale suggested by Saaty (2008) defines nine levels of relative importance, ranging from "equally important attributes" (number "1") to "one attribute extremely more important" than the other (number "9"). Further discussion is not reported here for brevity. See (Saaty, 2008) and (Zio, 1996)



for details on Saaty's verbal expressions of importance and (Saaty, 2008), (Alexander, 2012), (Saaty and Vargas, 2012) for details about AHP method and construction of pairwise matrices.

Figure 2 Hierarchical tree-based AHP model for the assessment of the relative trustworthiness of risk assessment models

For the tangible basic "leaf attributes" T_{ijk}, a quantitative evaluation $T_{M_l,T_{ijk}}$ can be given by *direct inspection* and analysis of the models. Instead, if the basic leaf sub-attributes cannot be given a direct numerical evaluation (or if the analyst does not feel confident in carrying out this task), the scaling system explained above (i.e., scores from 1 to 9) can be adopted to provide a (semi-quantitative) relative evaluation of the "leaf attributes" $T_{M_l,T_{ijk}}$ with respect to the risk models M_l available (guidelines are provided in Appendix B of this paper for relatively evaluating the basic leaf sub-attributes). After obtaining the weight for each criterion with respect to the corresponding upper level criteria, their "global" weighting for with respect to the top goal T can also be obtained by multiplying its weight by the weights of its upper parent elements in each level. For example, the "global" weight $W_{global}(T_{ijk})$ of basic "leaf" sub-attribute T_{ijk} with respect to the "top" attribute (goal) T is given by $W(T_{ijk}) \cdot W(T_{ij}) \cdot W(T_i)$. For example, in the hierarchy tree of Figure 1, the "global weighting" of the consistency of data (denoted by T_{221}) with respect to level of trustworthiness is obtained by multiplying its weight by the weight of quality of data (denoted by T_{22}) and the weight of strength of knowledge (denoted by T_2): $W(T_{221}) \cdot W(T_{22}) \cdot W(T_2) = W_{global}(T_{221})$. Finally, the (relative) trustworthiness $T(M_l)$ of a model M_l is evaluated using a weighted average of the corresponding leaf attributes $T_{M_l,T_{ijk}}$:

$$T(M_l) = \sum_{i=1}^{n_T} \sum_{j=1}^{n_{T_i}} \sum_{k=1}^{n_{T_{ij}}} W_{global}(T_{ijk}) \cdot \frac{T_{M_l,T_{ijk}}}{\sum_{l=1}^{n} T_{M_l,T_{ijk}}} \quad (1)$$

where $T_{M_l,T_{ijk}}$ is the numerical value that the basic "leaf" sub-attribute $T_{T_{ijk}}$ takes with respect to model M_l (for example, for attribute $Q = T_{111}$ variable $T_{M_l,T_{111}}$ equals the number of equations and correlations contained in M_l); n is the number of models to be compared; n_T, n_{T_i}, and n_{T_ij} are defined above.

Several considerations need to be made on the proposed approach. Clearly, there is no claim that the trustworthiness assessment method is comprehensive and complete. Attributes similar to those considered here have been already proposed and adopted in relevant works of literature: see, e.g., Flage & Aven (2009); Aven (2013b); Bani-Mustafa *et al.* (2018), where the strength of knowledge is assessed in terms of "phenomenological understanding", availability of reliable data", "agreement among peers" and "reasonability of assumptions". In addition, the enumeration of some model leaf attributes (e.g., approximations, assumptions, formulas…) may seem an "artifact" of presentation or interpretation, in absence of a protocol rigorously constructed to this aim that could lead to lack of consistency and consensus in the experts' judgments. On the other hand, the following aspects should be considered. First, such a type of approach has been already used for evaluating attributes in relevant models, e.g., evaluation of phenomenological understanding, availability of reliable data, reasonability of assumptions and agreement among peers, demonstrating the feasibility (Flage & Aven, 2009). Second, the issue of enumerating model assumptions and evaluating their quality have already been treated in several papers: see, e.g., (Aven, 2013b); (Boone et al., 2010); (Berner and Flage, 2016); (Khorsandi and Aven, 2017). Then, most important, notice that the "direct enumeration" is not the only way to provide numerical values $\mathrm{T}_{\mathrm{M_l},\mathrm{T_{ijk}}}$ for the basic "leaf" attributes $\mathrm{T}_{\mathrm{T_{ijk}}}$ with respect to model M_l. As mentioned above, if the analyst does not feel confident, e.g., in "counting" assumptions, formulas and correlations, he/she may resort to semi-quantitative scales (e.g., scores from 1 to 9), in order to provide a relative evaluation of a "leaf" attribute $\mathrm{T}_{\mathrm{T_{ijk}}}$ with respect to the different risk models M_l's available (see for example the enumerating protocols in Appendix B, based on technical reports and experts' feedback). Finally, if the assessor does not feel comfortable with the assumption evaluation presented in the guidelines, she/he is free to use some other established methods, such as the NUSAP pedigree for assessing the quality of the assumptions (Van Der Sluijs *et al.*, 2005), (Boone *et al.*, 2010), (Kloprogge, Van der Sluijs and Petersen, 2011) or the assumptions deviation risk (Aven, 2013b); (Berner and Flage, 2016), (Khorsandi and Aven, 2017).

3.3. Uncertainty in the calculation of the inter-level priority weights in AHP

In this Section, some technical details related to the calculation of inter-level priority (weights) and the scoring of attributes in AHP are given. Most importantly, some issues associated to this assessment are addressed (i.e., the combination of the judgments from different experts and the enhancement of their consistency).

3.3.1. General balanced scale for pairwise comparison in AHP

As it has been illustrated above, in the AHP method experts typically use a "linear" scale from 1 to 9 to evaluate the strength (i.e., the contribution) of each criterion with respect to the other (see above). This scale is widely used in the literature and adopted by many scholars. However, this scale may not be suitable for assessing the level of trustworthiness, since it graduates linearly, which yields an uneven dispersion of weights. This, in turn, results in a misrepresentation of experts' real judgments and, therefore, in inaccurate estimates (Salo and Hämäläinen, 1997). As a consequence, many other scales have been introduced in the literature, that are more suitable for treating this kind of problems. In general, the verbal graduation of the scales has not been a concern in the literature. Instead, mapping these verbal graduations into numbers is what concerns the scholars. Actually, the criteria for selecting a scale must take into account the context of the problem (Salo and Hämäläinen, 1997). In this paper, the consistency of experts' evaluations is not a problem, since the pairwise comparison matrices are constructed iteratively to enhance their consistencies. In addition, we choose a "balanced scale" due to its ability to overcome the problem of the uneven dispersion of the local priorities (weights), which could lead to inaccurate estimates (Salo and Hämäläinen, 1997). In particular, we adopt the *generalized balanced scale* to ensure the equal dispersion of priorities for a large number of criteria (Goepel, 2018). In this scale, the priority vectors are equally dispersed (far apart from each other) for all n (Goepel, 2018):

$$w = \frac{9+(n-1)x}{n(n+8)} \quad (2)$$

where w is the priority, n is the number of criteria being compared, x is the number of judgments, $x = 1,2, \dots .9$. The scale r for these priorities is calculated as the following (Goepel, 2018):

$$r = \frac{9+(n-1)x}{9+n-x} \quad (3)$$

By way of example, let us assume that the expert is constructing a matrix to compare three attributes ($n = 3$). Then, the scales corresponding to the Saaty's verbal expressions (levels of importance) are calculated by Eq. (3) and found to be: $\frac{11}{11}, \frac{13}{10}, \frac{15}{9}, \frac{17}{8}, \frac{19}{7}, \frac{21}{6}, \frac{23}{5}, \frac{25}{4}, \frac{27}{3}$.

