UNIVERSITY OF GENOA
Department of Electrical, Electronic and Telecommunication Engineering
and Naval Architecture (DITEN)
PhD in Science and Technology for Electronic and Telecommunication
Engineering - Curriculum: Interactive and Cognitive Environments

Knowledge Acquisition Analytical Games: games for cognitive systems design

Francesca de Rosa

This dissertation is submitted for the degree of
Doctor of Philosophy

TUTOR: Dr Anne-Laure Jousselme SUPERVISOR: Prof. Alessandro De Gloria

February 2020

Coordinator of the PhD Course: Prof. Mario Marchese

Acknowledgements

I would like to acknowledge NATO Allied Command Transformation (ACT) support in the development and deployment of the Reliability Game, through the Data Knowledge Operational Effectiveness (DKOE) project at the NATO STO Centre for Maritime Research and Experimentation (CMRE). The deployment of the MARISA Game, instead, has been conducted under the MARISA project. The MARISA project has received funding from the European Union's Horizon H2020 research and innovation programme under grant agreement No 740698.

I am deeply grateful to my supervisor Prof. Alessandro De Gloria who guided and inspired me, leading by example with his professionalism and immense dedication to research.

I would like to express my gratitude to my tutor Dr Anne-Laure Jousselme for her support in these years, especially in the formulation of the research topic.

I would like to extend my sincerest thanks to NATO STO CMRE colleagues. I would like to thank Dr Catherine Warner, Director CMRE, and Dr Sandro Carniel, Head of Research Department CMRE, for their support in this activity. I am grateful to Dr Raffaele Grasso for the useful discussions on machine learning techniques. I am also very grateful to Mr Luca Morlando, Dr Elena Camossi, Dr Francesco Baralli and Mr Simone Vasoli for playing the games, the general discussions and their friendship.

My sincere gratitude goes to Cdr. Andrea Iacono and Cdr. Paolo Lombardi for their friendship and for always being available to support my research with enthusiasm, playtesting the games and sharing with me their valuable domain knowledge.

I would like to thank the NATO SAS-114 RTG team members for the useful discussions on the reliability concept, and the members of the NATO SAS-139 RTG for the interesting discussions on analytical wargaming.

My very special thanks goes to my family. I thank *Om* Luigi and *Tante* Stefania for their support since the incipit of this work. I am very grateful to Luigi, Federica and my parent-in-law, Bianca and Alberto, for their sincere affection.

Finally, I would like to dedicate this work to the persons without whom I would not have been able to start and complete this journey. To my parents, Sandra and Emanuele, who have supported me emotionally and physically, taking care with love of me, my children and my

husband. A special thank to my father also for all the time spent sharing his deep domain knowledge. To my children, Vittoria and Emanuele, who have helped me with their joy and patience, while *I was doing my homework*. Above all, I would like to thank from the deepest of my heart my husband Federico for his patience, understanding and unconditional love, which have lighted up every single day of this journey.

Abstract

Knowledge discovery from data and knowledge acquisition from experts are steps of paramount importance when designing cognitive systems. The literature discusses extensively on the issues related to current knowledge acquisition techniques. In this doctoral work we explore the use of gaming approaches as a knowledge acquisition tools, capitalising on aspects such as engagement, ease of use and ability to access tacit knowledge. More specifically, we explore the use of analytical games for this purpose.

Analytical game is not a new class of games, but rather a set of platform independent simulation games, designed not for entertainment, whose main purpose is research on decision-making, either in its complete dynamic cycle or a portion of it (i.e. Situational Awareness). Moreover, the work focuses on the use of analytical games as knowledge acquisition tools. To this end, the Knowledge Acquisition Analytical Game (K2AG) method is introduced. K2AG is an innovative game framework for supporting the knowledge acquisition task. The framework introduced in this doctoral work was born as a generalisation of the Reliability Game, which on turn was inspired by the Risk Game. More specifically, K2AGs aim at collecting information and knowledge to be used in the design of cognitive systems and their algorithms. The two main aspects that characterise those games are the use of knowledge cards to render information and meta-information to the players and the use of an innovative data gathering method that takes advantage of geometrical features of simple shapes (e.g. a triangle) to easily collect players' beliefs. These beliefs can be mapped to subjective probabilities or masses (in evidence theory framework) and used for algorithm design purposes. However, K2AGs might use also different means of conveying information to the players and to collect data. Part of the work has been devoted to a detailed articulation of the design cycle of K2AGs. More specifically, van der Zee's simulation gaming design framework has been extended in order to account for the fact that the design cycle steps should be modified to include the different kinds of models that characterise the design of simulation games and simulations in general, namely a conceptual model (platform independent), a design model (platform independent) and one or more implementation models (platform dependent). In addition, the processes that lead from one model to the other have been mapped to design phases of analytical wargaming. Aspects of game validation and player

experience evaluation have been addressed in this work. Therefore, based on the literature a set of validation criteria for K2AG has been proposed and a player experience questionnaire for K2AGs has been developed. This questionnaire extends work proposed in the literature, but a validation has not been possible at the time of writing. Finally, two instantiations of the K2AG framework, namely the Reliability Game and the MARISA Game, have been designed and analysed in details to validate the approach and show its potentialities.

Table of contents

List of figures

List of tables

Acronyms

AIS	Automatic Identification System
BN	Bayesian networks
BPA	Basic probability assignment
CARS	Crew Awareness Rating Scale
CISE	Common Information Sharing Environment
CPT	Conditional probability table
CSO	Company Security Officer
C-SAS	Cranfield Situation Awareness Scale
DAG	Directed Acyclic Graph
DataCron	Big Data Analytics for Time Critical Mobility Forecasting
DBN	Dynamic Bayesian network
EM	Expectation Maximisation
EC	European Commission
EEZ	Exclusive Economic Zone
EXP	Experiment
G	Game mechanic
GBDG	Geometric Belief Data Gathering
GECKA	Game Engine for Common/sense Knowledge Acquisition
GEQ	Game Experience Questionnaire
G/P/S	Gameplay/Purpose/Scope
GX	Game Experience
HF	human factors
HMM	Hidden Markov Model
H2020	Horizon 2020
ISAGA	International Simulation and Gaming Association
IUU	Illegal, unregulated and unreported fishing
KA	Knowledge acquisition
KEBN	Knowledge Engineering for Bayesian networks

KE	Knowledge Engineering
K2AG	Knowledge Acquisition Analytical Game framework
K2AGQ	K2AG evaluation Questionnaire
K2AGs	Games following the Knowledge Acquisition Analytical Game framework
LRIT	Long Range Identification and Tracking system
MARISA	Maritime Integrated Surveillance
MARS	Mission Awareness Rating Scale
MDE	Model/Driven Engineering
MEEGA	Model for the Evaluation of Educational Games
MLE	Maximum Likelihood Estimation
MSDBN	Multi-source dynamic Bayesian network
MV	Motor Vessel
M&S	Modelling and simulation
NASA-TLX	NASA Task Load Index
PE	Player Experience
POL	Patterns of Life
QUIS	Questionnaire for User Interaction Satisfaction
R	Round
RGBN	Reliability Game Bayesian network
SA	Situational Assessment
SABARS	Situation Awareness Behaviorally Anchored Rating Scale
SACRI	Situation Awareness Control Room Inventory
SAGAT	Situation Awareness Global Assessment Technique
SALSA	SA of en-route air traffic controllers in the context of automation
SAR	Search and Rescue
SARS	Situation Awareness Rating Scales Technique
SART	Situation Awareness Rating Technique
SASHA	Situation Awareness for Solutions for Human Automation [...]
SAW	Situational Awareness
SLOC	Sea Line of Communication
SMG	Smuggling of goods
SPAM	Situation Present Assessment Method
TBM	Transferable Belief Model
T&E	Testing and evaluation
VMS	Vessel Monitoring System
VTS	Vessel Traffic System

2TBN Temporal Bayes net

Definitions

Accuracy: "closeness of agreement between a test result or measurement result and the true value" [108]

Analytical game: a platform independent simulation game, designed not for entertainment, which main purpose is research on decision-making, either in its complete dynamic cycle or a portion of it

Analytical wargame: wargame for research

Cognitive technologies: technologies aiming at making sense of information, with the goal of supporting human cognitive abilities of inferring, predicting and taking decisions

Confidence: the state of feeling certain about the truth of something

Content design: “creation of characters, items, puzzles and missions” [22]

Game: "a form of play. It is an activity involving one or more players who assume roles while trying to achieve a goal. Rules determine what the player are permitted to do, or define constraints on allowable actions, which impact on the available resources, and therefore influence the state of the game space. Games deal with well-defined subject matter (content and context)" [119]

Gaming: "the common term, encompassing the terms [...] games [and] simulation" [119]

Geometric Belief Data Gathering: method used in the K2AGs to collect data on players' beliefs

Knowledge acquisition: the process of locating, collecting, and refining knowledge for the development of knowledge based systems [96]

Knowledge Acquisition Analytical Game: a game framework including analytical games which serves as knowledge acquisition tool for the design of cognitive technologies

Knowledge cards: cards used in K2AGs to convey messages to the player

Model: "a representation and abstraction of anything such as a real system, a proposed system, a futuristic system design, an entity, a phenomenon, or an idea" [1]

Modelling: process of creating a model

Precision: "closeness of agreement between independent test/measurement results obtained under stipulated conditions" [108]

Reliability: the degree of confidence that can be put on a specific source of information

Serious game: computerised games not designed primarily for pure entertainment

Simulation: "a functional model that imitates the behaviour of a reference system" [119]

Simulation game: "essentially a case study [, but] with the participants on the inside" [112]

Situational Assessment: process to attain Situational Awareness [74]

Situational Awareness: “the perception of the elements in the environment within a volume of time and space, comprehension of their meaning and the projection of their status in the near future” [74]

System design: “creation of rules and underlying mathematical patterns” [22]

Trueness: "closeness of agreement between the expectation of a test result or a measurement result and a true value" [108]

Wargame: "warfare model or simulation whose operation does not involve the activities of actual military forces, and whose sequence of events affects and is, in turn, affected by the decisions made by players representing the opposing sides" [163]

World design: “the creation of the overall backstory, setting and theme” [22]

Glossary

Analytical game . xi, xiii, 3, 21–26, 110, 113, 114

Analytical plan . 27, 39, 43

Artificial intelligence . 2, 7

Bayesian network . ix, xii, xiii, 2, 3, 6, 8, 10, 53, 55, 56, 76–81, 83–86, 96–100, 103, 105, 106, 108–111, 115

Belief function . 11, 12, 55

Cognitive informatics . 1

Cognitive technologies . 1

Confidence . xi–xiii, 31, 45–47, 49, 54, 60, 61, 63, 64, 66, 71–75, 90, 92, 114

Design . 5, 13, 14, 16, 18, 20, 53, 54, 56, 57, 65, 66, 75, 95, 96, 98, 99, 106, 109

elicitation . 3, 5–7, 93, 98, 114, 115

Evidence . 7, 12, 55, 61, 96

Expert system . 2

Game science . 3, 13, 14, 34

Geometric Belief Data Gathering (GBDG) . 28–31, 113, 116

Intelligent system . 2, 5, 7, 42, 87, 113

Knowledge acquisition . 2–6, 22, 24, 25, 34, 42, 47, 53, 84, 98–100, 110, 114, 115

Chapter 1

Introduction

Cognitive technologies have evolved since the late 1980's and early 1990's. Those technologies aim at making sense of information, with the goal of supporting human cognitive abilities of inferring, predicting and taking decisions [243]. While pursuing this objective, cognitive technologies try to mimic human reasoning abilities and schemes. As defined in [243], cognitive computing can be interpreted as a paradigm of intelligent computing methodologies and systems based on cognitive informatics, which is an "enquiry of computer science, information science, cognitive science, and intelligence science that investigates [...] the internal information processing mechanisms and processes of the brain and natural intelligence, as well as their engineering applications in cognitive computing" [244].

While we move towards higher degrees of automation, many of the cognitive tasks on which decision-making is grounded are gradually delegated to systems to facilitate operators of different working environments (e.g. safety, security, crises management, health, first aid). Situational Awareness (SAW) is one of the main building block of the dynamic decision making processes [75] where cognitive technologies might provide a major contribution. SAW is a state of knowledge defined as "the perception of the elements in the environment within a volume of time and space, comprehension of their meaning and the projection of their status in the near future" [74]. SAW can be obtained through a cognitive process known as Situational Assessment (SA).

Cognitive technologies act as "enabler[s], facilitator[s], accelerator[s] and magnifier[s] of human capability, [but] not [as] its replacement" [215]. Therefore, to improve human-machine teaming and user acceptance the system underlying reasoning and communication schemes should be intelligible [49] and possibly intuitive to the human.

Many authors (e.g. [159, 94]) have underlined the importance of adopting a human-centered design approach for those systems, as the operational environments do not only include technological elements, but extend beyond hardware and software to include procedu-

ral and human elements (e.g. Christensen's system model [40]). In operational environment users might play several roles, possibly concurrently, such as "decision maker, monitor, information processor, information encoder and storer, discriminator, pattern recognizer [,] . . . ingenious problem solver" [173] or disseminator.

In this thesis we will focus on operational environments, such as maritime Command and Control systems, but the concepts introduced equally apply to other domains that can entail a high degree of complexity in terms of information quantity, information quality, information variety, communication means and communication formats. This might push the information processing tasks (e.g. perception, correlation, filtering, sense making) sometimes beyond human ability. Moreover, the applications discussed refer to expert systems (i.e. intelligent agent systems or knowledge-based systems), which are "computer programs that exhibit a similar high level of intelligent performance as human experts" [198]. More specifically, expert systems are intelligent systems designed on the bases of knowledge acquired from experts [68].

Important steps in the design of such systems are the knowledge discovery from data and knowledge acquisition (KA) from experts [205]. This design phase embraces the extraction, structuring and organisation of expert knowledge to be encoded in an intelligent system [205].

Intelligent systems are required to deal with an ever growing amount of information and to cope with the associated inherent uncertainty. Therefore, information fusion techniques are coming progressively into play to reduce the cognitive burden placed on the operator. Research is ongoing into the fusion community to better understand how to appropriately handle the volume, variety and uncertainty of data and information. Therefore, the artificial intelligence community is giving particular attention to expert systems able to perform probabilistic reasoning (e.g. [122]). An example of probabilistic expert systems are the ones based on Bayesian networks (BNs), such as clinical diagnosis decision support tools (e.g. [11, 231, 202]). Maritime surveillance represents a good example of operational environments that could highly benefit from the deployment of intelligent systems. In the maritime domain we are witnessing for instance an increase in the development of BNs expert systems (e.g. [95, 65, 127, 48, 176]).

BNs have proven to be an interesting approach in this respect and a lot of effort has gone in the development of efficient inference and learning algorithms [122]. However, to be able to fully take advantage of this computational technology several issues related to the knowledge engineering (KE) task need to be further researched. For example, issues in KE for Bayesian networks (KEBN) [122] include but are not limited to: expert limited time availability, costs, expert inability to verbalise tacit knowledge or incoherence in probability assignments. A

natural answer to this KE issue has been to use available data and at least partially automate the KEBN (i.e. structural and parameter machine learning). Therefore, KEBN might be either data driven or domain knowledge driven. However, in many applications knowledge engineers resort to a mixed approach, in which knowledge is partially provided by domain experts and partially by data. The use of domain knowledge is also very valuable with respect to the kind of reasoning schemes that are implemented into the systems. In fact, when elicited from experts the BNs tend to implement causal reasoning, making such schemes more transparent to the user and, therefore, engendering trust in the system itself.

Research is ongoing with respect to the development of algorithms with enhanced performances for learning from data (e.g. [51, 225, 261]) and on how to incorporate the expert knowledge into the learning algorithm (e.g. [260]). Additionally, effort has also been devoted to the development of improved expert knowledge elicitation methods to support KEBN. The term "improved" in this context refers to several aspects, such as expert time required, costs, complexity of the task and boredness of the task, which might negatively influence the *motivation to think* of the expert. This might exacerbate the well known issues related to knowledge elicitation such as biases, inconsistencies and inability of experts to express some knowledge [122]. In fact, often expert's knowledge is neither directly accessible nor easy to verbalise [216].

In this doctoral work we explore the use of gaming approaches as a knowledge acquisition tools, capitalizing on aspects such as engagement, ease of use and ability to access tacit knowledge. More specifically, we explore the use of analytical games for this purpose and we introduce the innovative KA method of the Knowledge Acquisition Analytical Game (K2AG).

The remainder of this thesis is organised as follows. Chapter 2 presents KA related work and the mathematical notions of uncertainty handling needed to better understand the work presented. Chapter 3 includes an overview of the different sub-disciplines of game science, with a special focus on games used as research tools. Chapter 4 introduces the concept of analytical games, which is not a new type of games, but rather a set of games that can be identified through a change in perspective through which the games are classified. This change is achieved by an holistic approach to game science, which allows to blur the traditional boundaries posed by the definitions adopted in the single sub-disciplines. Furthermore, in this chapter we introduce the Knowledge Acquisition Analytical Game (K2AG) framework, which is an innovative framework for the design of games to be used in the context of KA. Chapter 5 and Chapter 7 present two different instantiations of K2AG, namely the Reliability Game and the MARISA Game. These chapters explain the knowledge engineering problems underpinning the design of such games, their design and their validation

activities. Chapter 6, instead, shows how the Reliability Game is not only a useful KA tool, but might be regarded also as a useful human factor method to assess Situational Assessment and Situational Awareness. Finally, Chapter 8 presents the conclusions of this work and further steps.

Chapter 2

Knowledge acquisition and uncertainty handling frameworks

2.1 Knowledge acquisition and intelligent systems

Knowledge discovery from data and KA from experts are two fundamental steps in the design of intelligent systems. More specifically, KA is the process of extracting, structuring, and organizing domain knowledge from experts [205].

To support KA several elicitation techniques have been introduced. As described in [229], those techniques can be classified as:

(i) Protocol-generation techniques (e.g. interviews, reporting techniques and observational techniques);

(ii) Protocol analysis techniques;

(iii) Hierarchy-generation techniques (e.g. laddering);

(iv) Matrix-based techniques;

(v) Sorting techniques;

(vi) Limited-information and constrained-processing tasks; and

(vii) Diagram-based techniques.

Protocol-generation techniques consist of several kinds of structured and unstructured interviews, reporting techniques and observational techniques. Protocol analysis techniques are used to identify elements such as goals, decisions, relationships and attributes to be further

exploited in knowledge modelling. Hierarchical-generation analysis aims at developing taxonomies or more in general hierarchical structures (e.g. decision networks). While matrix-based techniques aim at capturing relations between items in the form of a grid (e.g. problems vs solutions), sorting techniques capture the way in which items are compared and ordered. The limited-information and constrained-processing analyse specific tasks performed under time or information constraints. Finally, diagram-based techniques take advantage of graphical representations (e.g. concept mapping) to capture the relations between elements and concepts (e.g. causal relations or temporal relations).

Details on the different taxonomies, techniques and different uses in relation to the kind of knowledge and experts can be found in the abundant literature on the topic (e.g. [19, 146, 200, 85]). With respect to KA for BNs, the most widely adopted method for the elicitation of the structure and parameters are interviews and questionnaires. For example, in order to define the network parameters, experts might be requested to provide conditional probabilities by answering questions such as: *"What is the probability of drug smuggling, given that the two ships of interest are performing a rendezvous?"*. While apparently this might be a simple task to perform, there are a number of problems related to probability elicitation from experts. More specifically, experts might provide biased or incoherent probabilities [122]. The potential biases include *overconfidence*, *anchoring* and *availability*. The first one refers to the tendency to assign probabilities higher than justified. The second one refers to the tendency of estimates to be weighted on the bases of previous estimates. The last bias, instead, consists in assigning probabilities higher than justifiable to events that we remember or are more salient. Incoherence in estimates might be related, for example, to probabilities not summing up to one or to the use of the same verbal probability expression for different numerical mappings. Several researchers have been looking at solving those issues by modifying the interview or questionnaire approaches. Methods proposed include the elicitation of single probabilities through the use of gamble-like methods (e.g. probabilities wheels, lotteries, certainty equivalent gamble, betting) or probability scales, such as numerical scales with verbal anchor [185, 227]. Moreover, given that the number of probabilities to be elicited for each BN node grows exponentially with the number of parent nodes, some research proposes techniques to generate full conditional probability tables (CPTs) from a reduced number of elicited assessments. Examples of such approaches are the Weighted Sum Algorithm [50], the Likelihood Method [117] and the EBBN Method [253]. It should be noted that alternatively to the above mentioned methods, the use of latent variables is proposed in order to control the number of parent nodes (therefore, to control the CPTs dimension) by introducing them at intermediate steps, encoding more abstract concepts (e.g. [80, 57]). For

further details on the different techniques the reader is referred to the specific literature on elicitation of experts' probabilities (e.g. [184, 211, 242, 245, 179, 255, 84, 254, 241, 251, 8]).