3.3.2. Quality and consistency of experts' judgments in AHP

Several factors affect the consistency and quality of experts' judgments. Steenbergen et al., (2013) identify three main factors related to the inconsistency in experts' judgments: (i) lack of prior knowledge on the problem; (ii) subjectivity of the judgments and delicacy of the subject; (iii) expert judgment not only on the criteria of their specialty, but also about all other criteria. They also suggest some recommendations to overcome the problem of inconsistency and uncertainties in the experts' judgments through (i) improving the quality of the information provided to select the experts; (ii) adopting an experts' judgments protocol to prioritize the criteria; (iii) improving the quality of information to experts, needed to prioritize criteria.

In general, it is recommended to consider multiple experts' opinions for assuring the quality, and overcoming inconsistency and uncertainty in the quantitative judgments in decision processes (Ferrell, 1985). The experts' opinions are usually combined using behavioral or mathematical aggregations. In the behavioral aggregation, the experts share information, discuss and agree upon a value (Ferrell, 1985). In the mathematical aggregation, the opinions of the experts are combined mathematically using, for example, arithmetic and geometrical means.

In AHP, in particular, the opinions of the experts are usually combined by weighted arithmetic or geometrical means. This, in turn, depends on the homogeneity of the experts' group structure. For example, if the expert group structure is homogenous and they are willing to act as a single individual, the experts are asked to make the pairwise comparisons individually and the weighted geometric mean is, then, used). On the other hand, if the experts' structure is not homogeneous, or attending conflicting viewpoints and interests, then the resulting individual priorities are aggregated using arithmetic means (Ossadnik, Schinke and Kaspar, 2016). It should also be highlighted that the conflicting points of view in the experts' judgments might be due to the low reliability of some experts. Therefore, the reliability of experts needs to also be considered. In this work, a mixture between behavioral (Ferrell, 1985), (Jenkinson, 2005) and mathematical (Seaver, 1976), (Ferrell, 1985), (Jenkinson, 2005) approaches is formulated, and Dempster-Shafer Theory-AHP (DST-AHP) is adopted to combine experts' judgments and enhance their consistency.

3.3.3. Procedural steps for applying the developed framework

In this step, the experts' opinions are elicited and aggregated based on a mixture between behavioral (Ferrell, 1985), (Jenkinson, 2005), and mathematical (Seaver, 1976), (Ferrell, 1985), (Jenkinson, 2005) approaches. It is endorsed to follow a procedural step for recruiting and preparing the assessors before starting the evaluation process (See for example, (Steenbergen et al., 2013), (Jenkinson, 2005)):

1. The assessors are asked to individually construct the pairwise comparison matrices (knowledge matrices) for evaluating the relative importance of the criteria;
2. For each tangible basic leaf sub-attributes T_{ijk}, a quantitative evaluation $T_{M_l,T_{ijk}}$ is given by direct inspection and analysis of the alternatives (models). Instead, if it cannot be given a direct numerical evaluation (or if the expert does not feel confident in carrying out this task), the scaling system explained in Section 3.3.1 can be adopted to provide a (semi-quantitative) relative evaluation of the leaf attributes T_{ijk} with respect to the risk models M_l available, and based on the guidelines provided in Appendix B.
3. The experts discuss among each other and explain why they choose each judgment for both the relative importance of the criteria and the scores of the tangible basic events in each model;
4. The assessors are, then, asked to reconsider their judgment and change it if necessary;
5. The consistency of each individual matrix is measured, and the matrix input is modified if necessary;
6. The eigenvector problems are solved, and the weights are determined;
7. The experts' judgments are combined mathematically using Dempster-Shafer Theory-AHP (DST-AHP) as explained in detail below.

I. Expert's reliability discounting

The first step for combining the weights using the DST method is the discounting one (Shafer, 1976), (Jiao *et al.*, 2016), which allows overcoming the problem of conflicting opinions and considering the doubt regarding the reliability of the source of information (the expert, in this case). In this step, the reliability of the expert is considered using Shafar's discounting technique (Shafer, 1976):

$$m_\delta(A) = \begin{cases} (\delta) \cdot m(A) & \forall A \subseteq \Theta,\ A \neq \Theta \\ (1-\delta) + (\delta) \cdot m(\Theta), & A = \Theta \end{cases}, \quad \delta \in [0,1] \tag{4}$$

where Θ represents the set of criteria to be compared, A is the proposition in the power set 2^Θ and is called the focal element, $m(A)$ is the basic belief assignment (BBA), $m_\delta(A)$ is the discounted belief assignment and finally, δ is the source (i.e., expert) reliability factor. A value of $\delta = 1$ means that the source is fully reliable and $\delta = 0$ menas that the source is fully unreliable. Note that the discounting process leads to generating a new focal set that contains all the criteria. By introducing this focal set, we are actually accounting for ignorance and uncertainty in the source (judgment).

II. Combination of experts' judgments

After discounting the BBAs (weights), Dempster's rule of combination is used to combine the experts' weightings of a given criterion (Shafer, 1976), (Jiao *et al.*, 2016):

$$m_{1,2}(C) = (m_1 \oplus m_2)(C) = \begin{cases} 0 & C = \phi, \\ \frac{1}{1-K} \cdot \sum_{A \cap B = C \neq \phi} m_1(A) \cdot m_2(B) & C \neq \phi, \end{cases} \tag{5}$$

where $m_{1,2}(C)$ is the new belief assignment resulting from the combination of the two BBAs (weights), $m_1(A)$ and $m_2(B)$, calculated from the two pairwise comparison matrices of the two experts, K is a measure of the amount of conflict between the belief sets (for the same focal set) from the two matrices and is given by:

$K = \sum_{A \cap B = \phi} m_1(A) \cdot m_2(B)$(6)

III. Pignistic probability transformation

Finally, the weights need to be transformed from the creedal to the pignistic level using the transferable belief model proposed by (Smets and Kennes, 1994):

$$w(x) = \sum_{C \subseteq \Theta, C \neq \phi} m(C) \frac{1_{C(x)}}{|C|}, \forall x \in \Theta \quad (7)$$

where $w(x)$ is the BBA of a single element (criterion) (can be used directly in Eq. (1)), 1_C is the indicator function of C: $1_C = 1, if\ x \in C\ and\ 0\ otherwise$, $|C|$ is the norm of C (the number of elements in the focal set). The mass functions obtained from the pignistic probability transformation represent the relative "believed weights" of the criteria indicated in Eq. (1). As explained earlier, in our case the BBA of a single element represents the local weights in the AHP method i.e., $W(T_i), W(T_{ij}), W(T_{ijk})$, which are, in turn, obtained based on the combination of experts' judgments. Please refer to (Bani-Mustafa et al., 2020) for more details on applying DST-AHP.

4. Case study

In this section, the hierarchical tree-based framework proposed is applied to a case study concerning the modeling of the residual heat removal (RHR) system of a nuclear power plant (NPP). In section 4.1, the system is described; in section 4.2, the characteristics of the two models used to represent the system (i.e. the Fault Tree-FT and the Multi-States Physics-Based Model-MSPM) are presented in some detail; finally, in section 4.3, the proposed approach is applied to evaluate the trustworthiness of the two models.

4.1. The system

The Residual Heat Removal (RHR) system of a typical PWR reactor is taken as reference. The RHR is mainly used to remove the decay heat (residual power) from the reactor cooling system and fuel during and after the shutdown, as well as supplementing spent fuel pool cooling in the shutdown cooling mode for some types of reactors (NRC, 2010). The main components of the RHR system are: pumps, heat exchangers, diaphragms and valves. According to previous studies, it was found that 23% of RHR system failures are due to pumps failures, 58% are due to valves failures, whereas the rest of RHR system failures are due to other components' failures (Coudray and Mattei, 1984).

4.2. Models considered

Two models have been considered for evaluating the reliability (resp., the failure probability) of the RHR system: a Fault Tree (FT) model (Section 4.2.1) and a Multi-State Physics-based Model (MSPM) (Section 4.2.2).

4.2.1. Fault Tree (FT) Model

The Andromeda software (Hibti *et al.,* 2012) has been used for the analysis of the RHR's components failure modes and criticalities (importance analysis). The analysis is based on a logical framework for understanding the different possible ways in which the components and the system can fail. The failure probabilities used in the FT analysis are based on field experience feedback.

4.2.2. Multi-State Physics-based Model (MSPM)

The Physics-based model (PBM) and multi-state model (MSM) paradigms are often used to describe the degradation processes of components and systems. Physics-based modeling aims to develop an integrated mechanistic description of the component/system life, consistent with the underlying degradation mechanisms (e.g. wear, stress corrosion, shocks, cracking, fatigue, etc.) by using physics knowledge and related mathematical equations. Multi-state modeling is built on material science knowledge, degradation and/or failure data from historical field records or degradation tests, to describe the degradation processes in a discrete way (Gorjian *et al.,* 2010), (Di Maio *et al.*, 2015). However, the state transition rate estimates are also based on physical models rather than operational data (Unwin *et al.*, 2011). In this light, a model that combines the PBM and MSM, namely the Multi-State Physics-based Model (MSPM) (Unwin *et al.*, 2011), has been proposed to describe comprehensively the process of transition and degradation (Di Maio *et al.*, 2015).

In particular, in the analysis of the present case study, the main critical components have been taken into account (i.e. pump, diaphragm, breaker, motor, contactor and valve). The MSM was used to model the pump, breaker, motor and contactor, while the PBM model was used to model the valve and diaphragm, taking into account the degradation dependency of the valve on the pump. Further technical details can be found in (Di Maio *et al.*, 2015), (Lin *et al.*, 2015).

The results of MSPM and FT (using Andromeda software) are given in Table 2. The analysis shows similar results in the first eight years and a difference appearing in the tenth year with a more rapid decline in the reliability values obtained by MSPM. This can be explained by MSPM's ability to consider the time-dependent degradation process, whose effects emerge late in time.

Table 2 Values of reliability

Time (years)	0	1	2	3	4	5	6	7	8	9	10
Reliability (FT)	1	0.779	0.607	0.473	0.369	0.288	0.224	0.175	0.143	0.107	0.083
Reliability (MSPM)	1	0.775	0.603	0.469	0.366	0.285	0.222	0.173	0.135	0.105	0.060

4.3. Evaluation of model trustworthiness

The analysis is carried out through two main steps: the first is an "upward" evaluation of the weight of each element in the hierarchy tree with respect to the top goal of model trustworthiness; the second is a "downward" assessment of the model trustworthiness by means of a numerical evaluation of the basic "leaf" elements for both FT and MSPM models, as shown in Figure 2.

With respect to the evaluation of the weights, experts were asked to fill the pairwise comparison matrices in order to evaluate the importance of each attribute (criteria). By way of example and only for illustration purposes, Eq. (8) shows a pairwise comparison matrix of the "leaf" sub-attributes $\mathrm{Q} = \mathrm{T}_{111}$, $\mathrm{Mp} = T_{112}$ and $\mathrm{Dr} = T_{113}$ of level $s = 4$. The attributes relative importances with respect to the parent attribute (level of details) are evaluated using the balanced scale. Note that three attributes are compared in this case. By Eq. (3), the scales corresponding to the nine qualitative levels of importance (Saaty's verbal expressions) are found to be $\frac{11}{11}, \frac{13}{10}, \frac{15}{9}, \frac{17}{8}, \frac{19}{7}, \frac{21}{6}, \frac{23}{5}, \frac{25}{4}, \frac{27}{3}$.

$$\begin{array}{c} \\ \mathrm{T}_{111} \\ \mathrm{T}_{112} \\ \mathrm{T}_{113} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{T}_{111} & \mathrm{T}_{112} & \mathrm{T}_{113} \end{array} \\ \begin{bmatrix} 1 & 15/9 & 1 \\ 9/15 & 1 & MF_{23} \\ 1 & 15/9 & 1 \end{bmatrix} \end{array} \tag{8}$$

By solving the eigenvector problem for this matrix, we obtain the flowing BBAs: $m(\mathrm{T}_{111}) = 0.385$, $m(T_{112}) = 0.230$, $m(T_{113}) = 0.385$. Note that the BBAs (weights) of the three attributes in the example sum to one: $\sum_{k=1}^{3} W_{11k} = 1$.

I. Expert's reliability discounting

The next step is to discount the BBAs given by the experts. In this example, two experts are invited to assess the weights of the trustworthiness attributes. The reliability of the two experts is assumed to be $\delta = 0.85$ and $\delta = 0.70$. In the previous example presented in Eq. (8), the weights of the expert are discounted using Eq. (4). The results are reported in Table 3.

Table 3 Discounted weights from two experts with two reliabilities

Focal set	$m_\delta(A)$ for Expert 1 ($\delta = 0.85$)	$m_\delta(A)$ for Expert 2 ($\delta = 0.70$)
$\{\mathrm{T}_{111}\}$	0.327	0.194
$\{\mathrm{T}_{112}\}$	0.196	0.253
$\{\mathrm{T}_{113}\}$	0.327	0.253
$\{\mathrm{T}_{111}, \mathrm{T}_{112}, \mathrm{T}_{113}\}$	0.15	0.30
Note that the expert can choose focal sets of single criterion, e.g., $\{\mathrm{T}_{111}\}$ or distinct group of criteria, e.g., $\{\mathrm{T}_{111}, \mathrm{T}_{112}\}$ if he/she thinks, to the best of his/her knowledge, that this focal set is comparable to the universal set that contains all the criteria. This allows accounting for the uncertainty in the judgment.		

II. Combining experts' judgments

In this step, the experts' judgments presented in Table 3, are combined using Eq. (5). Table 4 shows how to combine the judgments of the two experts.

Table 4 Dempster's rule of combination matrix

Expert 2 / Expert 1	$m_\delta(T_{111})$	$m_\delta(T_{112})$	$m_\delta(T_{123})$	$m_\delta(T_{111}, T_{112}, T_{123})$
$m_\delta(T_{111})$	$m_\delta(T_{111})_1$	ϕ_1	ϕ_2	$m_\delta(T_{111})_2$
$m_\delta(T_{112})$	ϕ_3	$m_\delta(T_{112})_1$	ϕ_4	$m_\delta(T_{112})_2$
$m_\delta(T_{123})$	ϕ_5	ϕ_6	$m_\delta(T_{113})_1$	$m_\delta(T_{113})_2$
$m_\delta(T_{111}, T_{112}, T_{123})$	$m_\delta(T_{111})_2$	$m_\delta(T_{112})_2$	$m_\delta(T_{113})_2$	$m_\delta(T_{111}, T_{112}, T_{123})_1$
*Please note that the element ij in the Table represents the multiplication of the elements $1j \times i1$, e.g., $m_\delta(T_{111}) \times m_\delta(T_{111}) = m_\delta(T_{111})_1$; $m_\delta(T_{111}) \times m_\delta(T_{111}, T_{112}, T_{113}) = m_\delta(T_{111})_2$				

From Eq. (6), $K = 0.399$.

From Eq. (5):

$$m_{1,2}(T_{111}) = \frac{0{,}191}{1 - 0.399} = 0.318$$

The same steps are repeated, and the combined weights for the other focal elements are found to be: $m_{1,2}(T_{112}) = 0.243$, $m_{1,2}(T_{113}) = 0.364$, $m_{1,2}(T_{111}, T_{112}, T_{123}) = 0.075$

III. Pignistic probability transformation

In this step, the weight on the pignistic level is found by Eq. (7):

$$w_{1,2}^{\delta}(T_{111}) = m_{1,2}^{\delta}(T_{111}) + \frac{m_{1,2}^{\delta}(T_{111}, T_{112}, T_{123})}{3} = 0.318 + \frac{0.075}{3} = 0.343$$

Similarly, $w_{1,2}^{\delta}(T_{112}) = 0.268$ and $w_{1,2}^{\delta}(T_{113}) = 0.389$. All results are reported in Table 5. Table 5 shows the weighting factors obtained: in particular, the weights of each attribute with respect to the corresponding "upper level" parent (i.e., $W(T_i), W(T_{ij})$ *and* $W(T_{ijk})$).