2.2 Uncertainty and intelligent systems

The artificial intelligent community has slowly recognised how intelligent systems should not only reason logically, but should also be able to cope with uncertainty [122]. The term uncertainty encompasses three distinct concepts, namely ignorance, physical randomness (indeterminism) and vagueness [122].

Several mathematical frameworks for uncertainty handling exist. Probably the most widely adopted in artificial intelligence applications is the Bayesian one, which allows to reason about and with human opinions in terms of strengths of beliefs, which are treated as subjective probabilities [122].

In the engineering community, more and more the Bayesian framework (e.g. BNs) is adopted as uncertainty representation and inference tool due to the good balance between expressiveness and tractability [154]. However, limitation in the uncertainty expressiveness (e.g. impossible distinction between ignorance and indeterminism) might restrict its applicability. Other frameworks are proposed to overcome some of the limitations of the Bayesian one.

Another uncertainty framework is the Evidence theory, also known as Dempster-Shafer Theory or Belief Function Theory [60, 204]. This framework under specific conditions can be considered a generalization of the probabilistic Bayesian reasoning and classical logic, extending both classical set theory and probability theory.

While the reader is referred to the abundant literature on the Bayesian framework and Evidence Theory, the next sections provide the basic notions relevant to the work described in this thesis.

In this doctoral work, we do not aim at analysing which framework better suits different applications, as it is assumed that this is a design choice happening outside the knowledge acquisition task. In fact, this work focuses on providing more efficient and effective knowledge acquisition methods. While the final goal is to demonstrate their applicability to the collection of data useful for the modelling in different mathematical frameworks, due to time constraints the analysis has concentrated on applications within the Bayesian framework. The modelling and analysis of the data within the evidential framework is not included in this doctoral thesis, as it will be finalised in the next future. However, a brief overview on the theory is included in order to provide the ground for some details explained in Section 4.3.3. We refer the reader to more specific readings such as [60, 204, 210] for additional details.

2.3 Bayesian networks

In the two case studies presented in Chapter 5 and Chapter 7 we will refer to Bayesian networks (BNs). A Bayesian network is a Directed Acyclic Graph (DAG) that supports reasoning under uncertainty within the Bayesian framework. This graphical structure is defined through a set of random variables, that are represented through nodes, and their direct relationships, that are encoded through arcs. These arcs could represent different kinds of connection, such as causal ones. Regardless of the kind of reasoning the network is modelling, arcs represent conditional dependencies, while the absence of such arcs translates in conditional independence of the variables represented by the nodes. When an arc between the nodes X and Y is specified as $X \rightarrow Y$, then the node X is defined as the parent node of Y.

Let us define a set of random variables $X = \{X_1, \ldots, X_n\}$ with n number of BN nodes and denote the universe of disclosure of such variables by Ω_i, where t_i is the number of possible states for X_i and $i = 1 \ldots n$.

We can order the nodes such that the parents of a node X_i, also known as parent configuration, are a set of variables $pa(X_i) = \{X_q, \ldots, X_{i-1}\} \subseteq \{X_1, \ldots, X_{i-1}\}$.

Consequently, by definition the BN will represent the following joint probability distribution p over $\Omega_1 \times \ldots \times \Omega_n$:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i | pa(X_i)) \tag{2.1}$$

BNs allow to quantitatively model those relationships through conditional probability distributions associated to each variable (usually represented as CPTs). In BNs a node is conditionally independent of its non-parent ancestors, therefore

$$p(X_i | X_1, \ldots, X_{i-1}) = p(X_i | pa(X_i)) \tag{2.2}$$

For further details on BNs the reader is referred to the basic literature on the topic, such as [120]. The KE task [122] entails the definition of: (a) the relevant variables; (b) their relations (i.e. BN *structure*) and (c) the conditional probability distributions (i.e. BN CPTs).

This last point consists in specifying for each X_i the expression in Equation 2.2. More specifically, we can consider each node described by a CPT denoted through a vector θ of parameters. These are defined as:

$$\theta_{i,j,k} = p(X_i = x_k^i | pa(X_i) = \mathbf{x}_j^i) \tag{2.3}$$

where $k = 1,\dots t_i$ are the possible states taken by X_i and $j = 1,\dots,q_i$ are the possible parent configurations of node X_i. Therefore, $\mathbf{x}_j^i$ represents the set of states taken by the variables in the j^{th} parent configuration.

2.4 Bayesian network parameter learning

The CPTs are generally specified by experts, learned from data or by a mixture of the two previous approaches. When the model parameters are learned from data, the task consists in finding the most probable θ that explains the data. The Expectation Maximisation (EM) algorithm is a general approximated approach to parameter learning from data which has been introduced in [61]. This algorithm allows to perform machine leaning with missing data, such as in the case of latent variables models.

As summarised in [152], we can denote the dataset of measurements as $D = \{X^1,\dots,X^M\}$. Each X^m, with $1 \leq m \leq M$, is a realisation of the BN, such that $X^m = (X_1^m, X_2^m,\dots,X_n^m)$. If we assume independent successive realisations, the joint log-likelihood can be expressed as:

$$L(D;\theta) = \sum_{m=1}^{M} L(X^m;\theta) = \sum_{m=1}^{M}\sum_{i=1}^{n} \log \theta_{i,pa(X_i^m),X_i^m} \tag{2.4}$$

When machine learning needs to be performed with incomplete data, such as in the case of latent variables models, we can employ the Expectation Maximisation (EM) algorithm. The EM is a general approximated approach, which has been introduced in [61]. We can consider latent variables models as described by a probability distribution $p(X,Z;\theta)$, where X is the vector of observed data and Z the vector of missing data. As computing the log-likelihood of the observed data is not mathematically tractable, the EM algorithm in its MLE formulation aims at maximising the expected value of the log-likelihood of the complete data.

After initialising the parameters to a value θ^0, the EM repeats iteratively an expectation step (*E-step*) and a maximisation step (*M-step*). In the first step the current parameters are used to complete the data, using probabilistic inference. In the second step the completed data are treated as observed data and a new set of parameters is learned. More specifically:

E-step: computes the expected value of the latent variables assuming θ^r fixed; generally the following auxiliary function is computed:

$$Q(\theta,\theta^r) = E[L(X,Z;\theta)|X;\theta^r] \tag{2.5}$$

M-step: maximises the log-likelihood function (i.e. maximising the auxiliary function) to estimate the new parameters (θ^{r+1}):

$$\theta^{r+1} = \arg\max_{\theta} Q(\theta, \theta^r) \tag{2.6}$$

The iteration stops when the improvement of the log-likelihood function is below a fixed threshold or the maximum number of iterations is reached.

This algorithm is used in the work described in Chapter 5 to learn the parameters of the source reliability variable, which is a latent construct in the analysed model.

2.5 Dynamic Bayesian networks

A dynamic Bayesian network (DBN) is an extension of Bayesian networks to model probabilistically time-series [58, 154]. We can denote with $U = \{U_1, \ldots, U_T\}$ a set of random variables, which can be partitioned in $U_t = (X_t, Z_t, Y_t)$ with $t = 1 \ldots T$, where t represents the time slice, X_t represent the observed variables (i.e. input variables), Z_t represent hidden (latent) variables and Y_t the output variables of a state-space model.

It has to be highlighted how in general the term *dynamic* refers to the dynamic nature of the system modelled, but the network is not changing over time.

A DBN can be defined as a pair, $(B_1, B \rightarrow)$, where B_1 is a BN defining the prior $p(U_1)$, and $B \rightarrow$ is a *two-slice* temporal Bayes net (2TBN), which encodes the conditional probability $p(U_t|U_{t-1})$ through a DAG. More specifically, if $i = 1, \ldots, n$ are the nodes in the time slice, this relation can be expressed as follows:

$$p(U_t|U_{t-1}) = \prod_{i=1}^{n} p(U_t^i|pa(U_t^i)) \tag{2.7}$$

For notional reasons we can assume that a first-order Markov assumption holds. Therefore, the parents of a node, $pa(U_t^i)$, can be in the same time slice or in the previous one. More in general, arcs can be defined across more than two slices. We can assume that the parameters of the conditional probability distributions are time-invariant. However, parameters can change over time. Therefore, either they are added to the state-space and treated in the network as random variables or they might be treated as hidden variables (selecting the set of parameters). By *unrolling* the 2TBN over T we obtain the *semantics* of the DBN. More specifically, we can define the joint probability distribution as:

$$p(U_{1:T}) = \prod_{t=1}^{T} \prod_{i=1}^{n} p(U_t^i|pa(U_t^i)) \tag{2.8}$$

2.6 Evidence theory

Let the frame of discernment $X = \{x_1, \ldots, x_n\}$ be defined as a finite set of exclusive (2.9) and exhaustive (2.10) hypotheses:

$$\forall (x_i, x_j) \in X^2, x_i \cap x_j = \emptyset \tag{2.9}$$

$$x^* \in X \tag{2.10}$$

where $\emptyset$ is the empty set and x^* is the unknown hypothesis. The powerset of X, which is the set of all possible subset of X, is denoted by $P(X)$, such that:

$$P(X) = \{\emptyset, x_1, x_2, \ldots, (x_1, x_2), \ldots, X\} \tag{2.11}$$

If $|X|$ denotes the cardinality of X, then $|P(X)| = 2^n$. Any possible subset of X, will be denoted with a capital letter (e.g. $A \subseteq X$).

The basic probability assignment (BPA) m is a mapping from $P(X)$ to $[0,1]$ that satisfies the following conditions:

$$m(\emptyset) = 0 \tag{2.12}$$

$$\sum_{A \subset X} m(A) = 1 \tag{2.13}$$

$m(A)$ is the exact belief committed to A, which is the belief that a particular x of X belongs exactly to A. The condition set in (2.12) is also known as the closed-world assumption as it refers to the case in which the true state of the world belongs to X, therefore the information received point towards the set of possibles defined in X.

From the BPA the belief function (Bel), plausibility function (Pl) and communality function (q) can be defined. More specifically the belief function is defined as:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{2.14}$$

The belief function satisfies the following axioms:

1. $Bel(\emptyset) = 0$;
2. $Bel(X) = 1$;

3. Superadditivity (e.g. for every positive n and every collection $A_1,\ldots,A_n$ of subsets of X, $Bel(A_1\cup\cdots\cup A_n)\geq\sum Bel(A)-\sum_{i\leq j}Bel(A_i\cap A_j)+\cdots+(-1)^{n+1}Bel(A_1\cap\cdots\cap A_n)$).

The plausibility function is defined as:

$$Pl(A)=\sum_{B\cap A\neq\emptyset}m(B) \tag{2.15}$$

And the communality function is defined as:

$$q(A)=\sum_{A\subset B}m(B) \tag{2.16}$$

In the original development of the theory by Dempster [60] belief and plausibility functions were interpreted as lower and upper bounds respectively of an unknown underlying probability function. Later Shafer [204] introduced an epistemic view, which has also been embraced by Smets' [210] Transferable Belief Model (TBM). This epistemic view interprets belief functions rather as an expression of subjective uncertainty.

In the TBM the closed-world assumption is relaxed in favor of the open-world assumption in which empty set is allowed to have a non-null mass, therefore allowing the concept that information might point outside the frame of discernment. Another important concept introduced by the TBM is the differentiation between two levels at which beliefs play out, namely the credal level and the pignistic level. At the credal level the beliefs are quantified and aggregated through the belief functions. Subsequently a probabilistic transformation is applied to the resulting belief functions at the pignistic level, which corresponds to the level at which decisions are taken. This allows the use of probabilistic decision theory. There are different kinds of transformations that can be applied (e.g. plausibility transformation [41]), however the pignistic transformation is outlined here as it is the one that is going to be used in future analysis.

Given a BPA m, the pignistic probability $BetP_m$ (Smets [209]), is a (additive) probability measure defined for all the subsets A of X by:

$$BetP_m(A)=\sum_{B\subseteq X}\frac{m(B)}{|B|}|A\cap B| \tag{2.17}$$

where $|B|$ is the cardinality of B.

The game outcomes modelling within the evidence framework is ongoing.

Chapter 3

Game science: wargames, serious games, simulation gaming

3.1 Background

With respect to gaming we often encounter the terms wargame, simulation game (also known as simulation/gaming or gaming simulation) and serious game. These terms somewhat correspond also to different communities, that embody different points of view with respect to gaming. Experts have naturally converged towards these different sub-disciplines, depending on the focus of their research. However, those domains are not mutually exclusive and often there is a strong overlap. Very often those point of view appear to be complementary. In the attempt to consolidate game science, more and more authors are trying to reconcile the different points of view, which should correspond to sub-disciplines, under a broader umbrella of a science grounded on a more comprehensive theoretical foundation [119].

Game science is a very variegated and scattered discipline [119]. An example, are the findings of a fifty-year long mapping activity that has been performed by the International Simulation and Gaming Association (ISAGA) [119]. This activity has categorised the efforts within game science along two main dimensions, namely foci of interest and areas of application. The first dimension includes (i) theory and methodology, (ii) design, (iii) research methods, (iv) system development and (v) assessment and evaluation. While the second dimension includes (i) administration (business and public management), (ii) environmental, (iii) entertainment, (iv) services (e.g. healthcare, education, banking), (v) resources (e.g. human resources, cultural resources and natural resources), (vi) human settlement and geography, (vii) international relations, (viii) military, (ix) religion and (x) technology (e.g. information technology).

While for in depth analysis of the issues about the definition of *game*, *gaming* and *simulation* the reader is referred to the relevant literature, in this work we will use the following working definition excerpt from [119]:

Game: "a form of play. It is an activity involving one or more players who assume roles while trying to achieve a goal. Rules determine what the player are permitted to do, or define constraints on allowable actions, which impact on the available resources, and therefore influence the state of the game space. Games deal with well-defined subject matter (content and context)";

Simulation: "a functional model that imitates the behaviour of a reference system";

Gaming: "the common term, encompassing the terms [...] games [and] simulation".

Two aspects that might have contributed significantly to jeopardising game science are the dual nature of the viewpoints on games and the rapid advent of recreational digital games [119]. The duality in viewpoints refers to (i) the insider versus outsider (i.e. participant vs spectator) perspective; and (ii) the analytical science versus design science perspectives on games.

More specifically, from an analytical perspective games are interpreted as useful research methods for development and testing. One of the more challenging aspects of using games from an analytical perspective is the verification and validation activity (see Section 4.5). From the design science perspective, instead, games are analysed and assessed from the viewpoint of game specifications, which link to operational requirements. Therefore, the games themselves are studied in order to evaluate their development and use [89, 125]. In this case the games are assessed as artifacts within operational contexts (i.e. usability and utility).

As previously mentioned, the sudden widespread of digital games for entertainment contributed to the fragmentation of the gaming disciplines, strongly impacting on games not for entertainment.

Games not for entertainment have a long tradition. In fact, the first methods of military operational research took the form of wargames. Although there are evidences about the use of wargames in the previous centuries, the *Kriegsspiel*, which was developed in Prussia in the early 1800s, is considered the first *modern professional* wargame. With the end of the World War II, experienced military personnel transitioned into business, applying wargaming techniques to business management. In the 1950s the simulation gaming discipline began to emerge and gaming started to be applied in other fields, such as sociology, international relations or social psychology. Finally, the significant advances in computer and information

science allowed to move from the traditional game platforms, such as board games, towards higher degrees of automation in games.

While the reader is referred to the vast amount of literature in the different sub-disciplines, in the next sections we will briefly provide a description of the state-of-the-art in the different sub-disciplines to better understand where the method introduced by this doctoral work lies with respect to them.

3.2 Game typology

As highlighted in [119], each sub-discipline has proposed several typologies and taxonomies of games, based on game factors such as form, functions, activities and processes. While we do not aim at reporting on all the work done on the topic, in this section we will summarise some relevant aspects.

One important distinction is the gaming form based on the type of game rules implemented into the game [207, 119]. In fact, games can be rule-based, principle-based or free-form. In the first ones the rules are pre-defined, not questionable and have to be strictly followed. In the second ones the game rules can be interpreted by the players on the bases of underling norms. The last ones only include some basic rules (i.e. time of the game beginning, stop rules and the role of the facilitator), also known as *rules of nature*, while the rest of the game is self-organising and other game rules get negotiated between players.

Another important distinction is made on the bases of the degree of automation within the game. More specifically, a game can be *manual*, *computer-assisted* or *computerised* (i.e. serious games) [64]. An interesting typology is presented in [223], where the classification framework for computerised business simulation gaming is based on the concepts: (i) degree of *control* over the simulation by the participants and\or computer and (ii) degree of *interaction* among participants and computer. The proposed framework is based on the computerised simulation classification presented in [46], that distinguishes between:

- Computer-directed simulation: low computer-participant interaction and high computer control;
- Computer-based simulation: high computer-participant interaction and high participant control;
- Computer-controlled simulation: high participant-participant interaction and high computer control;

- Computer-assisted simulation: high participant-participant interaction and high participant control.

As computer-directed simulations, do not include gaming elements, they are excluded from the game classification framework [223]. Here *control* and *interaction* are assumed as zero-sum measures. Therefore, the two following cases are incompatible: (i) high computer control plus high participant control and (ii) high computer-participant interaction plus high participant-participant interaction. Although this classification was proposed for business simulations gaming, it can be extended to other simulation games.

Several other dimensions can serve as classification means for games. For example, based on the *format* we can distinguish between psycho-motor skill games, intellectual skill games and games of chance [72]. Based on the *interactive entertainment genre* we might differentiate between action games, strategy games, role play games, real world simulations, construction games, management games, adventure games and puzzle games [7]. Finally, based on players' relations we might classify the games as competitive or cooperative games [119].

Some authors have presented typologies based on the *purpose* of the game. For example, in [217] it has been proposed to classify them as entertainment games, educational games, experimental games, research games and operational games.

This classification resembles the one proposed by other authors such as [141], which identify games on the bases of their educational, research or operational/practical functions. Similarly in [125] the following categories of simulation games are identified: (i) education and training; (ii) policy intervention (e.g. [70]); (iii) design methods for complex socio-technical systems (e.g. [124]).

An interesting classification has been proposed in the Gameplay/Purpose/Scope (G/P/S) taxonomy for serious games [66]. This model proposes game play, purpose and scope as relevant classification dimensions. The first dimension differentiates between play-based or game-based games. The former are games characterised by a lack of well-defined objectives and rules. The latter instead, have defined objectives and rules. The purpose dimension allows classifying games on the basis of their function (e.g., message broadcasting, training, data exchange). Message broadcasting games are the ones that have been developed with the aim of broadcasting a message (e.g., educative games, informative games, persuasive games and subjective games). Games for training are developed with the purpose of improving players (cognitive or physical) performances. Finally, data exchange games have the specific purpose of supporting data exchange, such as "collecting information from [...] players" [66]. The scope dimension refers to the game market (e.g., state and government, military and defense, healthcare, education, corporate, religious, culture and art, ecology, politics,

humanitarian, advertising, scientific research) and the target audience (e.g., professionals or general public).