Different methods for assessing the weights (relative importance) of each attribute are implemented and compared for illustration purposes. As illustrated in Sect. 4.3, the procedural steps of Sect. 3.3.4 have been implemented to obtain the "DST-AHP weights", first using Saaty's linear scale and, then, using the balanced scale. On the contrary, the "weighted averages" are obtained using the conventional AHP method without accounting for the uncertainty in the assessors' judgment. Note that in this case study, we will only use the global weights (W_{global}) obtained by DST-AHP method for assessing the level of trustworthiness: the weights are shown just for comparison and illustration of the use of DST-AHP.

As a way of example, let us take the modeling fidelity attribute (highlighted in grey). First, the modeling fidelity weight was evaluated to be 0.25 and 0.50 by expert 1 (E1) and expert 2 (E2), respectively, using Saaty's scale. On the contrary, it was evaluated to be 0.40 and 0.50 using the balanced scale. Note that, in general, the difference between the two experts' evaluations decreases for all attributes, using the general balanced scale. This can be explained by the equal dispersions of the weights achieved by the balanced scale, which represents better the real evaluation of the expert (more representative of his/her real judgment). On the other hand, the difference caused by using the DST combination, instead of the weighted average combination method, cannot be directly predicted. Take again the modeling fidelity weights obtained using Saaty's scale as an example: the use of the DST method for combining the experts' opinions resulted in a decrease of the weight (0.303) compared to the weight obtained by the weighted average (0.363). On the other hand, the weight of the level of details attribute increases using the DST method (0.807 compared to 0.750). This is, in fact, because the DST method redistributes the weights taking into account the uncertainty and (in)consistency present in the experts' evaluations'.

Table 5 Attributes weighting factors calculated using the AHP method/ Balanced Scale

Attribute	Saaty Scale (1-9)				General Balanced Scale			
	E1	E2	weighted average	DST-AHP	E1	E2	weighted average	DST-AHP
Modeling fidelity	**0.250**	**0.500**	**0.363**	**0.303**	**0.400**	**0.500**	**0.445**	**0.421**
Strength of knowledge	0.750	0.500	0.637	0.697	0.600	0.500	0.555	0.579
Level of details	0.750	0.750	**0.750**	**0.807**	0.600	0.600	0.600	0.634

Number of approximations	0.250	0.250	0.250	0.193	0.400	0.400	0.400	0.366
Number of equations and correlations	0.429	0.200	0.325	0.332	0.385	0.278	0.336	0.343
Number of model parameters	0.143	0.400	0.259	0.215	0.231	0.361	0.290	0.268
Number of dependency relations	0.429	0.400	0.416	0.453	0.385	0.361	0.374	0.389
Quality of assumptions	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Quality of data	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Impact of assumptions	0.833	0.750	0.796	0.855	0.700	0.600	0.655	0.708
Number of assumptions	0.167	0.250	0.204	0.145	0.300	0.400	0.345	0.292
Consistency of data	0.750	0.750	0.750	0.807	0.600	0.600	0.600	0.634
Amount of data	0.250	0.250	0.250	0.193	0.400	0.400	0.400	0.366
E1 reliability factor: 70% E2 reliability factor: 85%								

The second step consists of an "upward" calculation, for the evaluation of the basic "leaf" attributes for each model using Eq. (1). Based on the data, information and knowledge, available and used in the risk assessment analysis, four types of trustworthiness analysis have been implemented. First, two assessments have been performed through direct quantitative evaluation of the leaf attributes (the number of model parameters is counted for each model) and Saaty's and balanced scales for evaluating the weights. The two scales have been used again to relatively assess the trustworthiness based on a semi-quantitative evaluation of the leaf attributes, which is carried out through comparing the two models to each other and, then, assigning a relative score for each leaf attribute.

In order to do that, scaling guidelines have been defined based on several EDF's technical reports (Burns, 1980) and the feedback of experts, and scores of 1-9 have been defined (see Appendices B-D for details) to evaluate the attributes.

Given the relative scores selected and using Eq. (1), the trustworthiness evaluation has been performed for both models, as illustrated in Tables 6-7. In this step, the assessors are asked to give a score S for the "leaf" attributes only (as illustrated in Tables 6-7), directly or relatively, with the help of the predefined guidelines, using Saaty's linear scale or the balanced scale. In fact, nine levels are defined in the guidelines. These levels can be mapped into scores of 1,2, ... 9, using Saaty's linear scale, or by Eq. (3) into scores of $\frac{10}{10}, \frac{11}{9}, \frac{12}{8}, \frac{13}{7}, \frac{14}{6}, \frac{15}{5}, \frac{16}{4}, \frac{17}{3}, \frac{18}{2}$, using the balanced scale (given that there are two models to be compared). The trustworthiness can, then, be simply calculated as the weighted average of the "leaf" attributes. Please note that the weighted score S_w is calculated as $S_{x-w} = W_{global} \cdot$

On the basis of the "leaf" attributes, the level of trustworthiness T was, then, calculated by Eq. (1) and found to be 4.594 for FT (M1) and 5.110 for MSPM (M2), using Saaty's Scale (normalized to 0.473 and 0.527 respectively as illustrated in Table 8). On the other hand, the use of the balanced scale results in a trustworthiness of 2.601 for M1 and 3.273 for M2 (normalized to 0.443 and 0.557 respectively, as illustrated in Table 8). In the same perspective, we have, again, applied the same method to evaluate the model's trustworthiness T using the direct quantification of the leaf attributes. The results are reported in Table 7. Note that since the evaluation of the leaf attributes is direct, the scores would be the same for the two scales. The difference in the weighted score comes only from the difference in the global weights for each scale. The results of the trustworthiness evaluation for the two models (FT and MSPM) are reported in Tables 6-7. Table 8 shows all the normalized results.

Table 6 Trustworthiness analysis using relative evaluation

Attribute	W_{global}		**Fault tree model (M1)**				**MSPM model (M2)**			
			Saaty		***Balanced***		***Saaty***		***Balanced***	
	W_S	W_B	S_S	S_{S-w}	S_B	S_{B-w}	S_S	S_{S-w}	S_B	S_{B-w}
Trustworthiness	-	-	-	4.594	-	2.601	-	5.110	-	3.273
Modeling fidelity	-	-	-	-	-	-	-	-	-	-
Strength of knowledge	-	-	-	-	-	-	-	-	-	-
Level of detail	-	-	-	-	-	-	-	-	-	-
Number of approximations	0.058	0.154	6	0.351	3	0.462	7	0.409	4.000	0.617
Number of equations and correlations	0.081	0.091	3	0.244	1.5	0.137	8	0.650	5.667	0.518
Number of model parameters	0.053	0.072	3	0.158	1.5	0.108	7	0.369	4.000	0.287
Number of dependency relations	0.111	0.104	1	0.111	1	0.104	4	0.444	1.857	0.193
Quality assumptions	-	-	-	-	-	-	-	-	-	-
Quality of data	-	-	-	-	-	-	-	-	-	-
Impact of assumptions	0.298	0.205	3	0.894	1.5	0.307	3.333	0.993	1.833	0.375
Number of assumptions	0.050	0.085	5	0.252	2.333	0.197	6	0.303	3.000	0.254
Consistency of data	0.281	0.183	8	2.250	5.667	1.039	5	1.406	2.333	0.428
Amount of data	0.067	0.106	5	0.336	2.333	0.247	8	0.537	5.667	0.601