3.3 Games for research

3.3.1 Simulation gaming

Although, there is no formal agreed definition regarding simulation games, they can be expressed as "essentially a case study [, but] with the participants on the inside" [112]. The participants, in fact, have to play a specific role in a simulated environment [70], that imitates at different levels of abstraction the reference system. Therefore, simulation gaming is an experience involving human participants that features competition and rules [223, 221].

Often modelling & simulation (M&S) and gaming (i.e. simulation game or wargame) are used as synonymous, but it has to be noticed that those are two very distinct disciplines. M&S might be used before or after a simulation game as complementary research tool or might even be employed during a simulation game, but differently than M&S the simulation game itself requires human players [191]. The two disciplines deal with different problems. This might be summarised in the statement: M&S deals with *complication*, while wargame and simulation games more in general with *complexity* [192]. A simple explanation of the two terms can be found in [193], where *complicated systems* are described as systems composed by many non-static part, which operate in patterned ways. On the contrary *complex systems* are rich in non-static features, that might operate in patterned ways at some lower level, but do not have a patterned behaviour at interaction level. Moreover, the term should not be interpreted as a synonym of game theory [237]. Several researchers have illustrated the difference and complementary between gaming and game theory, and the reader is referred to the relevant literature (e.g. [217]).

Simulation games can have several purposes (Section 3.2), including research. Simulation gaming for research, can be used to experiment with complex systems in order to understand and predict their behaviours [69, 70].

In fact, they appear to have the ability to positively tackle important concepts related to complex systems, namely *complexity* of the system, *communication* between stakeholders, *consensus* between stakeholders, *creativity* (or innovation) and *commitment to action* by stakeholders [69, 129]. There are several characteristics that make simulation games useful research tools. For example, it has been highlighted how studying and experimenting complex problems in simulation gaming is easier and cheaper than in real settings. Moreover, they allow to set up controlled experiments and provide a *safe to fail* environment.

This last characteristic has been stressed as a fundamental advantage compared to other experimentation methods also by the wargaming community (e.g. [64]).

3.3.2 Wargames

Many definitions of wargame exist, however, one of the most widely accepted ones refers to a "warfare model or simulation whose operation does not involve the activities of actual military forces, and whose sequence of events affects and is, in turn, affected by the decisions made by players representing the opposing sides" [163]. As specified in [191], it is a simulation of "simplified [...] potential future (or perhaps past) warfare situation[s]". Therefore, it constitute a subclass of simulation games, specifically focusing on warfare aspects (i.e. prevention, fight, resolution of the situation or security assistance situations).

Beside the presence of players, other key elements that determine a wargame are the *adversarial* nature of it and the centrality of the *decision-making* process. This last aspect is emphasised by different authors, for example in [64, 63]. In fact, military wargames are tools used to examine warfighters decision-making processes at strategic, operational and tactical levels. Wargames might have either an educational-training or a research purpose. In the latter case the term analytical wargame is employed. A slightly different perspective is presented in [191], which defines wargames as "inherently a research tool", including in this definition the education, training, communication, experimentation and analysis components. In fact, it advocates that the player immersed in the research context of the wargame, gains knowledge regardless of the main objective of the wargame.

Many types of wargames have been developed along the decades, such as seminar wargames, matrix wargames or courses of action wargames. Each type of game, moreover, can present several variants related to factors such as the adjudication, control, number of sides, number of players, degree of computerisation, the representation of soft factors, amount of intelligence provided to the player and turn length. The reader is referred to the literature (e.g. [27, 224, 64]) for further details on the topic.

3.3.3 Serious Games

The first definition appeared in [6] and it refers to simulation gaming to improve education. Later the concept has been re-proposed in [197], where serious games are defined by linking serious purposes with the video gaming industry technologies. Along the years, several other definitions have been proposed [2] and it appears that there is a strong agreement on the fact that serious games are games not designed primarily for pure entertainment. Therefore, entertainment, enjoyment or fun are not their primary design objective. In fact, most of the

research and applications in the domain of serious games have focused on education, training and user learning objectives in several domains (e.g. [147, 118, 34, 190]). The reader is referred to the relevant literature and numerous surveys on the topic (e.g. [67, 130, 249, 230]) for further details. However, from the different definitions it appears that there are different perspective regarding the implementation platforms (i.e. computerised or manual) for serious games. In fact, some interpret the term serious game as encompassing any game with a purpose beyond mere entertainment, regardless of the implementation platform, while many adopt a more restrictive view. In the latter case serious games are identified as *computerised (or digital)* games with a serious purpose. Following the first perspective serious games would represent a superset of simulation games and wargames, while following the second perspective they would be identified as a subset of them.

Despite the clear dominance of education and training in this field, serious games can serve also other purposes. In fact, they can be used as communication tools or as research methods. The latter are known also as data-exchange serious games [66].

3.3.4 Games for research and experimentation

As described in Section 3.2 some authors generically refer to games designed with the purpose of collecting data, others distinguish games on the bases of the the aim of collecting data. For example, in [217] the following classification is proposed:

1. experimental games, aiming at testing theories and hypothesis without specific applications and context;
2. research games, aiming at collecting information regarding broad subject areas (i.e. forecasts), but without clear immediate application of the results;
3. operational games: aiming at collecting information to support decision making and policy implementation in well defined situations.

Further, the final goal of operational games is categorised in one of the following purposes: demonstrating principles, generating ideas, changing attitudes, testing models, forecasting, answering "what if" questions, providing dress rehearsals, establishing communication or testing personnel. However, often games do not fall only under one of the above mentioned categories. For example, there might be games that are both operational and research games, as their goal is to support decisions, planning, and policy implementation (like operational games), but are focused on several situations presenting the problem to be explored (like research games). In order to avoid unnecessary constraints, the author in this work will

use the term *analytical games* to refer to games designed to explore a research problem in its broader meaning, without differentiating between research, experimental, operational and pure data-exchange (collection of data) purposes. Moreover, regardless of the type and classification of the game, this could be designed as an experimentation method, if appropriately tailored to explore the effects (i.e. causality) of manipulating selected variables [203, 114]. Games used for experimentation can be used both to validate or generate theories and assumptions [207]

3.3.5 Games for research in engineering

Different fields of engineering have started looking at games, not only from an educational perspective, but also as supporting design tools. A literature review of games used in engineering research [230] shows how data-exchange serious games have been used to support the sharing of data between collaborating designers or between researchers and subject matter experts (e.g. [91, 21, 83, 123, 226]).

Two notable examples explore human problem-solving strategies to support computational algorithm optimisation in the context of protein structure design [43] and vehicle powertrain controller design [183]. The findings derived from the use of the two games have shown that human-derived strategies can be a valuable resource when used in conjunction with computational algorithms. A game that explores decision-making strategies to better understand engineers' biases and tendencies is proposed in [230]. Some games in the context of SAW have been developed (e.g. [88]). However, to the best of the author's knowledge, the focus of such games remains on training and message broadcasting, with the exception of [196] which presents a game for assessing team SAW and [137] which describe the use of simulation games (often serious games) for the analysis and design of complex systems. For example, one of those games (Yard Crane Scheduler [135, 136]), focuses on individual, shared and distributed SAW to understand how it impacts decision-making in operational planning of inter-modal transport operations in container terminals.

Chapter 4

Knowledge Acquisition Analytical Games

4.1 Analytical games for decision-making

From the previous chapter it is understandable that many types of game exist and several classifications of the same type of game are possible depending on the focus. The main criteria to classify the games appear to be the game platform (i.e. serious games), the domain of application (i.e. wargames) or the techniques used (i.e. simulation gaming). The purpose (i.e. training & education, communication, research) becomes a second order criterion. In this section we propose a mere shift of perspective, where the emphasis is rather on the purpose. This allows us to identify a set of games that studies decision-making, that we denote with the term analytical game for decision-making.

Figure 4.1 shows a graphical representation of this shift of perspective. It has to be noted that the dimension of the slices does not correspond to the popularity or usage of such games. As it can be seen games can be classified in two main groups, namely games for entertainment and games not for entertainment.

Both classes include simulation and non simulation games, based on the fact that they propose or not a simulated environment in which the player needs to be immersed. The simulation can present different degrees of abstraction and complexity, with respect to real environments. Moreover, the simulation game might investigate decision-making or not. On the bases of their purpose they can be classified in games for research, games for education and training and games for communication [66]. Simulation games for research, that investigate decision-making, include the class of analytical wargames. All those games can be provided through different platforms, therefore, they can be manual (or non-digital,

or not computerised or analog) games, partially computerised or fully computerised (or digital). In partially computerised games only parts of the game play are automated, for example the visualisation of game elements, the adjudication or data collection. The fully computerised games include data-exchange serious games, as per previous definition. The smaller upper slice in Figure 4.1 correspond to the analytical game for decision-making class. Those games are a superset of analytical wargames as they are specifically designed to explore decision-making. However, they do not necessary present all the characteristics that identify wargames, such as the centrality of warfighting aspects and the need to explore the full decision/making cycle. For example, the analytical games for decision-making that will be discussed in the next chapters analyse anomalous situations and analyse only some of the steps of the decision-making cycle. More specifically, those games focus on the Situational Assessment and Situational Awareness steps.

From above we can derive the following working definition:

analytical game for decision-making: *a platform independent simulation game, designed not for entertainment, which main purpose is research on decision-making, either in its complete dynamic cycle or a portion of it.*

Therefore, analytical games for decision-making are not a new kind of game, but a new term to better describe and contextualise this doctoral work, which focuses on the use of analytical games for decision-making in support to the design of intelligent systems.

From a platform perspective we can assume that the main platforms for analytical games can be categorised in three groups, namely manual, fully computerised or a mixture of both. Table 4.1, adapted from [169], summarises different game platforms and provides the respective definitions. This categorisation, originally proposed for educational games, is here extended to analytical games for decision-making.

4.2 Analytical games for knowledge acquisition

While analytical games are not a new kind of game, the use of them as KE and KA tools for information system design is an innovative concept. In fact, in the past years some games for KA have started to appear, but up to author's knowledge those games, differently than the method described in the next sections, do not primarily investigate decision-making, with the aim of designing decision support systems.

For example, in [121] three games (MovIE WIzard, Book Wizad and MovIE Gurus) for KA are presented. The aim of such games is to use human computing to discover in text narratives relations between entities, which are hard-to-extract automatically. The SpotTheLink

Fig. 4.1 Analytical games and other game types

Category			Description
Fully computerised game			Electronic game that involves human interaction with a user interface to generate visual feedback on an electronic device
	PC game	Stand-alone	Game played on a general-purpose personal computer
		Online	Game played on some form of computer network (Internet), using a personal computer
	Console game		Game played on a specialized electronic device that connects to a common television set or composite video monitor
	Mobile game		Game played on a mobile device, such as, phone, tablet media player, etc.
Manual game			Game that is not played on an electronic device
	Board game		Game that involves counters or pieces moved or placed on a pre-marked surface or "board", according to a set of rules
	Card game		Game using playing cards as the primary device with which the game is played
	Paper & pencil game		Game that can be played solely with paper and pencil
	Prop game		Game that is played using props (portable objects)

Table 4.1 Analytical game platforms (adapted from[169])

game [222], instead, was developed in order to provide a collaborative experience, able to motivate users in ontology alignment related tasks. More specifically, it looks at the definition of mappings between Semantic Web ontologies, which is still not a fully automated task. This game is a release of a more wider game framework, known as OntoGame [208]. This game framework aims at deriving best practices and guidelines for semantic-content-authoring technologies [222]. More in general, several methods have been developed with the aim of taking advantage of games to collect useful information [233]. An example are games specifically looking at image annotation (e.g. [234, 236, 235, 256, 151]), video and music annotation [208, 15, 134], semantic web (e.g. [138, 208, 140]) and commonsense KA (e.g. [39, 32, 236, 103, 128]). In the latest case, specific game engines, such as the GECKA (serious Game Engine for Common-sense Knowledge Acquisition) have been developed [31].

From the above mentioned research it appears that the potential use of games for KA has started to gain the attention from the scientific community. For example, [105] describes how games (i.e. games-for-modelling) could be a useful tool for the acquisition of highly

structured domain-specific knowledge to be used in model-based methods for AI. However, this work presents very generic results. In general, it seems that little efforts have been devoted to formalise the approaches and generalise the concepts.

This thesis introduces the *Knowledge Acquisition Analytical Game* (K2AG), an innovative game framework for KE and KA for cognitive system design. The game framework refers to analytical games for decision-making which specifically focus on information processing, Situational Assessment and Situational Awareness. Those games are used as experiments to collect knowledge to be employed in the design of AI algorithms and systems.

As described in Section 2.1, in fact, KA still suffer from several issues. Therefore, research is ongoing to improve KA techniques. The K2AG is an innovative KA technique, that steams from the concept that games are "a communication mode capable of linking tacit to formal knowledge by provoking action and stimulating experience" [86]. In the remainder of this thesis we will use the acronym K2AGs to refer to the games developed following the K2AG framework.

4.3 Knowledge Acquisition Analytical Games elements

Literature lists several basic elements of simulation games, which embody the designer decision on *how* to game. Those elements are: scenario, pulse, cycle sequence, step of play, rules, roles, model, decision sequence and linkage, accounting system, indicators, symbology and paraphernalia [69]. The term pulse identifies an event or problem that is introduced in the play to focus the attention of the player on the important aspects of the problem. The cycles sequences include both the micro and macro cycles that characterise any game (e.g. introductory cycles or evaluation processes). The steps of game represent the progression in the game activities, while the accounting system is the set of fixed game procedures introduced to handle consistently players' decisions. The indicators are the aspects of the accounting systems that are used to report on the progress within the game. Symbology is how the indicators are represented and, finally, the paraphernalia are the additional game elements that are required to execute the game (i.e. flashcards, charts, forms or pens).

While the consideration on the K2AG overall design will be discussed in the Section 4.4, this section will focus on some important components that strongly characterise K2AG method. More specifically, the scenario, the knowledge cards, the game questionnaires and a data gathering method that has been developed to collect data relevant to the KA issue. The knowledge cards are used to convey knowledge to the player and trigger their assessments, therefore, they can be interpreted both as paraphernalia and forms of pulse.

4.3.1 Narrative and scenario

Scenario design is one of the most important elements of the design phase. In fact, as described in [164] games are "participatory narrative experiences". More specifically, one of the powerful aspects of games is the phenomenon known as *entre deux* in literary theory, that is the entrance of the person exposed into the narrative to an in-between world, where the narrative is perceived as real and reality is perceived in the background [164]. This leads to the "suspension of disbelief" [42], in which the person believes what is not there. In analytical games, such as in wargaming and simulation gaming in general, players need not only to assist to the narrative, but are actually requested to be active while in the *entre deux*. The overall narrative is composed by a *presented narrative* that is developed by the designer and a *constructed narrative*, which results from the active role of the player (e.g. statements, decisions and actions) [164]. It has been observed how in this condition, especially in high-engagement games, participants do not only make choices, but have the tendency to speak and explain to others (i.e. other participants or facilitator) their choices and actions, which is of high value in the experimentation setting, such as K2AG.

It is quite intuitive to understand that the scenarios and the way they are presented (i.e. audio and visual cues) impact on the presented narrative. In order to create suspension of disbelief, scenarios need to be engaging and believable [164]. This will not only impact the above mentioned phenomenon, but also the psychological validity of the game (Section 4.5). Therefore, the environment proposed within the game, hence the scenario, needs to be sufficiently realistic. In order to design realist scenarios domain knowledge is essential. The knowledge about the domain of the designer of K2AGs should be complemented with literature reviews, experts interviews and potential other sources of information, such as media.

The scenarios presented in the two use cases are maritime based and both present anomalous behaviour elements to trigger the assessment of the players. In fact, those behaviours are often associated with illegal activities at sea. In order to provide adequate scenarios, those have been validated with maritime experts, both during the design and the playtest phases. Additional detail on the scenarios used in the use cases can be found in Section 5.4.1 and Section 7.3.1.

Scenarios need to be accurate enough to provide a context adequate to the kind of reasoning to be induced in the player through the provision of knowledge cards, but it is very important to recall the importance of the impact of the game venue, that is the physical space in which the game will take place [164].

4.3.2 Knowledge cards

Knowledge cards are the mean through which messages are conveyed to the player.

Messages contain data, information, meta-data and meta-information identified in the experiment management plan (or analytical plan) (Section 4.4). Those messages support or induce the reasoning and assessment of the player. From an experimentation perspective they constitute the "experiment treatment" and the elements contained in the knowledge cards can be assimilated to the experimental factors.

We often encounter the terms data, information, source of information, meta-information, meta-data in relation to communication and decision support systems in support of maritime SAW. This section provides a brief explanation of the different terms, to better understand what knowledge cards might convey.

The first communication model can be dated back to Aristotle (before 300 B.C.), who proposed a five elements linear model, composed by the *speaker*, *speech*, *occasion*, *audience* and *effect*. More recently many different models of communication have been proposed, such as the ones by Shannon & Weiver [206], Berlo [16], Lasswel [133] and Schramm [199]. Regardless of the differences, it is possible to observe that there are three constant elements in those models: (i) source (or speaker, sender); (ii) message (or speech); (iii) receiver (or audience).

In K2AGs the receiver is the player, the message is conveyed through knowledge cards and the source is part of the game scenario. Meta-information regarding it can be included in the message.

A message can be interpreted as the container that is used to vehiculate data, information and meta-knowledge (Figure 4.2). The terms data and information are often used interchangeably and many debates still exist on whether the difference between them is functional or structural. This work will not solve this open issue and we will adopt the computer science perspective for which data need to be processed and contextualised to obtain information [182]. It is important to note that the context becomes a key element to discriminate between data and information. Therefore, a *high-level* information in one context, might became *low-level* data in another [182]. Data and information are often associated to meta-knowledge, more specifically meta-data or meta-information, which might be explicit or implicit. In fact, the associated meta-knowledge might be embedded in the sentence with which an information is vehiculated (e.g. a specific word such as *probably*) or through non-verbal communication, such as facial expressions and body posture.

A message might take different forms, for example a string of digits, a sentence expressed in natural language or an image. This format is intrinsically interlinked to the source, channel and/or receiver. The source is the element that creates and/or submits a message, the channel

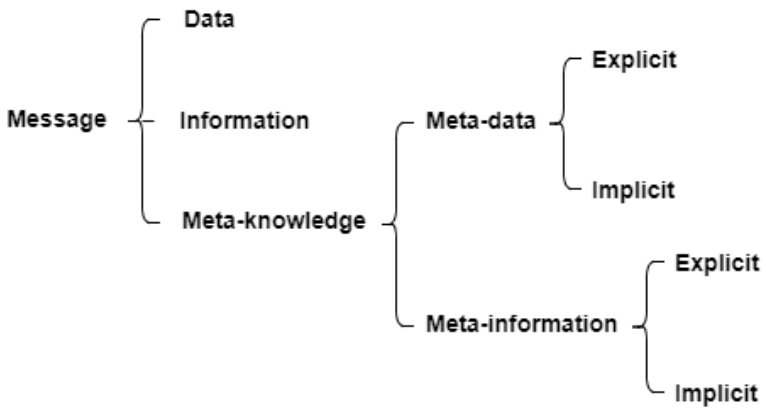


Fig. 4.2 The structure of message contents [53]

is the mean through which the message is dispatched and, finally, the receiver is the element to which a message is vehiculated.

The knowledge cards in addition to the message might contain also the data gathering area or a portion of it.

In the use cases discussed in this work knowledge cards contain messages expressed both in natural language and visual form. For additional details the reader is referred to the use case chapters.

4.3.3 Data gathering method

While there are many ways to collect data during games, depending on the data format (e.g. free text, numerical values or graphics), this section is dedicated to a specific data gathering method, called *Geometric Belief Data Gathering* (GBDG), that has been introduced in the K2AGs.