W_S: global weight using Saaty's scale S: score

W_B: global weight using balanced scale S_w: weighted score

Table 7 Trustworthiness analysis using *direct evaluation*

Attribute	W_{global}		**Fault tree model (M1)**				**MSPM model (M2)**			
			Saaty		***Balanced***		***Saaty***		***Balanced***	
	W_S	W_B	S	S_w	S	S_w	S	S_w	S	S_w
Trustworthiness	-	-		22.713		33.311	-	41.604	-	63.352
Modeling fidelity	-	-	-	-	-	-	-	-	-	-
Strength of knowledge	-	-	-	-	-	-	-	-	-	-
Level of detail	-	-	-	-	-	-	-	-	-	-
Number of approximations	0.058	0.154	7	0.409	7	1.080	7	0.409	7	1.080
Number of equations and correlations	0.081	0.091	1	0.081	1	0.091	9	0.731	9	0.823
Number of model parameters	0.053	0.072	8	0.422	8	0.574	18	0.949	18	1.291
Number of dependency relations	0.111	0.104	0	0.000	0	0	1	0.111	1	0.104
Quality assumptions	-	-	-	-	-	-	-	-	-	-
Quality of data	-	-	-	-	-	-	-	-	-	-
Impact of assumptions	0.298	0.205	3	0.894	3	0.614	3.33	0.992	3.33	0.682
Number of assumptions	0.050	0.085	4	0.202	4	0.338	3	0.151	3	0.254
Consistency of data	0.281	0.183	8	2.250	8	1.467	5	1.406	5	0.917
Amount of data	0.067	0.106	275	18.456	275	29.146	549.15	36.855	549.15	58.202

Table 8 Summary of the models trustworthiness values using relative scores and direct measures

	Scale	Fault Tree	MSPM
Normalized Trustworthiness (relative scores)	Saaty Scale	0.473	0.527
	Balanced Scale	0.443	0.557
Normalized Model Trustworthiness (direct measures)	Saaty Scale	0.353	0.647
	Balanced Scale	0.345	0.655

In Table 8, the results show that MSPM model is more trustworthy than the Fault tree model (using all types of scales and evaluation methods). However, this finding is more significant using the balanced scale and the direct scoring of the leaf attributes. This can be explained by the ability of the balanced scale of representing better the opinions of the assessors.

In general, these results confirm the expectations, with the MSPM outweighing the fault tree. This can be explained by the fact that MSPM is based on well-established physical models that represent the time evolution of the states of the components, taking into account their interactions with the environment and other components, which affect the process of degradation.

5. Discussion and Conclusion

In this work, we have developed a hierarchical tree-based decision-making framework to assess the relative trustworthiness of risk models. The contribution of this work lies mainly in originally integrating , in a systematic and practical framework, some existing techniques of literature for evaluating the level of trustworthiness of risk assessment models, and simultaneously treat the uncertainty and inconsistency of expert's judgment inevitable in this kind of analysis. The approach is based on the identification of specific attributes that are believed to affect the trustworthiness of the model. This is obtained through a hierarchical tree-based "decomposition" of the model trustworthiness into sub-attributes. The DST-AHP method has been used to assess the weights of the attributes presented in the hierarchical tree. Then, a weighted aggregation of the attributes is performed to evaluate the model trustworthiness. The method has been applied to a case study involving the Residual Heat Removal (RHR) system of a Nuclear Power Plant (NPP). Two models of different complexity (i.e., FT and MSPM) have been considered to evaluate the system reliability and the trustworthiness of these models has been compared.

FT trustworthiness has been found to score 4.594 out of 9, whereas MSPM has scored 5.110 out of 9 using Saaty's scale, or respectively 2.911 and 3.273 using the balanced scale. These results mean that MSPM provides more trustworthy risk estimates than FT, which can be explained by the fact that it takes into account components failure dependency relations. Also, although the results of the reliability analysis using the two models are quite similar at the beginning, differences appear at long times. This can be explained by the MSPM's ability to consider the degradation affecting the components whose effects emerge at long times, and it is considered another feature for which the MSPM model outweighs the FT in trustworthiness.

Although the results confirm the expectation, this should, however, be taken with caution and no definitive conclusions should be drawn, since the analysis by the two models are neither based on the same data set, nor the same amount of resources. The case study is instead an attempt to show the applicability and feasibility of the developed methodological framework.

Clearly, there is no claim that the trustworthiness assessment approach proposed is comprehensive and complete, as there exist other factors that affect the level of trustworthiness, which were not considered here. The method was, rather, a first attempt to systematically evaluate the models' relative trustworthiness. Obviously, it impossible to remove completely subjectivity and expert judgment is still present, the method provided is an attempt to cast such expert judgment in a systematic and structured framework. Also, further studies should be performed to define the scaling guidelines for attributes evaluation and to study how to integrate the level of trustworthiness in RIDM.

References

Alexander, M. (2012) 'Decision-Making using the Analytic Hierarchy Process (AHP) and SAS/ IML', *The United States Social Security Administration Baltimore*, pp. 1–12.

Aven, T. (2013a) 'A conceptual framework for linking risk and the elements of the data-information-knowledge-wisdom (DIKW) hierarchy', *Reliability Engineering and System Safety*, 111, pp. 30–36. doi: 10.1016/j.ress.2012.09.014.

Aven, T. (2013b) 'Practical implications of the new risk perspectives', *Reliability Engineering and System Safety*, 115, pp.

136–145. doi: 10.1016/j.ress.2013.02.020.
Aven, T. (2016) 'Risk assessment and risk management: Review of recent advances on their foundation', *European Journal of Operational Research*, 253(1), pp. 1–13. doi: http://dx.doi.org/10.1016/j.ejor.2015.12.023.
Aven, T. and Zio, E. (2013) 'Model output uncertainty in risk assessment', *International Journal of Performability Engineering*, 9(5), pp. 475–486.
Bani-mustafa, T., Zeng, Z., Zio, E., Vasseur, D., (2018) 'Strength of Knowledge Assessment for Risk Informed Decision Making', in *Esrel*. Trondheim.
Bani-Mustafa, T., Zeng, Z., Zio, E. and Vasseur, D. (2020) 'A new framework for multi-hazards risk aggregation', *Safety Science*. Elsevier, 121, pp. 283–302.
Berner, C. and Flage, R. (2016) 'Strengthening quantitative risk assessments by systematic treatment of uncertain assumptions', *Reliability Engineering and System Safety*, 151, pp. 46–59. doi: 10.1016/j.ress.2015.10.009.
Beynon, M., Cosker, D. and Marshall, D. (2001) 'An expert system for multi-criteria decision making using Dempster Shafer theory', *Expert Systems with Applications*, 20(4), pp. 357–367. doi: https://doi.org/10.1016/S0957-4174(01)00020-3.
Beynon, M., Curry, B. and Morgan, P. (2000) 'The Dempster–Shafer theory of evidence: an alternative approach to multicriteria decision modelling', *Omega*, 28(1), pp. 37–50. doi: https://doi.org/10.1016/S0305-0483(99)00033-X.
Bjerga, T., Aven, T. and Zio, E. (2014) 'An illustration of the use of an approach for treating model uncertainties in risk assessment', *Reliability Engineering and System Safety*, 125, pp. 46–53. doi: 10.1016/j.ress.2014.01.014.
Boone, I., Van Der Stede, Y., Dewulf, J., Messens, W., Aerts, M., Daube, G. and Mintiens, K. (2010) 'A method to evaluate the quality of assumptions in quantitative microbial risk assessment', *Journal of Risk Research*, 13(3), pp. 337–352. doi: 10.1080/13669870903564574.
Burns, R. D. (1980) 'Wash 1400—Reactor safety study', *Progress in Nuclear Energy*. Elsevier, 6(1–3), pp. 117119–117140.
Coudray, R. and Mattei, J. M. (1984) 'System reliability: An example of nuclear reactor system analysis', *Reliability Engineering*. Elsevier, 7(2), pp. 89–121.
Cox, T. and Lowrie, K. (2015) 'Special Issue: Foundations of Risk Analysis'. WILEY-BLACKWELL 111 RIVER ST, HOBOKEN 07030-5774, NJ USA.
Danielsson, J., James, K. R., Valenzuela, M. and Zer, I. (2016) 'Model risk of risk models', *Journal of Financial Stability*. Elsevier, 23, pp. 79–91. doi: 10.1016/j.jfs.2016.02.002.
Dezfuli, H., Stamatelatos, M., Maggio, G., Everett, C., Youngblood, R., Rutledge, P., Benjamin, A., Williams, R., Smith, C. and Guarro, S. (2010) 'NASA Risk-Informed Decision Making Handbook'.
Droguett, E. L. and Mosleh, A. (2008) 'Bayesian methodology for model uncertainty using model performance data', *Risk Analysis*. Wiley Online Library, 28(5), pp. 1457–1476.
Eiser, J., Bostrom, A., Burton, I., Johnston, D. M., McClure, J., Paton, D., van der Pligt, J. and White, M. P. (2012) 'Risk interpretation and action: A conceptual framework for responses to natural hazards', *International Journal of Disaster Risk Reduction*, 1(1), pp. 5–16. doi: 10.1016/j.ijdrr.2012.05.002.
EPRI (2012) *Practical Guidance on the Use of Probabilistic Risk Assessment in Risk-Informed Applications with a Focus on the treatment of Uncertanity*. Palo Alto, California.
EPRI (2015) *An Approach to Risk Aggregation for RisK-Informed Decision-Making*. Palo Alto, California.
Ferrell, W. R. (1985) 'Combining individual judgments', in *Behavioral decision making*. Springer, pp. 111–145.