More specifically, the GBDG method takes advantage of geometrical features of simple shapes (i.e. a triangle) to collect human belief assessments that can be readily modelled within different mathematical frameworks (e.g. Bayesian or Evidential). The data gathering area might be of different sizes. For example, it can be a small portion of the knowledge cards where the players put a sign in the position that corresponds to their belief or it can be the game board itself, such as in the case of the Reliability Game. The basic concept is to select a geometrical form with as many vertices as the hypothesis towards which the players need to provide an assessment. Therefore, in case the game requests to state beliefs relative to three different hypotheses, the corresponding GBDG shape will be a triangle (Figure 4.3), while in the case of two hypotheses the triangle would degenerate in a segment (Figure 4.4).

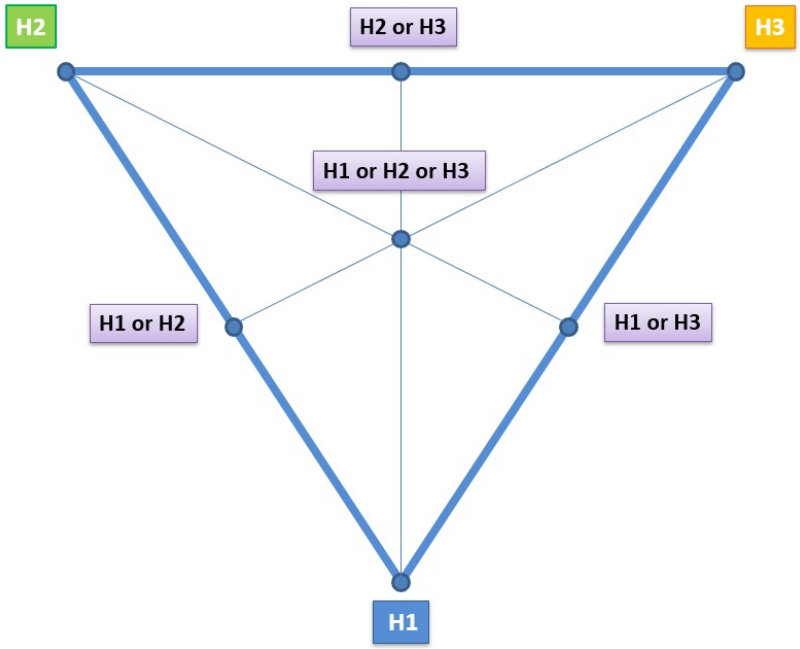


Fig. 4.3 GBDG for three hypothesis: a triangle

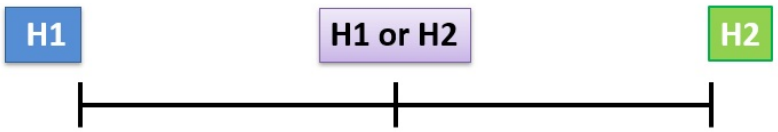


Fig. 4.4 GBDG for two hypothesis: a segment

The hypotheses are displayed in the corners and axes are included to connect all the vertices, in order to allow to collect beliefs toward any subset of the hypotheses.

The position on the GBDG shape chosen by the player (i.e. positioning the knowledge card on the board) captures the weight of belief that the information contained in a knowledge card provides toward some subsets of hypotheses. For example, if we consider the triangle, the selection of the lower corner indicates that the specific message provided by the knowledge card is pointing towards the hypothesis H_1 only. On the contrary, selecting H_1 *or* H_2 would indicate that the message is pointing towards both hypotheses (excluding H_3) and that the player could not discriminate between the two. The player can state the belief using in general every point on the triangle and internal axes. However, for experimentation reasons graphical anchors can be provided, in order to help the belief statement or to force participants to use only specific points on the bases of the experiment requirements. It is important to notice that the anchors here would be only graphical. In fact, no verbal or numerical anchors are included, as one of the strength of the method is to allow to express beliefs as a magnitude moving

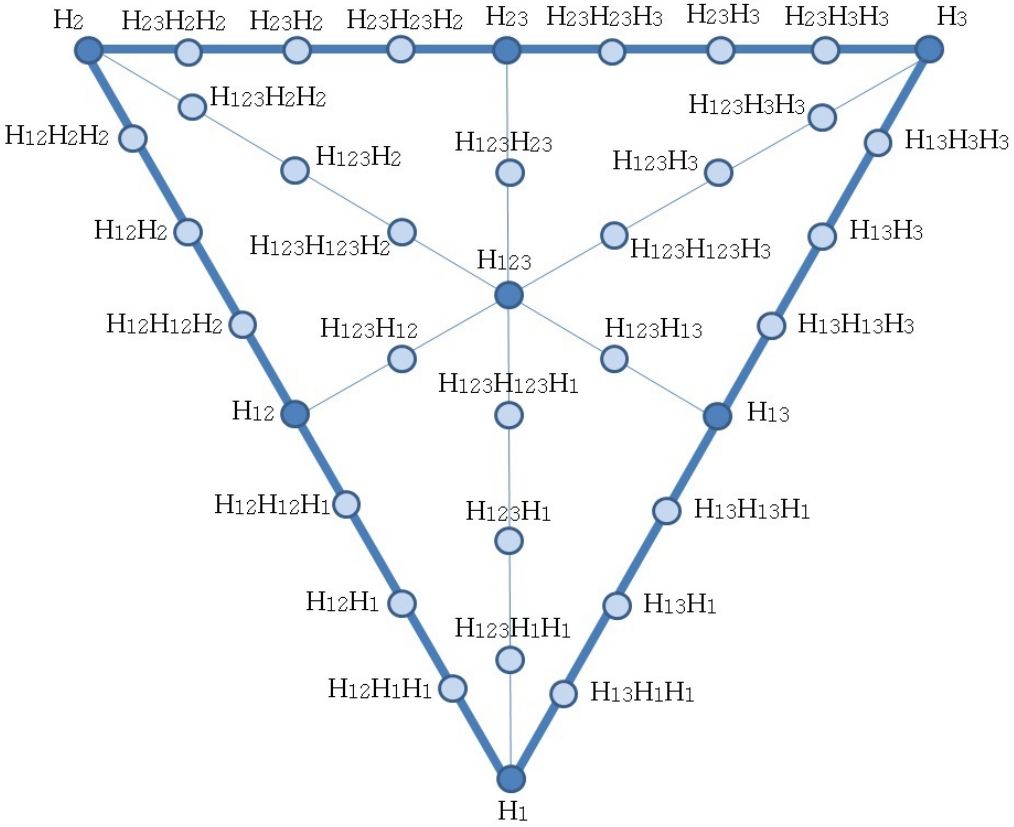


Fig. 4.5 Triangle positions codes

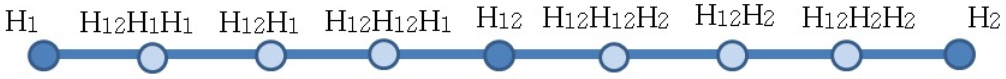


Fig. 4.6 Segment position codes

along the axes, disconnecting from numerical values and the mathematical interpretation of beliefs (e.g. subjective probability or masses).

Although numerical anchors are omitted on the GBDG shape, the post-processing of the data is relatively straightforward. In fact, the data can be modelled within different uncertainty frameworks, by transforming the position on the shape into subjective probabilities or masses. The mathematical formalisation of the GBDG will be subject to future analysis, however for a preliminary explanation the reader is referred to the literature on geometrical interpretation of probabilities and belief functions (i.e. probability simplex [45] and simplicial form of the belief space [47]). Table 4.2 present the transformation from the triangle positions (coded in

Figure 4.5) to subjective probabilities, while Table 4.3 reports the transformation to masses (Section 2.6).

In case of two hypotheses the transformation of the segment positions (coded in Figure 4.6) to subjective probabilities is reported in Table 4.4, while the one to the mass functions is shown in Table 4.5. If the two hypotheses of the segment represent the two possible states of a Boolean variable ($H_1 = False$ and $H_2 = True$) we can see that the segment mapping to subjective probabilities coincides with the probability scales often used in experiments. Therefore, the GBDG can be regarded as a generalisation of the more traditional approaches. In fact, GBDG allows to account for additional hypotheses and different modelling of the answers.

The GBDG data collection technique is an innovative feature of K2AG. To the best of author's knowledge, in fact, using geometrical shapes is an original way of answering a SAW related query and directly record the participant belief, while minimizing the invasiveness of the procedure. From the experiments it actually appears to support the assessment process, instead of interrupting it to answer to the query. Further research will, therefore, be devoted to this aspect. Moreover, the gamified approach creates an engaging context, as showed by the participant feedback regarding the games. This has a direct impact on participants' message processing mechanism. In fact, it has been demonstrated that engagement enhances the information elaboration motivation, leading to a more in-depth consideration.

Future research will investigate the advantages of the GBDG methodology compared to other methods and the limitations in terms of number of hypotheses that can be handled with such graphical representation.

K2AGs can obviously use different data gathering methodologies if deemed more appropriate to the problem at hand or a mixture of them. For example, in the Reliability Game, in addition to the triangle on the game board, the source quality ranking is recorded directly on the knowledge cards and the confidence rating is provided verbally to the facilitator.

4.3.4 Forms and questionnaires

A pre-game questionnaire is provided to the players in order to collect relevant personal data, such as demographic data (i.e. age, gender, nationality) and expertise data, such as educational level, status (i.e. civilian or military) and years of relevant experience. Additional questionnaires could be used to try to better characterise the players with respect to the experiment objectives. To this end the author is exploring the possibility of using well established questionnaires used in psychology to characterise the natural tendency of persons to enjoy thinking and, therefore, engage in thoughtful thinking (e.g. Need for Cognition [28]).

	Probabilities		
Position	$p(H_1)$	$p(H_2)$	$p(H_3)$
H_1	1.000	0.000	1.000
H_2	0.000	1.000	0.000
H_3	0.000	0.000	1.000
H_{12}	0.500	0.500	0.000
H_{13}	0.500	0.000	0.500
H_{23}	0.000	0.500	0.500
H_{123}	0.333	0.333	0.333
$H_{123}H_1$	0.667	0.167	0.167
$H_{123}H_1H_1$	0.833	0.083	0.083
$H_{123}H_{123}H_1$	0.500	0.250	0.250
$H_{123}H_{123}H_2$	0.250	0.500	0250
$H_{123}H_2$	0.167	0.667	0.167
$H_{123}H_2H_2$	0.083	0.833	0.083
$H_{123}H_3H_3$	0.083	0.083	0.833
$H_{123}H_{123}H_3$	0.250	0.250	0.500
$H_{123}H_3$	0.167	0.167	0.667
$H_{123}H_{13}$	0.416	0.167	0.416
$H_{123}H_{23}$	0.167	0.416	0.416
$H_{123}H_{12}$	0.416	0.416	0.167
$H_{23}H_2H_2$	0.000	0.875	0.125
$H_{23}H_2$	0.000	0.750	0.250
$H_{23}H_{23}H_2$	0.000	0.625	0.375
$H_{23}H_{23}H_3$	0.000	0.375	0.625
$H_{23}H_3$	0.000	0.250	0.750
$H_{23}H_3H_3$	0.000	0.125	0.875
$H_{13}H_3H_3$	0.125	0.000	0.875
$H_{13}H_3$	0.250	0.000	0.750
$H_{13}H_{13}H_3$	0.375	0.000	0.625
$H_{13}H_{13}H_1$	0.625	0.000	0.375
$H_{13}H_1$	0.750	0.000	0.250
$H_{13}H_1H_1$	0.875	0.000	0.125
$H_{12}H_1H_1$	0.875	0.125	0.000
$H_{12}H_1$	0.750	0.250	0.000
$H_{12}H_{12}H_1$	0.625	0.375	0.000
$H_{12}H_{12}H_2$	0.375	0.625	0.000
$H_{12}H_2$	0.250	0.750	0.000
$H_{12}H_2H_2$	0.125	0.875	0.000

Table 4.2 Transformation of triangle positions into subjective probabilities

As for any other type of experiment with humans it important to take into consideration ethical and security aspects. The latter one is relevant mainly in the case of defence or law enforcement related experiments. Therefore, at a minimum the players are requested to sign

	Masses							
Positions	m(∅)	m(H_1)	m(H_2)	m(H_{12})	m(H_3)	m(H_{13})	m(H_{23})	m(H_{123})
H_1	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
H_2	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
H_3	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
H_{12}	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
H_{13}	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
H_{23}	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
H_{123}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
$H_{123}H_1$	0.000	0.500	0.000	0.000	0.000	0.000	0.000	0.500
$H_{123}H_1H_1$	0.000	0.750	0.000	0.000	0.000	0.000	0.000	0.250
$H_{123}H_{123}H_1$	0.000	0.250	0.000	0.000	0.000	0.000	0.000	0.750
$H_{123}H_{123}H_2$	0.000	0.000	0.250	0.000	0.000	0.000	0.000	0.750
$H_{123}H_2$	0.000	0.000	0.500	0.000	0.000	0.000	0.000	0.500
$H_{123}H_2H_2$	0.000	0.000	0.750	0.000	0.000	0.000	0.000	0.250
$H_{123}H_3H_3$	0.000	0.000	0.000	0.000	0.750	0.000	0.000	0.250
$H_{123}H_{123}H_3$	0.000	0.000	0.000	0.000	0.250	0.000	0.000	0.750
$H_{123}H_3$	0.000	0.000	0.000	0.000	0.500	0.000	0.000	0.500
$H_{123}H_{13}$	0.000	0.000	0.000	0.000	0.000	0.500	0.000	0.500
$H_{123}H_{23}$	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.500
$H_{123}H_{12}$	0.000	0.000	0.000	0.500	0.000	0.000	0.000	0.500
$H_{23}H_2H_2$	0.000	0.125	0.875	0.000	0.000	0.000	0.000	0.000
$H_{23}H_2$	0.000	0.000	0.750	0.000	0.250	0.000	0.000	0.000
$H_{23}H_{23}H_2$	0.000	0.000	0.625	0.000	0.375	0.000	0.000	0.000
$H_{23}H_{23}H_3$	0.000	0.000	0.375	0.000	0.625	0.000	0.000	0.000
$H_{23}H_3$	0.000	0.000	0.250	0.000	0.750	0.000	0.000	0.000
$H_{23}H_3H_3$	0.000	0.000	0.125	0.000	0.875	0.000	0.000	0.000
$H_{13}H_3H_3$	0.000	0.125	0.000	0.000	0.875	0.000	0.000	0.000
$H_{13}H_3$	0.000	0.250	0.000	0.000	0.750	0.000	0.000	0.000
$H_{13}H_{13}H_3$	0.000	0.375	0.000	0.000	0.625	0.000	0.000	0.000
$H_{13}H_{13}H_1$	0.000	0.625	0.000	0.000	0.375	0.000	0.000	0.000
$H_{13}H_1$	0.000	0.750	0.000	0.000	0.250	0.000	0.000	0.000
$H_{13}H_1H_1$	0.000	0.875	0.000	0.000	0.125	0.000	0.000	0.000
$H_{12}H_1H_1$	0.000	0.875	0.125	0.000	0.000	0.000	0.000	0.000
$H_{12}H_1$	0.000	0.750	0.250	0.000	0.000	0.000	0.000	0.000
$H_{12}H_{12}H_1$	0.000	0.625	0.375	0.000	0.000	0.000	0.000	0.000
$H_{12}H_{12}H_2$	0.000	0.375	0.625	0.000	0.000	0.000	0.000	0.000
$H_{12}H_2$	0.000	0.250	0.750	0.000	0.000	0.000	0.000	0.000
$H_{12}H_2H_2$	0.000	0.125	0.875	0.000	0.000	0.000	0.000	0.000

Table 4.3 Transformation of triangle positions into mass function

an informed consent, which explains the aims of the experiment, the data protection policies and the right to withdraw the experiment.

	Probabilities	
Position	$p(H_1)$	$p(H_2)$
H_1	1.000	0.000
H_2	0.000	1.000
H_{12}	0.500	0.500
$H_{12}H_1H_1$	0.875	0.125
$H_{12}H_1$	0.750	0.250
$H_{12}H_{12}H_1$	0.625	0.375
$H_{12}H_{12}H_2$	0.375	0.625
$H_{12}H_2$	0.250	0.750
$H_{12}H_2H_2$	0.125	0.875

Table 4.4 Transformation of segment positions into subjective probabilities

	Masses			
Positions	$m(\emptyset)$	$m(H_1)$	$m(H_2)$	$m(H_{12})$
H_1	0.00	1.00	0.00	0.00
H_2	0.00	0.00	1.00	0.00
H_{12}	0.00	0.00	0.00	1.00
$H_{12}H_1H_1$	0.00	0.75	0.00	0.25
$H_{12}H_1$	0.00	0.50	0.00	0.50
$H_{12}H_{12}H_1$	0.00	0.25	0.00	0.75
$H_{12}H_{12}H_2$	0.00	0.00	0.25	0.75
$H_{12}H_2$	0.00	0.00	0.50	0.5
$H_{12}H_2H_2$	0.00	0.00	0.75	0.25

Table 4.5 Transformation of segment positions into mass function

A post-game questionnaire is also provided to the players in order to collect feedback on different aspects of the game. A specific questionnaire has been developed for the K2AG, which has been employed only in the MARISA Game, as it was not available at the time of the Reliability Game. The K2AG evaluation questionnaire is an integral part of the evaluation framework of games used for KA, which is further detailed in Section 4.6.

4.4 K2AG design

4.4.1 Game design

Game design has been widely discussed in game science research, but most of the work has focused on the design of the game artifact per se. Recently, research has recognised the need to consider different levels of design, namely the design of the game artifact and the design of the game in relation to the socio-technical systems issues it tries to inform or support [119].

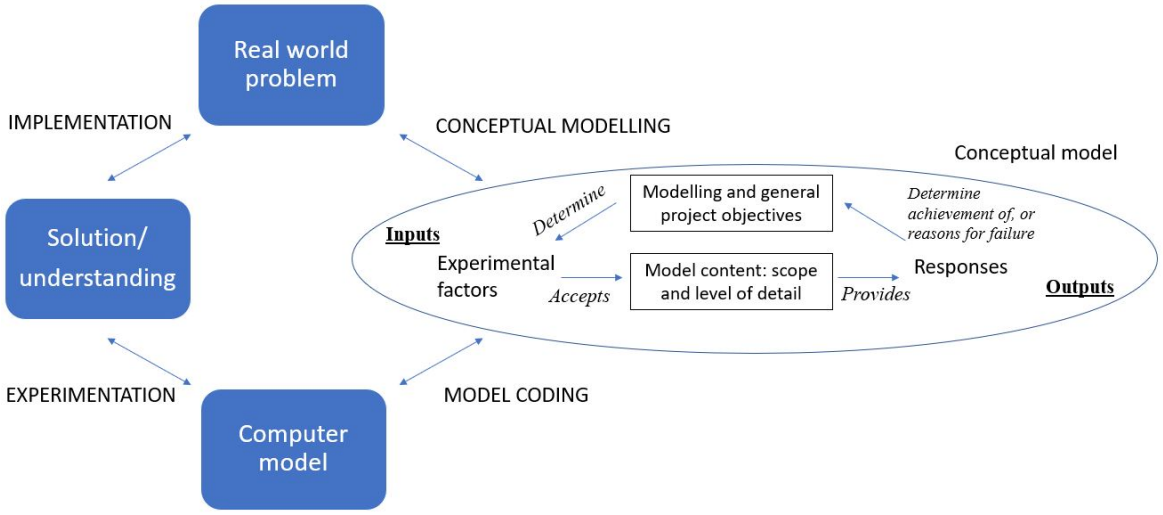


Fig. 4.7 The simulation project life-cycle as proposed in [187]

In this thesis both levels of design are taken into consideration. In fact, in both use-cases presented, both the artifact elements and the underlying KE problem are discussed. The next sections, instead, discuss high level design approaches relevant to the K2AG method and propose a unified perspective for the K2AG framework.