Flage, R. and Aven, T. (2009) ‘Expressing and communicating uncertainty in relation to quantitative risk analysis’, *Reliability: Theory & Applications*. Интернет-сообщество Gnedenko Forum, 4(2–1 (13)).

Flage, R. and Aven, T. (2015) ‘Emerging risk – Conceptual definition and a relation to black swan type of events’, *Reliability Engineering & System Safety*, 144(August), pp. 61–67. doi: 10.1016/j.ress.2015.07.008.

Franek, J. and Kresta, A. (2014) ‘Judgment scales and consistency measure in AHP’, *Procedia Economics and Finance*. Elsevier, 12, pp. 164–173.

Goepel, K. (2018) ‘JUDGMENT SCALES OF THE ANALYTICAL HIERARCHY PROCESS – THE BALANCED SCALE’, in *International Symposium of the Analytic Hierarchy Process 2018, Hong Kong, HK*. doi: 10.13033/isahp.y2018.033.

Goerlandt, F. and Montewka, J. (2014) ‘Expressing and communicating uncertainty and bias in relation to Quantitative Risk Analysis’, *Safety and Reliability: Methodology and Applications*, 2(13), pp. 1691–1699. doi: 10.1201/b17399-230.

Gorjian, N., Ma, L., Mittinty, M., Yarlagadda, P. and Sun, Y. (2010) ‘A review on degradation models in reliability analysis’, in Kiritsis, D. et al. (eds) *Engineering Asset Lifecycle Management: Proceedings of the 4th World Congress on Engineering Asset Management (WCEAM 2009), 28-30 September 2009*. London: Springer London, pp. 369–384. doi: 10.1007/978-0-85729-320-6_42.

Herbsleb, J., Zubrow, D., Goldenson, D., Hayes, W. and Paulk, M. (1997) ‘Software quality and the capability maturity model’, *Communications of the ACM*. ACM, 40(6), pp. 30–40.

Hibti, M., Friedlhuber, T. and Rauzy, A. (2012) ‘Overview of the open psa platform’, in *Proceedings of International Joint Conference PSAM*.

IAEA (2006) *Determining the Quality of Probabilistic Safety Assessment (PSA) for Applications in Nuclear Power Plants*. Vienna: INTERNATIONAL ATOMIC ENERGY AGENCY. Available at: http://www-pub.iaea.org/books/IAEABooks/7546/Determining-the-Quality-of-Probabilistic-Safety-Assessment-PSA-for-Applications-in-Nuclear-Power-Plants.

INSAG (2011) *A Framework for an Integrated Risk Informed Decision Making Process*. Vienna: INTERNATIONAL ATOMIC ENERGY AGENCY (INSAG). Available at: http://www-pub.iaea.org/books/IAEABooks/8577/A-Framework-for-an-Integrated-Risk-Informed-Decision-Making-Process.

Jenkinson, D. (2005) *The elicitation of probabilities: A review of the statistical literature*. Citeseer.

Jiao, L., Pan, Q., Liang, Y., Feng, X. and Yang, F. (2016) ‘Combining sources of evidence with reliability and importance for decision making’, *Central European Journal of Operations Research*. Springer, 24(1), pp. 87–106.

Khorsandi, J. and Aven, T. (2017) ‘Incorporating assumption deviation risk in quantitative risk assessments: A semi-quantitative approach’, *Reliability Engineering & System Safety*, 163, pp. 22–32. doi: https://doi.org/10.1016/j.ress.2017.01.018.

Kloprogge, P., Van der Sluijs, J. P. and Petersen, A. C. (2011) ‘A method for the analysis of assumptions in model-based environmental assessments’, *Environmental Modelling and Software*. Elsevier Ltd, 26(3), pp. 289–301. doi: 10.1016/j.envsoft.2009.06.009.

Lin, Y.-H., Li, Y.-F. and Zio, E. (2015) ‘Fuzzy reliability assessment of systems with multiple-dependent competing degradation processes’, *IEEE Transactions on Fuzzy Systems*. IEEE, 23(5), pp. 1428–1438.

Lin, Y. (2016) ‘A holistic framework of degradation modeling for reliability analysis and maintenance optimization of nuclear safety systems’. Université Paris-Saclay.

Lin, Y. H., Li, Y. F., Zio, E., Lin, Y. H. and Li, Y. F. (2013) ‘Multi-State Physics Model for the Reliability Assessment of a Component under Degradation Processes and Random Shocks Multi-State Physics Model for the Reliability Assessment of a Component under Degradation Processes and Random Shocks'.

Lopez Droguett, E. and Mosleh, A. (2014) 'Bayesian Treatment of Model Uncertainty for Partially Applicable Models', *Risk Analysis*, 34(2), pp. 252–270. doi: 10.1111/risa.12121.

Di Maio, F., Colli, D., Zio, E., Tao, L. and Tong, J. (2015) 'A multi-state physics modeling approach for the reliability assessment of nuclear power plants piping systems', *Annals of Nuclear Energy*, 80, pp. 151–165. doi: 10.1016/j.anucene.2015.02.007.

Di Maio, F., Turati, P. and Zio, E. (2015) 'Prediction capability assessment of data-driven prognostic methods for railway applications', in *Proceedings of the third European conference of the prognostic and health management society*.

Nasa (2013) 'STANDARD FOR MODELS AND SIMULATIONS-NASA-STD-7009', (I), pp. 7–11.

Nicolas Zweibaum & Jean-Pierre Sursock (2014) *Addressing multi-hazards risk aggregation for nuclear power plantsthrough response surface and risk visualization*. Palo Alto, California.

NRC (2010) *Reactor Coolant System and Connected Systems*, *NUREG-0800 Standard Review Plan for the Review of Safety Analysis Reports for Nuclear Power Plants*. Washington: NRC.

Oberkampf, W. L., Pilch, M. and Trucano, T. G. (2007) 'Predictive capability maturity model for computational modeling and simulation', *cfwebprod.sandia.gov*. Available at: https://cfwebprod.sandia.gov/cfdocs/CCIM/docs/Oberkampf-Pilch-Trucano-SAND2007-5948.pdf%5Cnfile:///Users/markchilenski/Documents/Papers/2007/cfwebprod.sandia.gov%0A/Oberkampf/cfwebprod.sandia.gov%0A 2007 Oberkampf.pdf%5Cnpapers://31a1b09a-25a9-4e20-879d-4.