4.4.2 Simulation games design framework

As described in [228], which focuses on simulation-based serious gaming, one of the main design frameworks for simulation games is the one proposed in [90], which introduces an overall design concept based on the following phases: *initialisation*, *design*, *construction* and *use*. This process has been further mapped to the simulation design process proposed in [187, 188]. The iterative nature of the development of simulation-based projects is shown in Figure 4.7, while the mapping of the simulation game design framework and the simulation design can be observed in Figure 4.8. The main processes in this case are identified as the *conceptual modelling*, the *model coding*, the *experimentation* and the *implementation*. From the figures it is possible to observe how the *initialisation* and the *design* phase need to feed *conceptual modelling*.

Several authors have addressed the issue of defining what a *conceptual model* is (e.g. [259, 14, 131, 158]). More specifically, [187] defines a *conceptual model* as "a non-software specific description of the computer simulation model (that will be, is or has been developed) describing the objectives, inputs, outputs, content, assumptions and simplifications of the model". With the term objective, the author does not only refer to the overall purpose of the

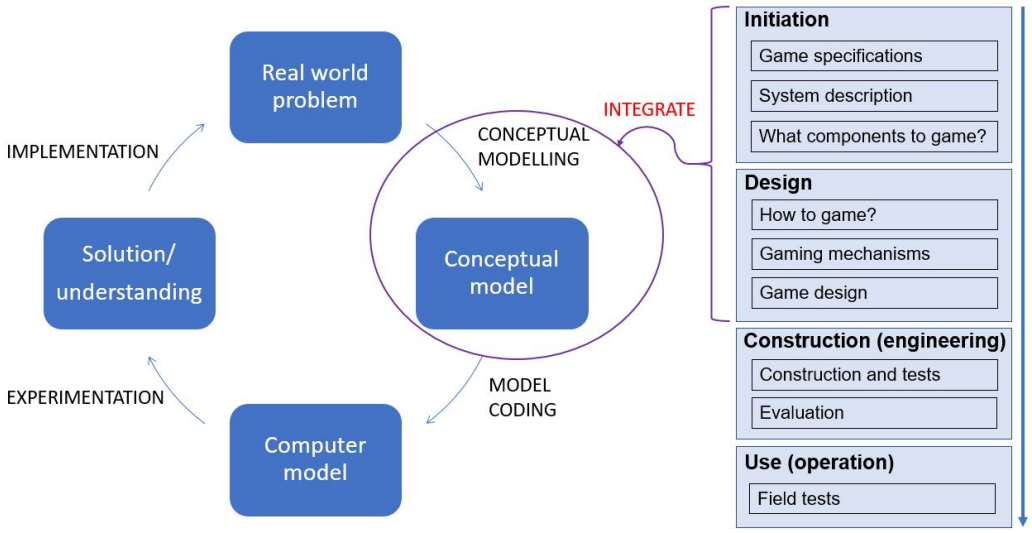


Fig. 4.8 Design framework for simulation adapted to simulation-based serious gaming (excerpt from [228])

model, but also to the project objectives, which include the time-scale, nature of the model and the model use. The inputs, also defined as experimental factors, are the factors that are altered during the simulation in order to obtain the desired observations. The model content, instead, include the scope of the model, namely its boundaries, and the level of detail. Finally, the outputs serve a double scope, namely understanding if the modelling objectives are met and in case of a negative outcome the reason for not meeting them. It has to be considered that, although software independent, the *conceptual model* might have to be adjusted as a consequence of the choices made in the following steps of the design process. Verification and validation in this functional model of game design are interpreted as parallel activities performed along the whole cycle .

4.4.3 Wargames design framework

Given that the K2AGs are a set of games that include analytical wargaming, the same overall management process adopted for analytical wargames (e.g. [27, 224]) also applies to them. This management process is shown in Figure 4.9. This figure depicts the main phases, the progression and the feedback loops between phases. This process and the different steps within the phases have been developed in order to provide a systematic and disciplined design and execution process [178], with the aim of ensuring scientific rigor, making "it possible to replicate the game, repeat the game, or iterate on some aspect of the game" [178].

Figure 4.10a to Figure 4.10c, show in detail the steps within the *define*, *design* and *develop* macro phases of a wargame life-cycle.

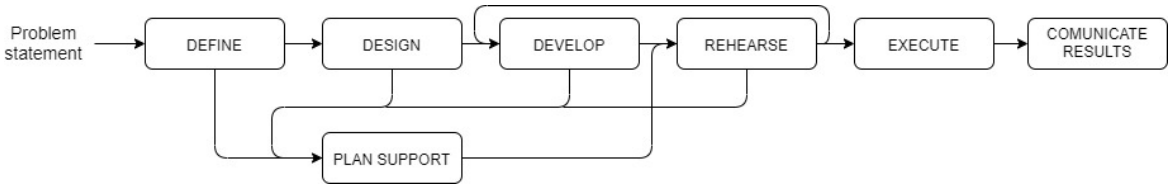
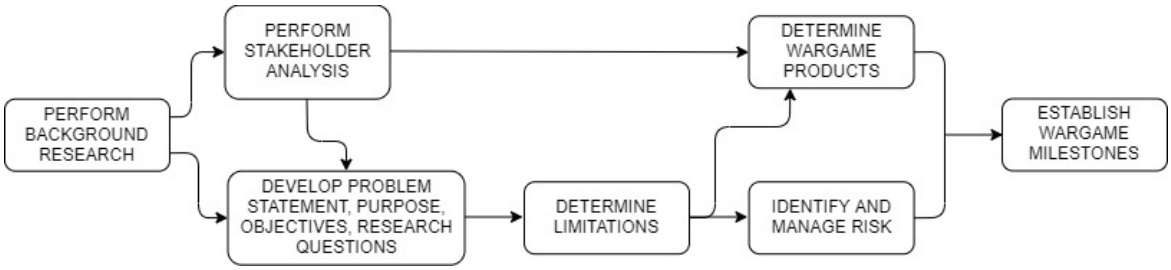
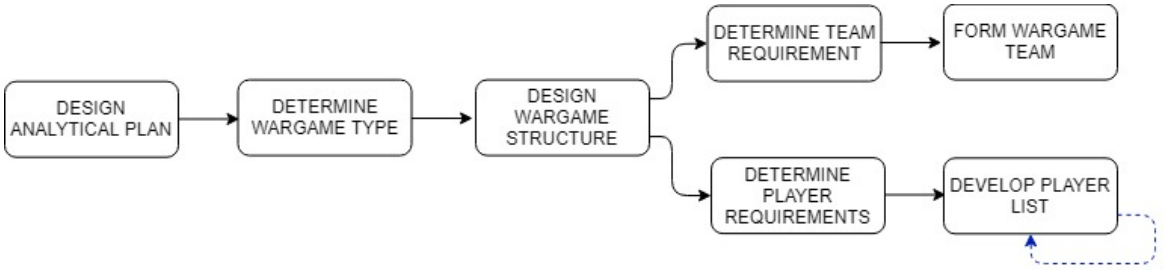


Fig. 4.9 Wargame project management process [224]



(a) Wargame project management process step: *define*



(b) Wargame project management process step: *design*

(c) Wargame project management process step: *develop*

Fig. 4.10 Steps of the wargame design life cycle (adapted from [224])

4.4.4 Model-Driven Engineering and simulations design

An interesting perspective is proposed in [240]. According to the interpretation of simulation engineering as part of software engineering, some authors propose to apply Model-Driven Engineering (MDE) to M&S (e.g. [35]). MDE differentiate between three different models.

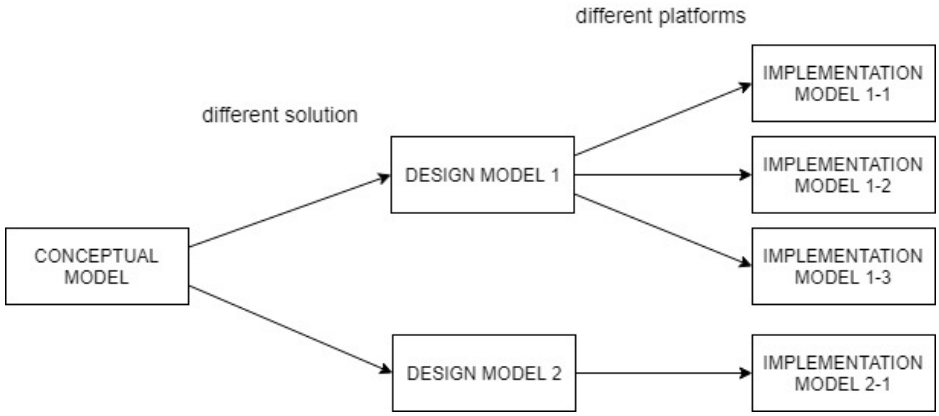


Fig. 4.11 The modelling process from a *conceptual model* to implementation models [240]

The first one, the domain model or *conceptual model*, results from the analysis phase. The second one, namely the platform-independent *design model*, results from the design phase. Finally, the platform-specific *implementation model* is the output of the implementation phase. As can be seen in Figure 4.11 a one-to-many relation exists between (i) the *conceptual model* and the *design models* and (ii) each *design model* and the *implementation models*.

Although the interpretation of the meaning of *conceptual model* might vary between the M&S, the software engineering communities and the gaming one, this distinction of the three different kinds of models is very important. In fact, this perspectives proposes to consider in a first instance a solution or computation independent model. In this phase the system design choices are not included. Moreover, the model focus is on the domain and experts' perspectives. Starting from the *conceptual model* several platform-independent computational solutions can be derived, namely the *design models*. Those can be implemented through several implementation models, corresponding to different platforms. The final selection will be guided by the appropriate design choices, such as the ones on architectural style, nonfunctional criteria, performances and adaptability.

4.4.5 K2AG design framework

The distinction between the three models is here applied to the K2AG framework. Moreover, given that K2AGs are simulation games that can be either computerised or not, we will refer to an adaptation of the definition in [187], where the *conceptual model* of a K2AG is "a platform independent description of the simulation model, describing the objectives, inputs, outputs, content, assumptions and simplifications of the model".

In a first instance it might appear that there is a perfect correspondence between: (i) *conceptual modelling* and the *define* (wargame) phase; (ii) *design modelling* and the *design* (wargame) phase; (iii) *model implementation* and the *develop* (wargame) phase.

However, the *design* (wargame) phase includes a specific step referring to the design of the analytical plan, which encompasses the analysis of the research objectives and questions, in order to define the information to be collected, and the assessment on how such information will be analysed. However, those aspects are central to the *conceptual modelling* task. More specifically, those aspects correspond to the definition of the experimental factors, as described in [187]. This entails that from a K2AG perspective, this is the phase in which the formal link with the KE problem is established and formalised. Similarly, other steps of the *design* (wargame) phase do not correspond to the *design modelling*, but rather to the *implementation modelling*. For instance, the determination of player requirements, the determination of team requirements, the development of a player list and the creation of the wargame team highly depend on the implementation platform. Therefore, these elements should be included in the *implementation modelling* phase.

Figure 4.12 illustrates the adaptation of the design framework proposed by [228] to the K2AG method. In this figure we can observe how the design phase is actually composed by the three distinct steps of *conceptual modelling*, *design modelling* and the *implementation modelling*. The figure highlights that the design could lead to different *design model* and how for each *design model* there could be more *implementation models*. For example, this is the case when a game has both a digital and a board game version. The outcomes for the different *implementation models* might not only inform directly the *solution/understanding* step, but also other *implementation*, *design* or *conceptual models*. The arrows in the picture show the feedback loops within the cycle. A feedback loop between the *solution/understanding* and the conceptual model has been introduced in addition to the ones of the original model. This is because after the *solution/understanding* not always the next phase is the *solution implementation*. In fact, further refinements to the experiment might be needed before implementing the solution or a different model might need to be implemented (i.e. game needing a digital version for collecting additional data). Therefore, a new design cycle iteration might have to be started. The same applies to the feedback loop introduced between the implementation models and the conceptual model. Finally, the relevant steps of the single design process are expanded, through the mapping with the design phases proposed for wargaming.

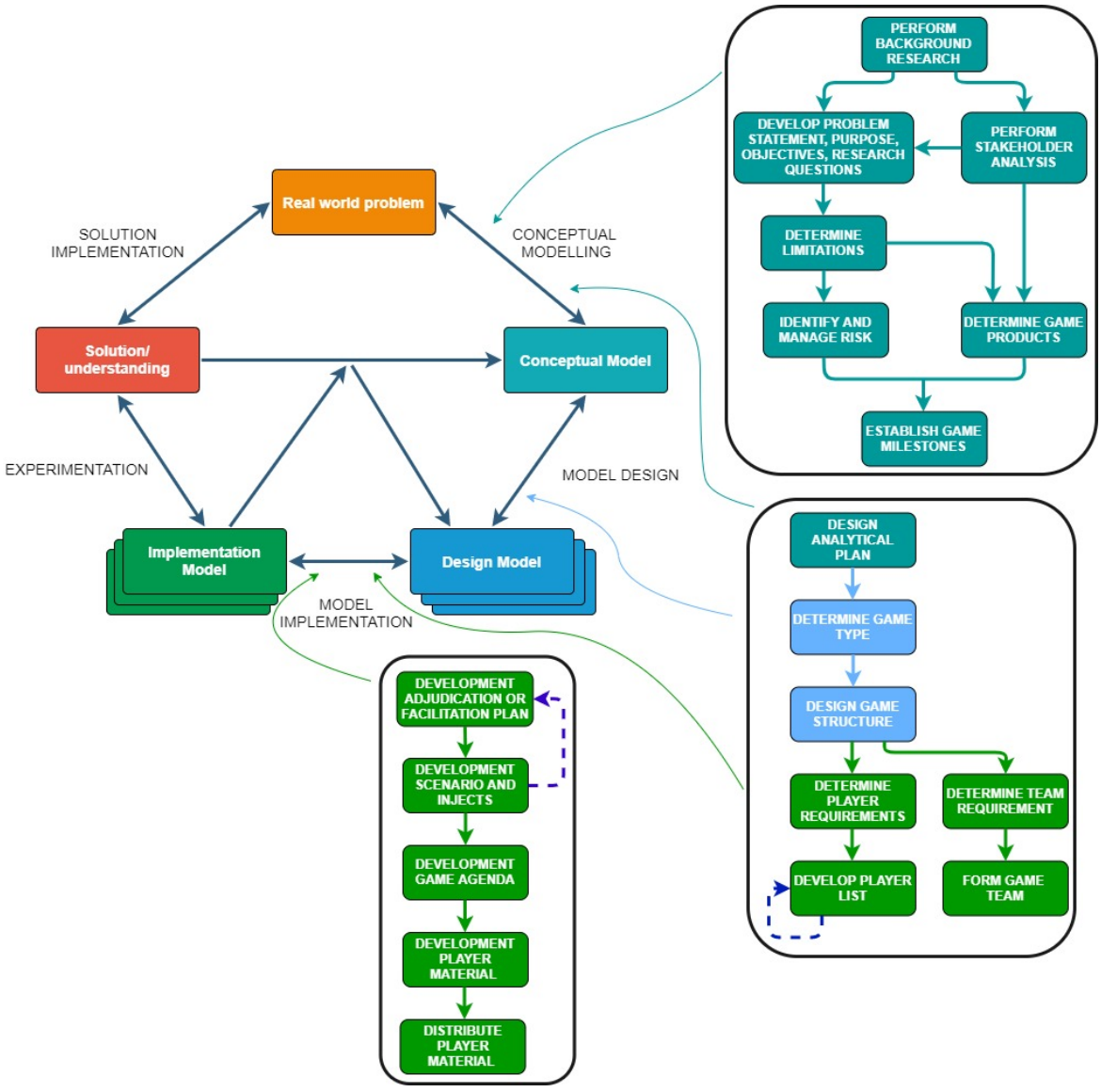


Fig. 4.12 K2AG design life-cycle

4.5 Verification and validation

Verification and validation are a critical part of any design process. Verification of K2AGs is performed all along the design cycle and culminates in the play-test phase, after the implementation modelling step. This allows to ensure that all the specified requirements are fulfilled. The concept of validation in some contexts refers to validity measurements (e.g., measurements of accuracy). Therefore, often the term evaluation is preferred to the validation one in the gaming communities. In fact, many authors (e.g. [166]), argue that validation of games in the traditional sense, namely "confirmation through the provision of objective evidence that the requirements for a specific intended use or application of a system have been fulfilled" [250] (where the term "system" refers to the game to be evaluated), is not a straightforward activity. This is not only exacerbated by the fact that often the success and quality of (war)games of high profile (consequence and investment) is the result of organisational politics, but also by the lack of a consolidated set of epistemological theory and principles [191].

As summarised in [129], there are mainly two different perspectives when it comes to validation of simulation games. One perspective adopts standardised validity criteria. An example is the use of the adapted validation principles proposed by Raser [181], namely the *psychological validity*, the *process validity*, the *structural (or construct) validity* and the *predictive validity*. *Psychological validity* refers to the need for a simulation game to provide a sufficiently realistic environment in order to generate the desired player behaviour. *Process validity* refers to the need to have game processes that are isomorphic to the ones in the reference system. *Structural validity* refers to the need to have a game structure (i.e. theory and assumptions implemented) congruent to the reference system. Finally, *predictive validity* refers to the ability of the game to provide good estimates of what could happen in the reference system, given similar conditions. The other perspective (e.g. [166, 165]), interprets validation as working systematically with the support of experts and ensuring that the design of the game is in line with the motivation and objectives of the research project, supporting the selection and use of research instruments and methods [232], without specific criteria to be met.

In addition, following [111], which discusses defence experiments with a special focus on M&S, validity can be interpreted as "fit for purpose", where this last concept is assimilated to the one of adequacy. Further to that, it summarises four main requirements that need to be met to ensure the experiment validity:

1. ability to employ the new capability (i.e. ensure that the capability works properly, that the players can use it and that the capability is actually exercised);

2. ability to detect change during the experiment;

3. ability to isolate the reason for change;

4. ability to relate results to actual operations.

For further details on the requirements, the potential threats and possible prevention to them, the reader is referred to the original document.

In the present work we argue that both perspective should constitute the basis for the the validation of K2AGs. For K2AGs psychological and process validity could hardly be achieved without the strong involvement of domain experts. Moreover, structural and predictive validity are linked to the motivation and objectives of the KA experiment.

When it comes to the motivation and objectives K2AGs are designed to serve as research methods to collect knowledge to be encoded in intelligent systems. To this end it is important to ensure that the design ensures findings that contribute to the intended knowledge [178]. Therefore, as for any experiment, during the design and execution of the K2AGs rigour needs to be enforced. As explained in [178], scientific rigour relates to the concepts of validity, reliability, replicability. Following the interpretation proposed by the author, the concept of validity can be interpreted as the extent to which the K2AG "generates findings that actually measure what the researcher intended to measure" [178]. The method reliability, not to be confused with the concept of source reliability (Section 5.2), refers to the ability of the measurement procedure to produce the same measure, when used in the same way. Finally, the replicability concept refers to the "extent to which another equally capable researcher could duplicate an analysis using the same data and reach the same conclusions" [178].

For K2AG validation this work proposes to assess the criteria summarised in Table 4.6. Those are a follow-on adaptation of Raser criteria to K2AGs. The main adaptations drive from the understanding that the reference system in the case of K2AGs are human mental models, knowledge and reasoning schemes. A branch of modern epistemological theory asserts that knowledge is the result of a mental model, built with the aim of solving problems [191]. The validity of the mental model, and therefore of the knowledge, is directly related to the utility of the knowledge in the problem solving task. Hence, as for computer simulation validity [201], valid knowledge (i.e. model) is the one that has a range of accuracy inline with the intended application [191].