Ossadnik, W., Schinke, S. and Kaspar, R. H. (2016) 'Group aggregation techniques for analytic hierarchy process and analytic network process: a comparative analysis', *Group Decision and Negotiation*. Springer, 25(2), pp. 421–457.

Paté-Cornell, M. E. (1996) 'Uncertainties in risk analysis: Six levels of treatment', *Reliability Engineering & System Safety*, 54(2), pp. 95–111. doi: 10.1016/S0951-8320(96)00067-1.

Paulk, M. C., Curtis, B., Chrissis, M. B. and Weber, C. V (1993) 'Capability Maturity Model for Software, Version 1.1', *Software, IEEE*, 98(February), pp. 1–26. doi: 10.1.1.93.1801.

Saaty, T. L. (1980) *The Analytic Hierarchy Process*, *McGraw HillIne*.

Saaty, T. L. (2008) 'Decision making with the analytic hierarchy process', *International Journal of Services Sciences*, 1(1), p. 83. doi: 10.1504/IJSSCI.2008.017590.

Saaty, T. L. and Vargas, L. G. (2012) *Models, methods, concepts & applications of the analytic hierarchy process*. Springer Science & Business Media.

Salehpour-Oskouei, F. and Pourgol-Mohammad, M. (2018) 'Sensor placement determination in system health monitoring process based on dual information risk and uncertainty criteria', *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*. SAGE Publications Sage UK: London, England, 232(1), pp. 65–81.

Salo, A. A. and Hämäläinen, R. P. (1997) 'On the measurement of preferences in the analytic hierarchy process', *Journal of Multi-Criteria Decision Analysis*. Wiley Online Library, 6(6), pp. 309–319.

Seaver, D. A. (1976) *Assessment of group preferences and group uncertainty for decision making*. DECISIONS AND DESIGNS INC MCLEAN VA.

Shafer, G. (1976) *A mathematical theory of evidence*. Princeton university press.

Simola, K. and Pulkkinen, U. (2004) *Risk Informed Decision Making A Pre-Study*. Finland: Nordisk Kernesikkerhedsforskning.

Van Der Sluijs, J. P., Craye, M., Funtowicz, S., Kloprogge, P., Ravetz, J. and Risbey, J. (2005) 'Combining Quantitative and Qualitative Measures of Uncertainty in Model-Based Environmental Assessment: The NUSAP System', *Risk Analysis*. Wiley Online Library, 25(2), pp. 481–492. doi: 10.1111/j.1539-6924.2005.00604.x.

Smets, P. and Kennes, R. (1994) 'The transferable belief model', *Artificial intelligence*. Elsevier, 66(2), pp. 191–234.

Steenbergen, R. D. J. M., van Gelder, P., Miraglia, S. and Vrouwenvelder, A. (2013) 'Safety, reliability and risk analysis: beyond the horizon', in. CRC Press, pp. 3355–3361.

Triantaphyllou, E. and Shu, B. (1998) 'Multi-criteria decision making: an operations research approach', *Encyclopedia of Electrical and Electronics Engineering*, 15, pp. 175–186. Available at: http://univ.nazemi.ir/mcdm/Multi-Criteria Decision Making.pdf.

Unwin SD, PP Lowry, RF Layton, Jr, P. H. A. M. T. (2011) 'Multi-State Physics Models of Aging Passive Components in Probabilistic Risk Assessment', in *In International Topical Meeting on Probabilistic Safety Assessment and Analysis*. Wilmington, North Carolina: Amercian Nuclear Society, La Grange Park, IL., p. vol. 1, pp. 161–172.

Veland, H. and Aven, T. (2015) 'Improving the risk assessments of critical operations to better reflect uncertainties and the unforeseen', *Safety Science*, 79, pp. 206–212. doi: http://dx.doi.org/10.1016/j.ssci.2015.06.012.

Xu, L. and Yang, J.-B. (2001) *Introduction to multi-criteria decision making and the evidential reasoning approach*. Manchester School of Management.

Zeng, Z., Di Maio, F., Zio, E. and Kang, R. (2016) 'A hierarchical decision-making framework for the assessment of the prediction capability of prognostic methods', *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*. SAGE Publications, 231(1), pp. 36–52. doi: 10.1177/1748006X16683321.

Zio, E. (1996) 'On the use of the analytic hierarchy process in the aggregation of expert judgments', *Reliability Engineering and System Safety*, 53(2), pp. 127–138. doi: 10.1016/0951-8320(96)00060-9.

Zio, E., Cantarella, M. and Cammi, A. (2003) 'The analytic hierarchy process as a systematic approach to the identification of important parameters for the reliability assessment of passive systems', *Nuclear Engineering and Design*, 226(3), pp. 311–336. doi: 10.1016/S0029-5493(03)00211-5.

Appendix A: Synthetic review of the methods in the literature

Table A.1 Synthetic review of the methods in the literature

Method	Use and objective	Characteristics and criteria	Methods
Predictive Capability Maturity Model (Oberkampf, Pilch and Trucano, 2007)	Assesses the level of maturity of computational modeling and simulation methods.	Semi-quantitative assessment of maturity with respect to six criteria: (i) representation and geometric fidelity; (ii) physics and material model fidelity; (iii) code verification; (iv) solution verification; (v) model validation; (vi) uncertainty quantification and sensitivity analysis.	Experts' knowledge.
Prediction capability of a prognostic method (Di Maio *et al.*, 2015),	Assesses the prediction quality of prognostic tools.	An indicator of prognostic performance assessed qualitatively and quantitatively given: a. The RUL model predication quality, which is assessed "*quantitatively*" based on: (i) Timeliness weighted error bias; (ii) sample mean error; (iii) mean absolute percentage error; (iv) mean square error; (v) sample median	Experts' knowledge; weighted average of criteria within AHP.

(Zeng *et al.*, 2016)		error; (vi) performance; (vii) weighted prediction spread; (viii) sample standard deviation; (ix) root mean square error; (x) prediction spread. b. The trustworthiness of method, which includes: (i) reliability; (ii) resources requirement; (iii) mathematical modeling adequacy; (iv) validity.	
Modeling and Simulation (M&S) credibility model (NASA, 2013)	Assesses the credibility of M&S tools.	Credibility assessed semi-quantitatively based on: (i) M&S development, including verification and validation; (ii) M&S operations, including input pedigree (a record of traceability from the input data source), results uncertainty and results robustness; (iii) supporting evidence, including the use history, M&S management and people qualifications.	Scoring protocols and experts' knowledge.
Knowledge assessment (Flage and Aven, 2009)	Expresses the knowledge on which risk assessment is based.	SoK qualitatively assessed as minor, moderate or significant, based on: (i) phenomenological understanding of the problem; (ii) availability of reliable data; (iii) reasonability of assumptions made; (iv) agreement (consensus) among experts (i.e., low value-ladenness).	Evaluation protocols and experts' knowledge.
Assumption deviation risk (Aven, 2013b), (Berner and Flage, 2016)	Assesses the possibility and criticality of risk deviations.	Semi-quantitative rough evaluation of the uncertainty associated to an assumption and the sensitivity of the model output to such assumption.	Experts' knowledge and local, one-at-a-time sensitivity analysis.
Evaluation of model uncertainty, credibility and applicability (Droguett and Mosleh, 2008 and 2014)	Assesses model uncertainty.	Comparison of model predictions and real data, within a Bayesian framework.	Bayesian methodology where information about models are available in the form of homogeneous and nonhomogeneous performance data (pairs of experimental observations and model predictions).