Game design is an iterative process, which should include at least one playtest phase. Therefore, depending on the kind of players available for such task, in addition to the testing of aspects such as game elements and game mechanics (i.e. verification), the designer could include intermediate validation steps, in order to assess all or a portion of the proposed

Criterion	Definition	Possible assessment method
Psychological validity	Need for a simulation game to provide a sufficiently realistic environment in order to generate the desired player behaviour	Free feedback from experts; evaluation of the player experience
Process validity	Need to have game processes that are isomorphic to the ones in the reference system	Free feedback from experts; evaluation of the player experience
Structural validity	Need for a game structure which is "fit for purpose" with respect to the analytical plan	Ability to detect change during the experiment; ability to isolate the reason for change as per analytical plan; evaluation of the player experience
Predictive validity	Refers to the validity of the knowledge elicited, based on the knowledge utility as per analytical plan	Ability to detect change during the experiment; ability to isolate the reason for change as per analytical plan; ability to use the results in the actual design of algorithms or systems; evaluation of the algorithm/system designed with the use of the collected data

Table 4.6 K2AG validity criteria and possible assessment methods

validation criteria. For example, if the participants to the playtest are domain experts they could provide a first assessment of the psychological and process validity.

4.6 Evaluation of the player experience

The term player experience has been introduced to identify the adoption of user experience concepts to digital games, which are interpreted as a specific category of software [247]. Although most of the work performed on player experience (PE) specifically focus on computerised games, given that player experience refers to the quality of player-game interaction [156], the same concept can be extended to non-digital games. Therefore, it applies to the different forms of K2AG. This interaction is investigated both during and after the game.

It is interesting to notice the differentiation between the concept of game experience (GX), player experience and player type. In fact, the player experience concept is preferred to game experience one, as the experience is made by the person that plays the game. In human-computer interaction research a shift from usability to user experience has occurred, similarly in game research a shift has occurred from game experience to player experience [247]. Player experience and player type are two distinct concepts. In fact, the first one is

a dynamic construct [247], while the second one is a static construct or trait of the person [157, 247], not related to the specific game under evaluation.

When it comes to experience related to games, three macro-categories are considered: the quality of the game itself (i.e. game system experience), the quality of the human-game interaction (i.e. individual player experience) and the quality of the human-game interaction in context (i.e. social, temporal and spatial) [155].

The evaluation of the game system experience can be traced back to the traditional game testing phase, which includes software testing in case of digital-games [247]. For the player experience, both individual and contextual component, several assessment methods exist [155, 247]. Those methods are categorised in physiological methods (i.e. electo-encephalography, electro-myography, electro-dermal activity and heart rate), psychological methods (i.e. persona models, player models, surveys, verbal reports, interviews and think aloud) and behavioural methods (i.e. eye tracking, game logs, reaction time, reaction quality, observation and video recordings). Several psychological models of player experience have been proposed and the reader is referred to the relevant literature for a detailed description (e.g. [247]).

For the design of K2AG it has been decided to focus on the use of surveys as player experience evaluation as those are the cheapest, easiest and quickest assessment methods. However, K2AG could benefit also from additional evaluation methods. In fact, the K2AG evaluation Questionnaire (K2AGQ), as other post-experiment surveys, might suffer from the issue of relying on players' memories [247] and correlation with performance within the game.

Several questionnaires have been developed (i.e. Game Experience Questionnaire [177], MEC Spatial Presence Questionnaire [238], Spatial Presence Experience Scale [98], Game Engagement Questionnaire [25], EGameFlow [82] and the Core Elements of the Gaming Experience Questionnaire [30]). For an overview on such tools the reader is referred to [247].

The K2AGQ sections and items are based on two evaluation questionnaires for games with an educational purpose, namely the Game Experience Questionnaire (GEQ) [107] and the MEEGA+ model [167, 168]. The MEEGA+, is an update to the Model for the Evaluation of Educational Games (MEEGA) [195], which is one of the most used evaluation frameworks for educational games [29, 170].

The motivation for the selection of the GEQ model as the bases for the K2AGQ is linked to the fact that this model is an important tool that emphasises the multifaceted nature of the game experience. Differently than other methods (e.g. Game Engagement Questionnaire [25]), it does not focus on engagement as a player characteristic, but rather on player experience as a game evaluation instrument [161]. The GEQ questionnaire is

used to evaluate the game experience along seven quality dimensions, namely competence, sensory and imaginative immersion, flow, tension/annoyance, challenge, affect (positive and negative), psychological involvement (empathy and negative feelings), behavioural involvement, experience (positive and negative), tiredness and return to reality [107].

The MEEGA+ model decomposes the evaluation along two quality factors, namely player experience and perceived learning. This last quality factor has been omitted from the K2AG evaluation model as those games do not primarily focus on learning objectives. The original MEEGA model included a third quality factor (motivation) that in the MEEGA+ has been converted in a sub-factor of player experience. More specifically, the player experience is declined along the eight dimensions: focused attention, fun, challenge, social interaction, confidence, relevance, satisfaction and usability. This last one is further divided in learnability, operability, aesthetics, accessibility and user error protection (see Table 4.7). Those sub-dimensions are not listed in the K2AGQ, but they are addressed under the usability section of the questionnaire (Table 4.17).

The K2AGQ includes the following dimensions: overall attitude, the sensory and imaginative immersion, flow, challenge, confidence, relevance, satisfaction, workload, usability and social interaction.

Following GEQ and in contrast with MEEGA+, the immersion and flow dimensions have been introduced explicitly in the K2AGQ, because those are very important quality factors to evaluate the game experience [25, 247]. As summarised in [25], the term *immersion* describes the experience of becoming engaged in the game play and retaining different levels of awareness of the surroundings [12] or the feeling of being part of the game [252]. Flow, instead, is defined as the feeling of enjoyment in rewarding activities, when there is a balance between the challenge and skills [149, 150].

Table 4.8 to Table 4.17 show the different sections of the K2AGQ. In addition to this sections, an area is provided for free text general feedback.

The question items of the *overall attitude* dimension (Table 4.8) have been compiled on the bases of some GEQ - Core Module questions. This section summarises what other authors call fun, positive affects and negative affects.

The *sensory and imaginative immersion* dimension (Table 4.9) includes items from the GEQ - Core Module and one item from the *focused attention* dimension of the MEEGA+ questionnaire.

The *flow* dimension (Table 4.10) includes items from the GEQ - Core Module. However, some have been rephrased following the statements proposed in the *focused attention* dimension of the MEEGA+ questionnaire.

Dimension	Definition
Aesthetics	Evaluating, if the game interface enables pleasing and satisfying interaction for the user [109]
Learnability	Evaluating, if the game can be used by specified users to achieve specified goals of learning to use the game with effectiveness, efficiency, freedom from risk and satisfaction in a specified context of use [109]
Operability	Evaluating the degree to which a game has attributes that make it easy to operate and control [109]
Accessibility	Evaluating, if the game can be used by people with low/moderate visual impairment and/or color blindness [109]
User error protection	Evaluating, if the game protects users against making errors [109, 82]; applied only for evaluation of digital games
Focused Attention	Evaluating the attention, focused concentration, absorption and the temporal dissociation of the students [115, 195, 246]
Fun	Evaluating the students' feeling of pleasure, happiness, relaxing and distraction [177, 195]
Challenge	Evaluating how much the game is sufficiently challenging with respect to the learner's competency level [219, 195]; the increase of difficulty should occur at an appropriate pace accompanying the learning curve; new obstacles and situations should be presented throughout the game to minimize fatigue and to keep the students interested
Social Interaction	Evaluating, if the game promotes a feeling of a shared environment and being connected with others in activities of cooperation or competition [82, 195]
Confidence	Evaluating, if students are able to make progress in the study of educational content through their effort and ability (e.g., through tasks with increasing level of difficulty) [115, 195]
Relevance	Evaluating, if students realize that the educational proposal is consistent with their goals and that they can link content with their professional or academic future [115, 195]
Satisfaction	Evaluating, if students feel that the dedicated effort results in learning [115, 195]

Table 4.7 MEEGA+ quality dimension definition (excerpted from [167])

The *challenge* dimension (Table 4.11) includes one item from the GEQ - Core Module and the three items adapted from the *challenge* dimension of the MEEGA+ questionnaire. The item from the GEQ - Core Module, has been rephrased in order to present a positive connotation in line with the other statements of the questionnaire. In fact, the original statement was presented in its negative form.

The *confidence* dimension (Table 4.12) includes the two items from the *confidence* dimension of the MEEGA+ questionnaire. Moreover, it includes a specific question regarding

the impact of the facilitation on the players' confidence. Facilitation is generally a factor that is overlooked in the evaluation questionnaires, probably because most of the research has been performed with respect to digital games, which do not include the human facilitation component.

The *relevance* dimension (Table 4.13) includes the four items from the *relevance* dimension of the MEEGA+ questionnaire. However, given that the statements refer to learning objectives, they have been rephrased to align to the main objective of K2AGs, namely KA.

The *satisfation* dimension (Table 4.14) includes the three items from the *satisfaction* dimension of the MEEGA+ questionnaire, adapted to fit the K2AG objectives. Additionally, two statements from the GEQ - Core Module have been included.

The GEQ - Core Module includes statements that refer to the the amount of effort needed to participate into the game. The amount of physical and cognitive demand posed on the player is an important factor to be considered in players experience, which up to author's knowledge, has not been included explicitly in other questionnaires. Cognitive workload is especially relevant in K2AGs, for which physical workload is minimal, but high cognitive demand can be posed on the player, given that players are repeatedly requested to perform assessments related to new information. In order to appropriately account for *workload* a specific section of the questionnaire (Table 4.16) has been introduced, which reports the statements of the NASA Task Load Index (NASA-TLX) [106], which is a widely used and validated subjective workload assessment tool.

The *usability* dimension (Table 4.17), refers to the game as a system for which the design needs to ensure a proper interaction with the user. Therefore, the statements included in this section of the questionnaire are an adaptation to the game concept of the more general Questionnaire for User Interaction Satisfaction (QUIS) [38, 97], which is a tool to assess subjective satisfaction for human-computer interfaces.

Finally, the *social interaction* dimension (Table 4.15), is adopted as in the MEEGA+ questionnaire.

All the items are rated, as in the MEEGA+, on a five-point Likert scale, with verbal anchors (1 - not at all, 3 - moderately, 5 - extremely)[56].

Due to the limited number of participants to the MARISA Game an extensive validation of the questionnaire has not been possible at the time of writing. Moreover, it has to be mentioned that the feedback questionnaire delivered during the MARISA Game was a preliminary version, which did not include the *social interaction* dimension. Future research will be devoted to the consolidation and validation of the K2AGQ.

Overall attitude	1	2	3	4	5
I had fun					
I felt content					
I felt good					
I felt bored					
I enjoyed the game					
It gave me a bad mood					
I felt annoyed					
I felt pressured					

Table 4.8 K2AG post-game questionnaire section on overall attitude

Sensory and imaginative immersion	1	2	3	4	5
There was something interesting at the beginning of the game that captured my attention					
I was interested in the game story					
I felt imaginative					
I felt I could explore things					
It felt like a rich experience					
I found it impressive					

Table 4.9 K2AG post-game questionnaire section on sensory and imaginative immersion

Flow	1	2	3	4	5
I was fully occupied with the game					
I was deeply concentrated in the game					
I was so concentrated in the game that I lost track of time					
I forgot about my immediate surroundings while playing the game					

Table 4.10 K2AG post-game questionnaire section on flow

Challenge	1	2	3	4	5
It was easy					
The game is appropriately challenging for me					
The game provides new challenges at an appropriate pace					
The game does not become monotonous as it progresses					

Table 4.11 K2AG post-game questionnaire section on challenge

Confidence	1	2	3	4	5
When I first looked at the game I had the impression that it would be easy					
The content and the structure of the game helped me to become confident that I would support the stated goal					
The facilitation approach of the game helped me to become confident that I would support the stated goal					

Table 4.12 K2AG post-game questionnaire section on confidence

Relevance	1	2	3	4	5
The game contents are relevant to my overall interests					
It is clear how the game contents are related to the stated goal					
The game is an adequate experimentation method for the project					
I prefer providing support to projects with games to supporting it with other means (e.g. interviews)					

Table 4.13 K2AG post-game questionnaire section on relevance

Satisfaction	1	2	3	4	5
Completing the game gave me a satisfying feeling of accomplishment					
I felt competent					
I felt skillful					
I feel satisfied with the experience (e.g. supporting through the game the project with expertise)					
I would recommend this game to my colleagues					

Table 4.14 K2AG post-game questionnaire section on satisfaction

Social interaction	1	2	3	4	5
I was able to interact with other players during the game					
The game promotes cooperation and/or competition among players					
I felt good interacting with other players during the game					

Table 4.15 K2AG post-game questionnaire section on game social interaction

Workload	1	2	3	4	5
How much mental and perceptual activity was required (e.g. thinking, remembering, calculating, searching, etc.)? Was the task easy or demanding, simple or complex?					
How much physical activity was required (e.g. pushing, pulling, controlling, etc.)? Was the task easy or demanding, slack or strenuous?					
How much time pressure did you feel due to the pace at which the tasks or task elements occurred?					
Was the pace slow or rapid?					
How successful were you in performing the task? How satisfied were you with your performance?					
How hard did you have to work (mentally and physically) to accomplish your level of performance?					
How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?					

Table 4.16 K2AQ post-game questionnaire section on workload

Game usability	1	2	3	4	5
How easy it was to interpret (e.g. read and understand) the game items?					
How clear was the organization of the overall layout?					
How clear was the sequence of "screens" presented?					
How consistent was the use of terms throughout system?					
How much the game terminology is related to the task you are doing?					
How much is the position of messages consistent on the game layout?					
How clear are the requests for input by the player?					
How much are you kept informed of what the facilitator is doing?					
How helpful are the instructions that you receive when you make an error?					
How easy is it to learn to play?					
How easy is it to explore new features by trial and error?					
How easy is it to remember names and use of commands?					
How easy is it to perform the task in a straight-forward manner?					
How helpful are the help messages during the game?					
How clear are the supplemental reference materials?					
How fast is the game?					
How easy is it to correct your mistakes?					
How much are experienced and inexperienced users' needs taken into consideration?					
How good are the use of colors and sounds?					
How good are the feedbacks received during the game?					
How pleasant are the game response to errors?					
How good are the game messages and reports?					
How much are the game clutter and interface "noise"?					

Table 4.17 K2AG post-game questionnaire section on usability

Chapter 5

Case study: the Reliability Game

5.1 Motivation

In this chapter we present a specific K2AG implementation, called the Reliability Game, which aims at informing the design of multi-source information fusion systems. More specifically, this K2AG aims at collecting data to be used in further research of source factors impact on human SA and consequent SAW. The final goal is to understand the human information strategies and replicate it in fusion algorithms.

The term source factor is used in this chapter with the specific meaning of element that characterises a source of information, such as its type (e.g., radar, human operator and historical databases), quality, reliability or attractiveness. It should be noted that although the target audience is professionals [66], namely subject matter experts in maritime SAW, it could be extended to general public through the development of an appropriate scenario.

Details with respect to the notion of source reliability are included in Section 5.2. The design approach and choices are illustrated in Section 5.4. The game outcomes, which demonstrate the effectiveness of the game design, are described in Section 5.5, while Section 5.6 present a discussion on the results of the validation activity.

Further, this chapter illustrates how K2AGs appear to be a promising tool to support the KEBN, by discussing the use of data collected through the Reliability Game to learn the parameters of the source reliability variable, which in most cases is a latent variable.

A summary of related work on KA techniques and the Bayesian network formalism is provided in Chapter 2. This chapter, instead, expands on the KE task as follows. Section 5.3, presents the KE problem that the Reliability Game is trying to address. Section 5.7 details the Reliability Game KEBN. More specifically, in this section the design of the computational model for the Reliability Game is explained, together with an explanation on how the learning of the source reliability variable is performed trough Expectation-Maximisation. Section 5.8

reports on some discussions on those results. Finally, Section 5.9 presents an example of potential use of the results.

5.2 The reliability concept

For a proper consideration of reliability in the fusion process, it is helpful to understand what source reliability is and to define the underpinning elements that are central to its quantification. There is no universal definition of source reliability and even fields that have traditionally been working with multi-source information such as military intelligence neither have come to a definition, nor to a formalisation of the concept, nor to an agreement on the rating of the source reliability [53]. Following [5], reliability is defined as the "ability to rely on or depend on, as for accuracy, honesty and achievements". It is important to underline that the term ability does not represent an ability of the source itself, rather our bet on the ability to rely on it. Therefore it is our own estimate, which is a function of many factors including the capacity and/or willingness of the source of providing good information. In the field of intelligence source reliability is evaluated on the basis of past meta-knowledge and experience with the specific source. However, in general it might depend on several other factors, such as similarity, perceived expertise, attractiveness [23] of the source or experience with analogous sources (encapsulated in source type). In this work we will follow the working definition of source reliability proposed in [53]: the degree of confidence that can be put on a specific source of information. Confidence, in turn, can be defined as the state of feeling certain about the truth of something [3].

The purpose of the work described in the following sections is to understand how source factors, specifically source type and source quality, impact SA and SAW. The source reliability is treated as a latent variable, therefore is never specifically mentioned in the Reliability Game execution.

5.3 The knowledge engineering problem statement

In addition to the challenge of a desirable human-centered system design approach [94], the systems that support SAW have to deal with an ever increasing volume and velocity of the information, coupled with an increase of the variety of the information and corresponding sources with a potential lack of veracity. Information available in the maritime domain, for example, can come from a variety of sources. Common sources of information are the Automatic Identification System (AIS), the Long Range Identification and Tracking system (LRIT), radar, the Vessel Monitoring System (VMS), the Vessel Traffic System (VTS),

human operators, eye witnesses and social media. Data and information fusion technologies come into play to support operators' SA and reduce the information overload. To this end information aggregation (e.g. data fusion) approaches have proven to be effective, provided that the outputs are presented in an intuitive and actionable format that engenders trust [248, 142]. To get full advantage of the variety of sources beyond the ones traditionally in use, we need not only to combine them but also to correctly account for source factors in fusion processes [189, 174, 175]. With respect to source reliability, most mathematical fusion operators assume that the sources are fully reliable or at least equally reliable and therefore assign an equal weight on the resulting combined belief assessment [92]. In reality this assumption is not always satisfied and sources can differ in reliability. Several strategies within different uncertainty frameworks (e.g., Bayesian, belief functions) have been proposed to account for partially reliable sources. Generally, the consideration of source reliability in the fusion process relies on discounting, pruning or reinforcement operations [189, 92, 174, 175], allowing for instance to completely discard a piece of information provided by an unreliable source or to strengthen the weight of information originating from a highly reliable source. However, further research is needed to clarify some concepts related to source reliability, to clarify the semantics of source quality dimensions and to ensure that the implementation of those reliability accounting strategies in current support systems meets some criteria of understandability or intuition.

Within the Bayesian framework two different types of models have been proposed to account for source reliability. More specifically, one family of models is based on the assumption that reliability should be treated as an *exogenous* variable (e.g. [17, 18, 44]), while the other treats it as an *endogenous* variable (e.g. [20, 81, 87, 93]). This work follows the latter view as it aims at explicitly investigate how such variable behaves.

The hierarchical model in Figure 5.1 is proposed in [20]. In this model both the evidence report (Rep_H) and the source reliability (Rel) are captured explicitly, while the distinction between the real evidence and the report about the evidence is embedded in the relation between the hypothesis (H) and the evidence report (Rep_H).

In [132] and [79] the authors proposed the use of basic causal structures, defined as *idioms*, to construct BNs to reason about legal argument. One of those *idioms* is the *evidence-accuracy*, in which they propose an equivalent structure to the one in [20]. However, the variable that refers to source dimensions is defined as *accuracy*. The interesting aspect is that they further define *accuracy* of evidence as a function of *objectivity*, *competence* and *veracity*. Following this approach, we can extend the model proposed in [20] and account for the different components that build up reliability (Figure 5.2). We will refer to this kind of BN structure as the *reliability structure* in the remainder of the thesis.

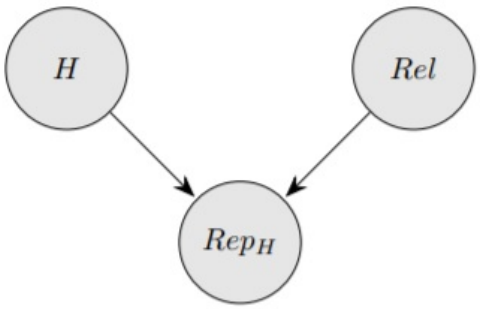


Fig. 5.1 Reliability as an endogenous variable [20]

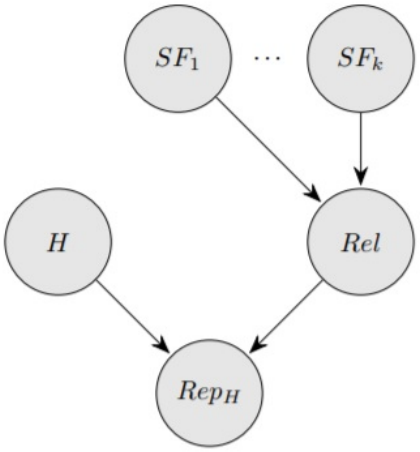


Fig. 5.2 Reliability and source factors as parents

While research has concentrated on the structure of BNs including source reliability, it appears that the characterisation of the strength of the relation between the variables (i.e. conditional probability distributions) has been overlooked. With this problem in mind the author will show how K2AGs can contribute to this research area.

5.4 The Reliability Game overall design

The Reliability Game core is reasoning under uncertainty with information provided by sources of different type and quality, which are assumed as two underpinning factors of source reliability. The aim of the Reliability Game is to capture the impact of source factors on human SA. One of the final goals is to inform the design of automated reasoners to be included in multi-source information fusion systems. This section summarises the design of the Reliability Game. More specifically, the following subsections provide details about

the world design (Section 5.4.1), the system design (Section 5.4.2) and the content design (Section 5.4.3). The world design is defined [22] as “the creation of the overall backstory, setting and theme”, while the “creation of rules and underlying mathematical patterns” is identified under the definition of system design [22]. Finally, with the term content design we refer to the “creation of characters, items, puzzles and missions” [22]. The Reliability Game design follows a mechanic-driven approach, which starts from the definition of the game core, followed by the selection of specific game mechanics (G). The following Gs have been identified in an early stage of development: (G1) assessment of hypotheses relative to a missing vessel and (G2) use of cards to communicate messages to the player. Those two mechanics were selected as they proved to be effective elements proposed in the Risk Game [113], which is a game that focuses on information quality dimensions, namely accuracy, precision and trueness. Following [108], the term accuracy can be defined as the "closeness of agreement between a test result or measurement result and the true value". The term precision refers to the "closeness of agreement between independent test/measurement results obtained under stipulated conditions" and the trueness refers to "closeness of agreement between the expectation of a test result or a measurement result and a true value", where the measurement is the information. Differently from the Risk Game, the Reliability Game assumes as fixed such dimensions in order to obtain a more rigid experiment control. This choice has been driven by the need to isolate the experiment variables (source type and source quality) impact on the final players’ belief assessment.

5.4.1 World design

The game is set in a maritime scenario and refers to a fictitious geographical area with sea portions under the sovereign of three different countries:

(L1) Right Land is a failed and poor state;

(L2) Centre Land has a good economy, although it suffers from disorders due to the vicinity to Right Land;

(L3) Left Land is a stable and rich country, thanks to the presence of oil field and extraction facilities within its Exclusive Economic Zone (EEZ) and to the Left Land Canal, which is a strategic waterway owned, managed and maintained by the Left Land government.

The player is part of Left Land Maritime Authority, which is the only authority with responsibilities within Left Land territorial waters. Therefore, it is responsible for maritime safety, maritime security, environmental protection, customs and port state control. More

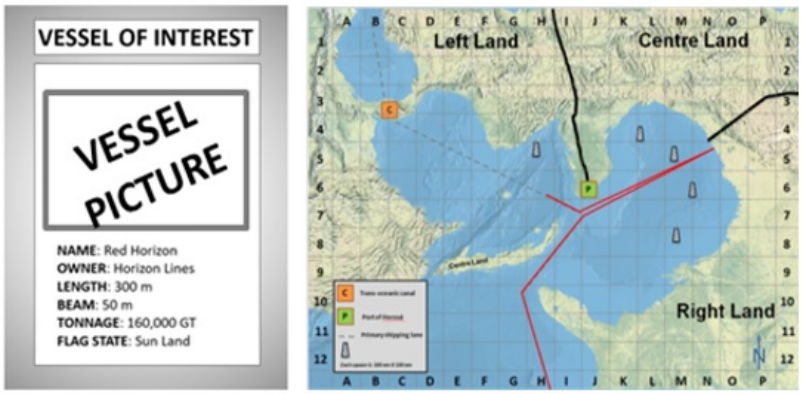


Fig. 5.3 MV Red Horizon information (vessel of interest) and its track before AIS contact loss as displayed in the scenario map by the red line

specifically, the player embodies the head of the monitoring department, who is informed by a subordinate that the Automatic Identification System (AIS) contact of the tanker ship MV Red Horizon (Figure 5.3) has been lost since six hours. The player is asked to assess what is currently happening to the ship in order to take further actions.

The player is presented with a set of three collectively exhaustive and mutually exclusive hypotheses and is asked to perform a belief assessment about what is happening to the ship on the basis of the incoming information. The three candidate hypotheses for this scenario are:

(h1) nothing is happening to the ship;

(h2) ship has a safety issue (i.e. an incident);

(h3) ship is connected to a security issue (i.e ship involved in oil smuggling).

Hypothesis h_1 would be explained by the fact that the AIS signal is not received due to a possible failure of the AIS and no intervention would be required. On the contrary h_2 or h_3 would trigger respectively a Search and Rescue (SAR) operation or a security operation. The action phase per se is not part of the game as the game stops after the SA phase. However, it represents the driver and motivation of the player.

Those hypotheses have been selected as each one represents a possible instantiation of three main high level classifications of events in connection to anomalous behaviours at sea. More specifically, the safety and security issue classes refer directly to high level institutional mandates of the Maritime Authorities and highly drive the decision-making cycle, as they demand different operational responses. Additional hypotheses selected within these classes could be adopted in the experiment (i.e. immigrant smuggling or piracy).

Source Type	Acronym
Automatic Identification System	AIS
Company Security (and Safety) Officer	CSO
Long Range Identification and Tracking system	LRIT
Service providing Patterns of Life on ship calls	PoLs Calls
Service providing Patterns of Life on ship routes	PoLs Routes
Maritime Safety Agency	Safety Agency
Intelligence	-
Radio operator	Radio
Vessel Traffic Service operator	VTS
Ship position prediction algorithm	Posit. Pred.
National ship reporting procedures	Rep. Proced.

Table 5.1 List of sources used in the Reliability Game

However, further analyses should be performed in order to better understand which is the maximum number of hypotheses that the player would be able to handle in the reasoning, without resorting to heuristic cues. Further research is also needed to assess the impact of the number of hypotheses on the data gathering method introduced with the Reliability Game. From a modelling perspective, instead, the addition of hypotheses would just result in a corresponding increase of the possible states of some modelled variables (Section 5.7).

5.4.2 System design

The Reliability Game is a single player game. Each game session is divided into four rounds ($R1$, $R2$, $R3$ and $R4$), in which a set of eleven cards is provided to the player. In each round the player is requested to assess what is happening to the ship on the basis of the available information and meta-information on source factors (source type and source quality) provided through cards. We will refer to the card as conveying a message (M), which is composed by the information (I) and associated meta-information about source factors (SF), namely source quality (Q) and type (T). The full list of source types is shown in Table 5.1.

The information provided might be true or false. Although it is not explicitly requested to assess information trueness, the player will implicitly assess this information dimension as a consequence of the game dynamics. A summary of the game state, intended as the picture of all relevant variables that may change during the play [22] is reported in Table 5.2. Table 5.3 summarises the game view, which is the portion of the game state that is visible to the player in each round [22]. With respect to the game view it can be noticed that each round is exactly the same (e.g., scenario, triggering event, information presented in the cards, order of cards), with the only exception of the meta-information about source factors. The sequence

Variable	Description	Frame
H	Hypothesis	$\{h_1, h_2, h_3\}$
M	Message conveyed by a card	$\{M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8, M_9, M_{10}, M_{11}\}$
I	Information conveyed by a card	$\{I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8, I_9, I_{10}, I_{11}\}$
IT	Information trueness	{True; False}
Q	Source quality	{1, 2, 3, 4, 5, Unknown}
T	Source type	{AIS, LRIT, CSO, Rep. Proced., Intelligence, Safety Agency, Posit. Pred., PoLs Routes, PoLs Calls,Radio, VTS}
C	Confidence level	{1, 2, 3, 4, 5, Unknown}

Table 5.2 Reliability Game state

Variable	Description	Round 1 ($R1$)	Round 2 ($R2$)	Round 3 ($R3$)	Round 4 ($R4$)
H	Hypothesis	Assessed[1]	Assessed	Assessed	Assessed
M	Message conveyed by a card	Provided[2]	Provided	Provided	Provided
I	Information conveyed by a card	Provided	Provided	Provided	Provided
IT	Information trueness	Assessed Implicitly[3]	Assessed Implicitly	Assessed Implicitly	Assessed Implicitly
Q	Source quality	Not provided	Provided	Assessed	Provided
T	Source type	Not provided	Not provided	Provided	Provided
C	Confidence	Assessed	Assessed	Assessed	Assessed

[1] Assessed = player has to assess the item and communicate it to the facilitator;
[2] Provided = item value provided to the player; Not Provided = item value not provided to the player;
[3] Assessed Implicitly = player has to assess the item but not to communicate it to the facilitator.

Table 5.3 Reliability Game view

in which cards are presented is kept constant with the purpose of controlling the information presentation order effect [257].

Each card needs to be positioned on a triangular game board (Figure 5.4), which is designed following the approach explained in Section 4.3.3.

Once all the eleven cards have been processed and positioned on the game board, the player is asked to rate the global confidence in the three hypotheses. The winning condition corresponds to the assignment of the highest confidence rate to the correct hypothesis. Details on the confidence rating can be found in next section. Figure 5.5 illustrates a diagram of a game session, explaining the main actions that the participant has to perform.

To summarise, the basic game mechanics are:

(G1) the assessment of hypotheses relative to a missing vessel;

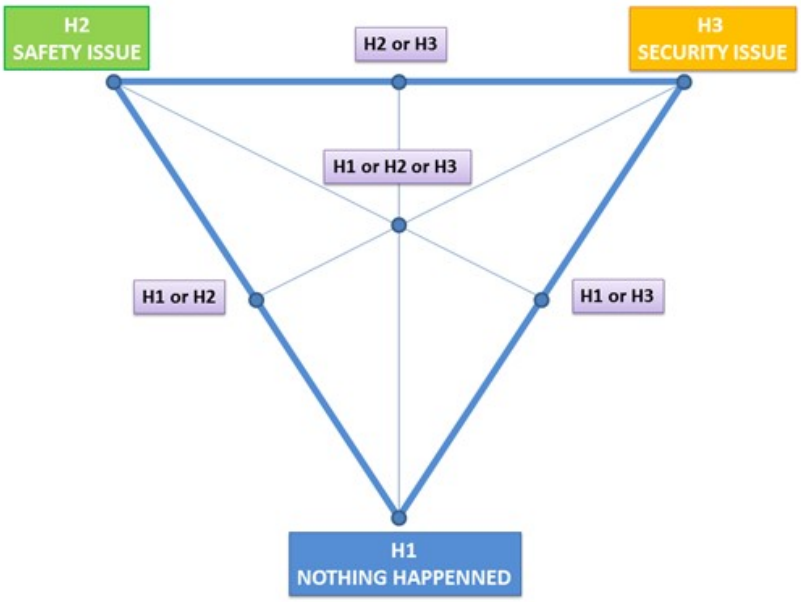


Fig. 5.4 Reliability Game board on which the cards need to be positioned

(G2) the use of cards to communicate messages to the player;

(G3) the investigation component;

(G4) the card positioning on the board to rate the support of a message towards hypotheses;

(G5) the shuffling of cards as a consequence of new evidence acquisition (optional);

(G6) the global confidence rating of the hypotheses at the end of each round.

5.4.3 Content design

At the start of the session the player is introduced by a facilitator both to the game core and to the game mechanics. During this introduction session the scenario, rules and different game elements (e.g. game board, scenario map, cards and flashcards) are presented to the player. The scenario map (Figure 5.2) depicts the geographical area and other relevant geographical contextual information, such as the location of borders, the location of oil installations and the presence of a primary shipping lane that crosses the Exclusive Economic Zone (EEZ) of Left Land, leading to the trans-oceanic channel. Moreover, it visualises the AIS track of the ship of interest before the contact was lost. The messages displayed on the cards are divided in three areas, namely the source type area, the source quality area and the information area as can be seen in Figure 5.6.

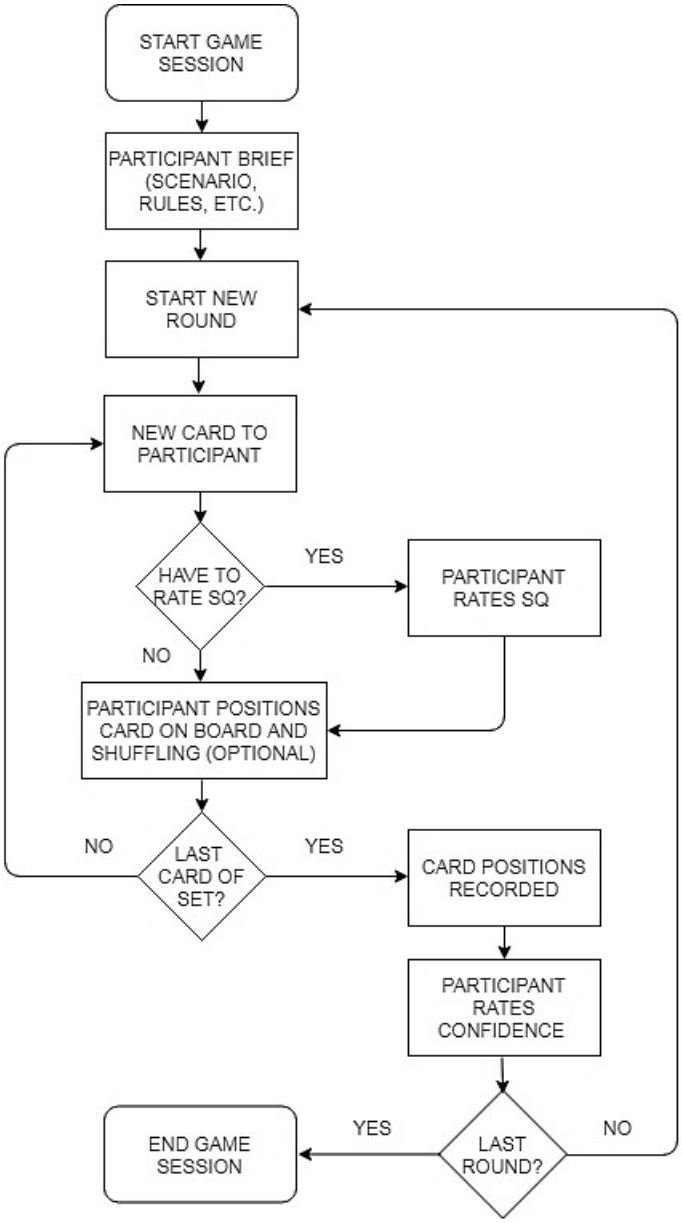


Fig. 5.5 Diagram of the a session of Reliability Game

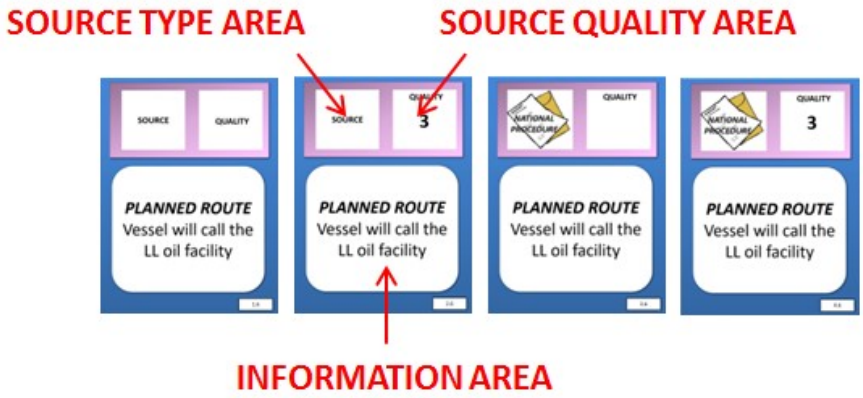


Fig. 5.6 Example of the presentation of the same message in the four different rounds

As previously mentioned, the only variation between the cards in the different rounds is in the meta-information on source type and source quality. As can be noted from the figure the information area is kept constant, while the source type area and source quality area are changing. Examples of message content in terms of the conveyed information, source type, source quality assigned (in $R2$ and $R4$) and information trueness is provided in Table 5.4. For instance, the message M_7 reads that the Company Security Officer, which quality is unknown, reports false information regarding the fact that nothing happened to the ship.

M	Conveyed information	Source type	Source quality assigned	Information trueness
M_1	Current ship position in X	AIS	4	False
M_2	Ship not answering to radio calls	Radio	5	True
M_7	Report that nothing is happening to the ship	CSO	U	False
M_8	Comparison of position X with usual ship routes	POLs Routes	3	True

Table 5.4 Example of Reliability Game messages

In addition to the message cards the player is also presented with flashcards supporting the player's rating and providing additional contextual information. Figure 5.7 shows the flashcard regarding the vessel of interest, the one on the source quality rating scale and finally the one on the confidence rating scale. While the first flashcard contains the relevant information on the ship (e.g., ship type, dimensions, flag state and ownership), the one on the source quality scale represents visually the suggested rating scale for the specific variable, which ranges from 1 to 5 or *Unknown*, with source quality 1 meaning low quality. A source quality scale of six levels has been selected to align with most of the existing standards of

Fig. 5.7 Flashcards – Vessel of interest, source quality levels and confidence levels in the analysis

source reliability rating in the intelligence domain [53]. The use of percentage ranges and the reliability verbal expressions was explicitly avoided, as studies demonstrated the subjective interpretation of the word reliability and of the matching between the verbal and numerical expression [53]. To provide an intuitive visual support to the understanding of the ranking a graphical representation of the scale has been included, which is inspired to the home energy efficiency rating chart [4].

At the end of each round players are asked to rate the global confidence in their analysis of the current situation. The levels relative to the confidence rating are analogous to the ones for the source quality, but provide an additional definition for confidence. The rating scale was selected with reference to the analysis performed in [53]. This review, in fact, shows that with respect to confidence the available rating scales differ in the proposed number of rating levels. Most scales vary from a five level scale to a three level scale. To minimise confusion and errors it was decided to adopt a six level scale in agreement with the source quality scale. Those levels correspond to the five levels present in the intelligence scales plus the Unknown value. This value is deliberately not included in intelligence scales as it is expected that intelligence analysts are able to state their confidence in an analysis[1], however it was deemed interesting its inclusion in order to verify if and how this value would be used if available. It is important to highlight that several standards defining confidence levels map the confidence terms with specific probability intervals. However, as there is no agreement on the correspondence, such mapping was not considered in the Reliability Game.