Appendix B: Method used to translate the hierarchical tree attributes into a semi-quantitative scale

The following table presents the guidelines adopted in this paper to translate the attributes of the hierarchical tree into a semi-quantitative scale. Such guidelines are defined based on discussions and suggestions provided by EDF analysts, with relevant experience in the problem ad case study at hand.

Table B.1 A semi-quantitative scale for the hierarchical tree attributes

Parameter	Translation "real number → scale 1/9"
Number of approximations ($Ap = T_{12}$)	Low number of approximation and low believed effect of their aggregate on the outputs: 9 Few approximations with low effect of their aggregate: 7 Moderate number of approximations with acceptable effect of their effect on the outputs: 5 High number of approximations with high effect of their aggregate on the outputs: 3

	High number of approximations with sever effect of their aggregate on the outputs: 1 The even number are left for the intermediate cases
Number of equations and correlations ($Q = T_{111}$)	1-2 equations : 1 3 equations : 2 4 equations or 1 (Boolean logic equation) : 3 . . >9 equations : 9
Number of state rates and model parameters ($Mp = T_{112}$)	0-2: 1 3-5: 2 . . >32: 9
Number of dependency relations considered ($Dr = T_{113}$)	0 dependency relations considered : 1 1%-12.5% of the failures rates are considered dependent on the failure of other components: 2 13.5%-25%: 3 26%-37.5%: 4 . . >88.5% All components failures are dependent on other components failures : 9
Number of assumptions ($As = T_{212}$)	Directly related to the actual number of assumptions used.
Impact (Sensitivity analysis and indications) ($I = T_{211}$)	The impact is related to the assumptions. The difference between the values of failure rate with and without the assumption should be estimated. A score between 1-9 is given for each assumption, and the final score is then averaged over all assumptions. 1. No repairs: assuming no component repairs, at time 500, we obtain a probability of failure which is 500 times higher as compared to the case when the repair is considered (Figs 9-12 (Lin, 2016)) 2. One directional dependency: assuming only one-direction dependency of the valve degradation from the degradation and vibration of the pump, decreases the valve reliability of about 3 times (Figs 9-21(Lin, 2016)) 3. Human error: In case of human error (omission in closing the manual valve), we obtain a probability of failure of RHR which is 1.096 times higher. Nevertheless, the human error probability is very small.

	4. No random shocks: assuming no random shocks results in a relative difference in the failure rate of the components. in particular, there is a reduction of (-2.99%-19823.08%) with respect to the case with the random shocks (Table II (Lin, 2016))
Consistency of data ($C = T_{221}$)	The expert should give a score between 1-9 evaluating of the consistency of data, taking into account the source of data, its compatibility and relevance to the components that need to be analyzed. As in the case study the data is collected from the same type of reactors 900 Mwe, it is highly consistent: the consistency is given a score of 8. However, we cannot guarantee a perfect consistency, as the information about a specific component might be collected from other components that are similar but slightly different: e.g., the failure rate of RHR pumps is calculated taking into account failures of all pumps in the reactor.
Amount of data (Number/amount of sources) ($Ad = T_{222}$)	The following classification is adopted according to the suggestions of EDF experts: > 25 reactor years of experience : 1 25-50: 2 51-100: 3 101-175: 4 176-275: 5 276-400: 6 401-550: 7 551-725: 8 Over 725: 9

Appendix C: Trustworthiness attributes evaluation for Fault Tree (FT) M1

Table C.1 Trustworthiness attributes evaluation for Fault Tree (FT)

Parameter	Direct score	Relative score	Note
$Ap = T_{12}$	7	6	7 minimal cut sets
$Q = T_{111}$	1	3	1 equation (Boolean logic): failure probability based on “rare event” approximation
$Mp = T_{112}$	8	3	8 failure rates for 8 basic events
$Dr = T_{113}$	0	1	No dependency relations considered
$As = T_{212}$	4	5	No repairs No dependency relations between components and failure mechanisms

			Human error No random shocks
$I = T_{211}$	3	3 4 4 1 Avg: 3	Based on the sensitivity analysis performed by (Lin, 2016) and the analysis performed using Risk Spectrum Software by EDF 1. No repairs: assuming no component repairs, at time 500, we obtain a probability of failure which is 500 times higher as compared to the case when the repair is considered (Figs 9-12 (Lin, 2016)) 2. No directional relation considered 3. Human error: In case of human error (omission in closing the manual valve) we obtain a probability of failure of RHR which is 1.096 times higher. Nevertheless, the human error probability is very small. 4. No random shocks: assuming no random shocks results in a relative difference in the failure rate of the components. in particular, there is a reduction of (-2.99%-19823.08%) with respect to the case with the random shocks (Table II (Lin, 2016))
$C = T_{221}$	8	8	The data are collected from application of SAFO (OMF-reliability-centered-maintenance-feedback computer assisted collection on 7 CP1-CP2 sites and report on data. As this data is collected from the same type of reactors 900 MWe it is highly consistent. On the other hand, we cannot guarantee a "perfect" consistency, as the information about a specific component might be collected from other, similar but possibly different, components: e.g., the failure rate of RHR motor operated valves is calculated taking into account failures of all motor operated valves in the reactor.
$Ad = T_{222}$	275	5	EDF internal reports on data collected between 1980 and 1992, or 275 years reactor for each component.

Appendix D: Trustworthiness attributes evaluation for Multi-State Physics-based Model (MSMP) M2

Table D.1 Trustworthiness attributes evaluation for Multi-State Physics-based Model (MSMP)

Parameter	Direct score	Relative score	Note
$Ap = T_{12}$	7	7	No relevant approximation
$Q = T_{111}$	9	8	4 multi-state models

			3 physical equations for valve and diaphragm behavior 2 threshold equations for D_v and D_D (denote respectively: the number of cycles of solicitation of the valve over time and the thickness loss of the pipe over time)
$Mp = T_{112}$	18	7	-5 transitions rates in the multi-state model - 11 parameters for physical equations for the valve and diaphragm - 2 parameters for the modeling of number of cycles and thickness loss (18 parameters in total)
$Dr = T_{113}$	1	4	1 dependency relation considered between the valve and the pump
$As = T_{212}$	3	6	No repairs 1 directional dependency: the dependency of the valve degradation on the pump degradation and vibration No random shocks
$I = T_{211}$	3.3333	3 6 1 Avg: 10/3	Based on the sensitivity analysis performed by (Lin, 2016): 1. No repairs: assuming no component repairs, at time 500, we obtain a probability of failure which is 500 times higher as compared to the case when the repair is considered (figs 9-12 (Lin, 2016)) 2. One directional dependency: assuming only one direction dependency of the valve degradation on the degradation and vibration of the pump decreases the valve reliability of about 3 times (Figs 9-21 (Lin, 2016)) 3. No random shocks: assuming no random shocks results in a relative difference in the failure rate of the components. in particular, there is a reduction of (-2.99%-19823.08%) with respect to the case with the random shocks (Table II (Lin, 2016))
$C = T_{221}$	5	5	The data are collected from internal technical reports: -Pump 621.95 years reactor (PWR 900 MWe, PWR 1300 MWe, PWR N4) PWR 900: 2 PWR 1300, N4: 2 -Breaker 420 Years reactor (PWR1300 MWe, CPY) CPY: 18 PWR 1300:19 -Contactor 528.21 years reactor (1300 MWe, CPY, PWR N4) CPY: 26 PWR 1300: 48 PWR N4-1400: 29 - Motor 626.42 years reactor (900 MWe, 1300 MWe, Palier PWR N4)

			CPY: 43 PWR 1300: 36 PWR N4-1400: 34 Even though the data collected in EDF internal reports comes from different sources with different types of reactors, it is still consistent as the different components are very similar.
$Ad = T_{222}$	549.15	8	-Pump : 621.95 years reactor -Breaker: 420 Years reactor -Contactor : 528.21 years reactor - Motor : 626.42 years reactor