[1] Private conversation with intelligence analyst

5.4.4 Game design constraints

The main constraints that had to be accounted for during the design phase can be categorised as physical constraints and cognitive constraints. The first ones are those acting on the physical elements of game or related to logistical aspects, while the later are the ones dealing with cognitive tasks to be performed by the player. The main physical constraints are the dimensions of the game elements such as the cards that had to be manageable, readable and had to be moved easily. In addition to this, another important limitation is that in a non-digital game not all item moves can be easily captured unless an external observer constantly records the moves (e.g. through notes or pictures). The main cognitive constraints are the number of cards that have to be provided to the player, the game session length and the need for supporting elements to compensate for the fact that in real world activities operators can rely on background knowledge and on the support of real systems (e.g. a display showing the AIS track of the lost ship). The size of the set of cards has been selected as a trade-off between the ability of the player to manage the set of cards and the attempt to minimise some effects that might impact the experiment results. Two notable effects are the random responding by the players [162] and the carryover effect [26]. The carryover effect takes place within subject experiments when one test might impact the following ones. In order to minimise the carryover effect due to memorising the information from one round to the following one (also referred to as practice effect) it was decided to have a card set size major than seven. In fact, it has been suggested that the storage capacity of the short-term memory of an average person is approximately seven items, plus or minus two [148]. The game session length is a relevant cognitive constraint, as the game had to be short enough to keep the players attention, avoiding mind-wandering effects. Mind-wandering refers to the effect of the mind not focusing on a specific topic for a long period of time, which might occur especially when engaged in attention-demanding tasks [145].

5.5 Game validation

5.5.1 Validation aspects

In the next sections a first analysis of the data is reported. This qualitative analysis has been a useful mean for the Reliability Game validation. More specifically, the game has been evaluated along the four validity criteria identified for K2AGs (Table 4.6). With respect to the psychological and process validity the main aspects have been derived from the structured and unstructured feedback received from the participants. With respect to the structural and predictive validity the analysis started from the concepts of a "fit for purpose" game with

respect to the stated objective. As previously mentioned the purpose of the Reliability Game is to collect data regarding source quality and source type impact on SA and SAW. Therefore, in order to evaluate the effectiveness of the game with respect to the above mentioned scope the main criteria are the observation of variations of card positions and confidence rating between rounds. Because the only input variation between rounds consists in the meta-information about source type and source quality, it is assumed that the two above mentioned criteria are able to capture the corresponding impact on SA as belief change. The following sections report the outcomes of the qualitative analysis of the first sample of data collected.

5.5.2 Experiment set-up

The game underwent a quick prototyping and play testing phase that allowed verifying the board design, the scenario, the information items proposed and the facilitation approach. After minor changes to some information items, a revised version has been issued. The collection of data is still ongoing, but we present herein data collected on a small, but relevant, sample of subject matter experts that allowed verifying the effectiveness of the proposed Gs.

At the time of the evaluation analysis the game had been played with twenty-one (21) players, whose demographics and characteristics are reported in Table 5.5. Participants' selection was performed on a voluntary base from maritime subject matter experts, with either civil or military status. The experimental set-up followed a within subject design, in which the participants have been exposed to four different conditions, namely the game rounds. The conditions variation corresponds to the game view summarised in Table 5.3. For each player the following in-game data has been collected:

($D1$) a picture of the final cards position at the end of each of the four rounds;

($D2$) the source quality rating during the third round;

($D3$) the confidence rating in the hypothesis at the end of each round.

In this experiment there was not an external observer constantly recording the item movements. Thus, only the final aggregation of beliefs at the end of each round has been recorded (D1), while the shuffling of cards has not been captured. However, this represents a minor issue as the cards shuffling resulted in a game mechanics seldom used by the players. Moreover, it will be completely superseded in a digital version of the game currently under development. Beside the in-game data collection above mentioned, a post-game data collection has been performed in the form of feedback questionnaire. The scope of this questionnaire was to assess participants' understanding of the game and perception with

Feature	Specification	Value
Gender	Male	100%
	Female	0%
Age	Average	46.5 years
	Standard Dev.	10.3 years
Status	Military	76%
	Civilian	24%
Nationality	Danish	14.28%
	France	4.76%
	German	19.04%
	Italian	33.33%
	Norwegian	4.76%
	Romanian	4.76%
	United Kingdom	14.28%
	United States	4.76%

Table 5.5 Participants demographics and characteristics

respect of this innovative gaming approach (e.g., relevance with respect to their mission, engagement, facilitation). It is important to note that this questionnaire has been provided as part of a broader feedback questionnaire and only eleven out of the players of the Reliability Game returned their answers.

5.5.3 Feedbacks and observations on the game design

The participant survey shows that the players perceived the game as engaging, realistic and relevant with respect to operational needs (Figure 5.9). From a facilitation point of view, it has been observed that it is important not only to introduce the players to the game rules and to have them familiar with the game dynamics, but also to clearly state and explain the game core to have the players feeling more comfortable and confident about the remaining part of the experiment. Most players actually were explaining their reasoning to the facilitator, which is considered of value for the refinement of next iterations of the game. Players showed to understand well the purpose of the game and the game mechanics, which appears to be intuitive and requiring a low level of pre-experiment training. It has to be underlined that in the Reliability Game there is not a proper pre-experiment training session. Instead, the rules are explained and then the facilitator guides the player when providing the first cards by asking after the card is positioned if the player confirms that the card supports the belief associated to the specific card position. In case of a negative answer the facilitator would help the player positioning the card in the corresponding location.

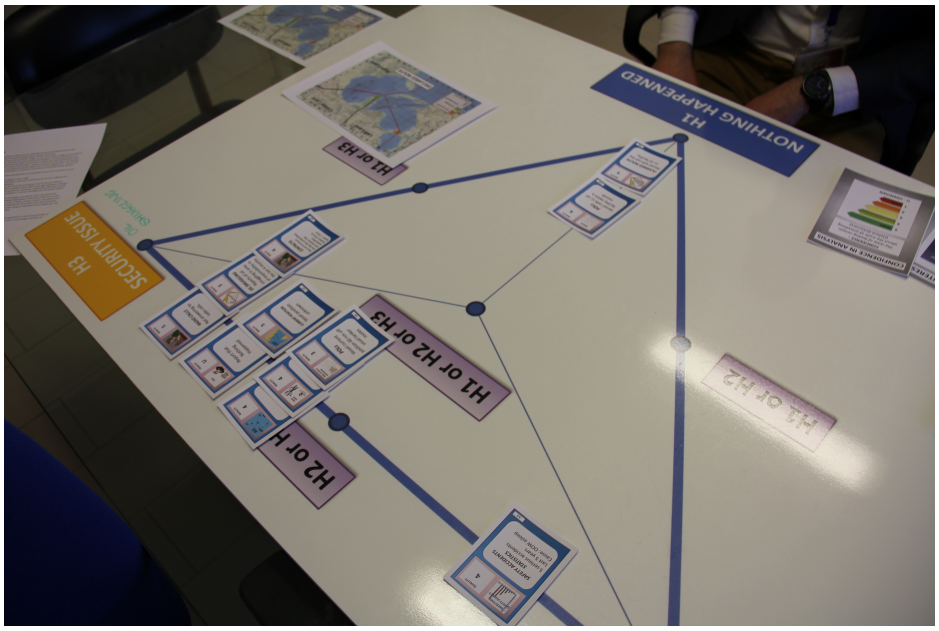


Fig. 5.8 Example of a picture (D1) collected at the end of a round

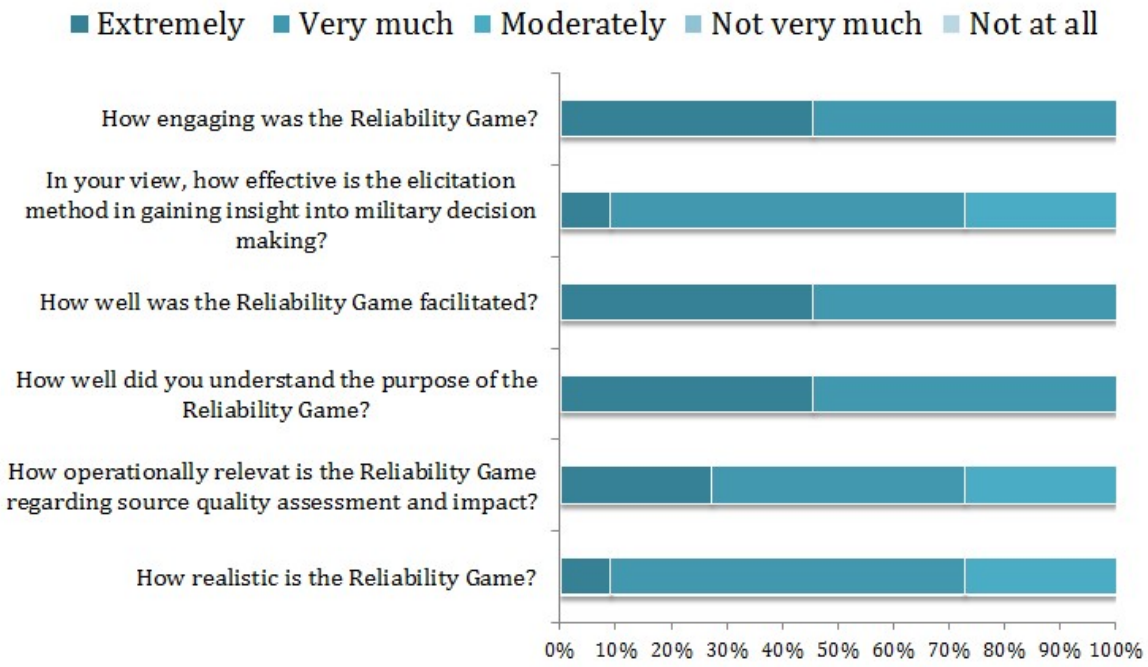


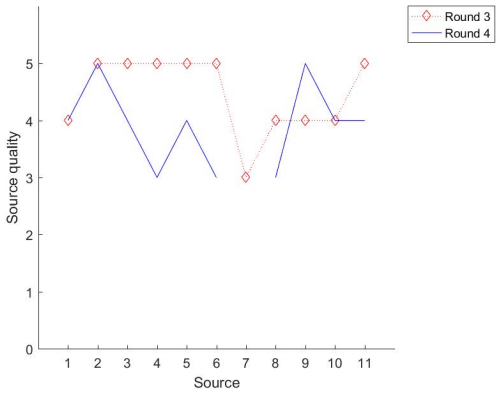
Fig. 5.9 Players' feedback questionnaire outcomes

5.5.4 Outcomes on source quality rating

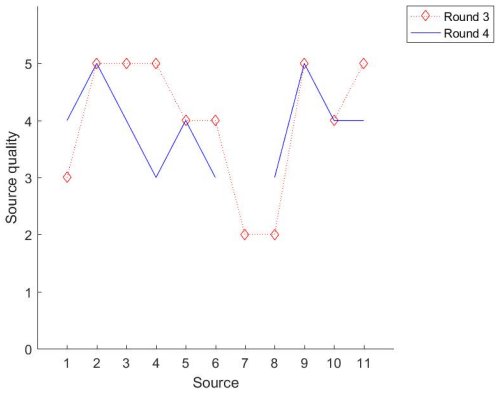
During the third-round of the game participants were requested to rate the source quality given the meta-information on source type which had been provided (Section 5.4.2). Figure 5.10 presents an example of the source quality rating by three different players (red diamonds), which is compared to the source quality values that are provided to the players in $R2$ and $R4$ (blue line). Empty values correspond to an *Unknown* rating. The three players presented different rating profiles. We can observe how Player A (Figure 5.10a) has a tendency to rate the source quality higher than the assigned source quality value. Player B (Figure 5.10b) demonstrates a tendency to variably rate the source quality higher, lower or equal to the assigned ratings. Finally, Player C (Figure 5.10c) shows a tendency to rate the source quality lower than the quality assigned.

Moreover, as shown in Figure 5.10 there was a difference between the participants' source quality ratings and the values provided to them in $R2$ and $R4$. This translates in a variation of the conditions between $R3$ (source type provided, source quality assessed) and $R4$ (source type provided, source quality provided).

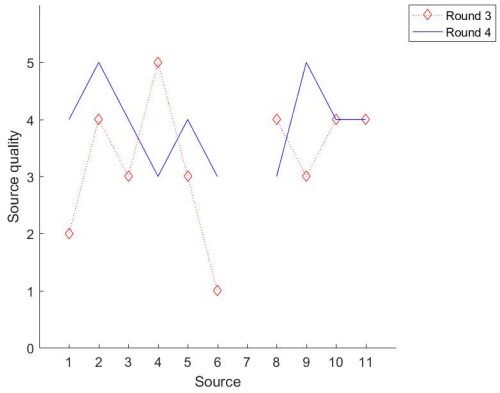
Figure 5.11 depicts the overall source quality assessment for each of the eleven cards. It is important to mention that although the source quality values with decimals (3.5 and 4.5) were not included in the original scale, one player requested to use them. From the figure it can be observed that with the exception of the rating of this player, the assessments on Source 1 and Source 8 are identical. This is an important observation as the degree of familiarity of the subject matter experts with the two sources is considerably different. In fact, Source 1 (Automatic Identification System) is widely available and commonly used in maritime surveillance. On the contrary, Source 8 (Vessel to Route Association algorithm) is more experimental and is still in its early stages of development. Other novel information sources are Source 5 and Source 10, namely a vessel position prediction algorithm and a maritime Patterns-of-Life on ship statistics service. Both sources present a certain degree of variation in the quality rating. However, it can be observed that the one of Source 5 is higher than the one for Source 10. This result suggests that non-conventional information sources are not necessarily considered of low quality. Moreover, from the verbal players' feedback it appeared that the players were drawing comparisons between the source's capacity and their own cognitive abilities (e.g., ability of associating a ship to a route). This observation concurs with some persuasion literature on source factors which has shown the impact of the perceived source similarity on human information assessment [23]. Source 7 (Company Security Officer) is the one exhibiting the highest degree of variability in the quality ratings. This source in $R2$ and $R4$ has an Unknown assigned quality. Only three players rated the source as such, while most of the players assigned a low-quality rating. As explicitly stated



(a) Source quality rating by Player A



(b) Source quality rating by Player B



(c) Source quality rating by Player C

Fig. 5.10 Example of source quality rating ($R3$) by three different players

by some players, this appears not to be related to the nature of the source (human vs sensor), but rather to a possible conflict of interest of the Company Security Officer who could retain or falsify information. This observation is also supported by the fact that Source 2, the human operator, has been rated of high quality.

An interesting result is the one related to the use of the value Unknown. In fact, it has been seldom used, even in the case in which the player had no knowledge of the type of source. More specifically, some players did not know the Company Security Officer or the Long Range Identification and Tracking system. They asked for information to the facilitator, who provided basic information, without disclosing details on the quality. Although the players were often reminded of the possibility of using the *Unknown* value, most of them did not do. This suggests that the players tended to estimate a source quality value even if the source is not known.

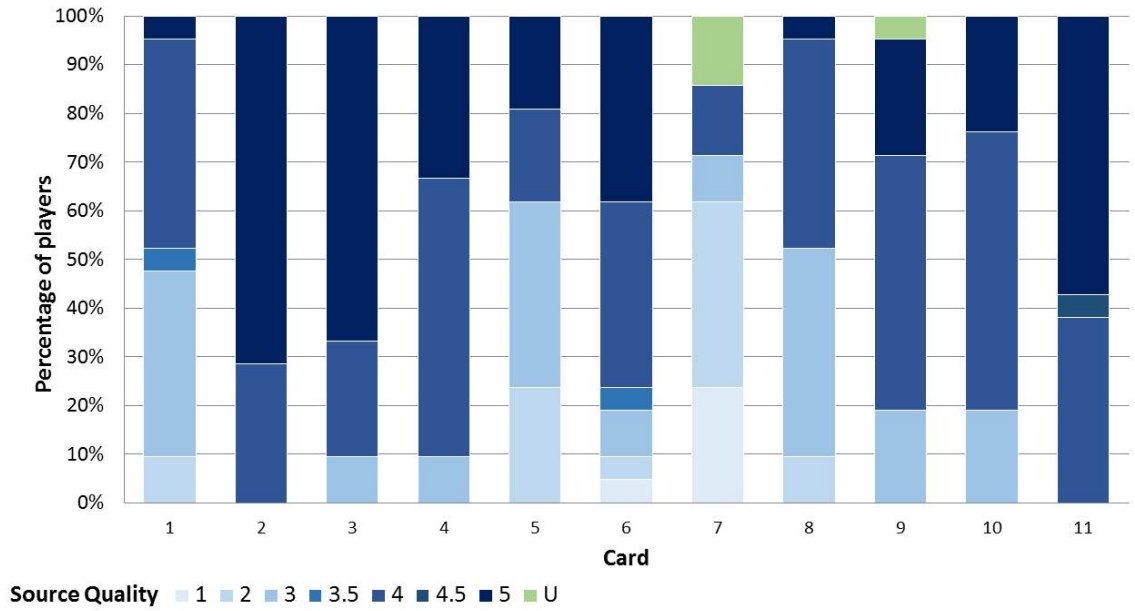
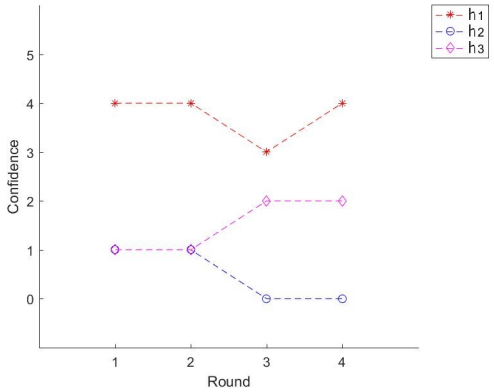


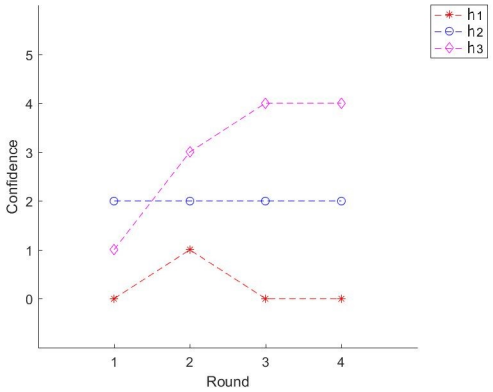
Fig. 5.11 Source quality ratings by card

5.5.5 Outcomes on confidence rating

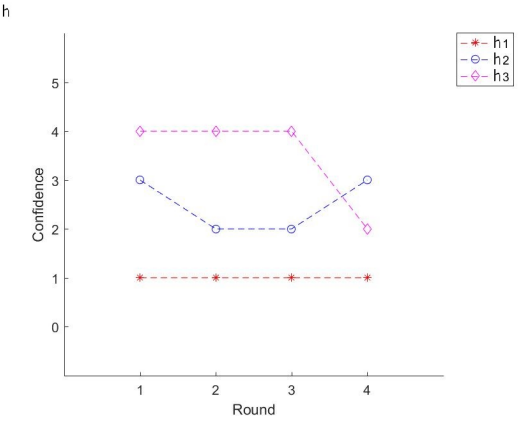
At the end of each round participants were requested to rate their confidence in the fact that the correct hypothesis might be h_1, h_2 or h_3. Figure 5.12 displays the confidence rating of different players. With respect to the relative confidence ratings, it can be observed that the sum of the confidence in the hypotheses is not constant between the rounds and that the variation of the confidence in one of the hypothesis does not imply the variation of the confidence in the others. From Figure 5.13, which is reporting a summary of the different confidence ratings of the participants, we can observe interesting results regarding the use of the scale presented. Equivalently to the case of source quality, one player asked to use a value with decimals. More specifically, the participant asked to introduce the value 2.5 as to express the concept of 50%, which is not possible with the original form of the scale. The proposed scale did not include the rating value 0, which conceptually corresponds to the exclusion of the hypothesis with high confidence. However, many players asked to use the value 0. On the contrary the value 5, corresponding to the conceptual opposite (full confidence that the specific hypothesis is the right one), has been rarely used. This result suggests that participants might more easily exclude hypotheses than being certain about them. Another possible interpretation is that they might feel more self-confident in excluding than being certain about the hypotheses. With the term self-confidence the author refers to the concept of self-assurance in personal judgment. Contrary to the intelligence scales a sixth



(a) Confidence rating by Player A



(b) Confidence rating by Player B



(c) Confidence rating by Player C

Fig. 5.12 Example of confidence rating by three different players

level was included in the original scale, namely *Unknown*, to allow the players the possibility to state their inability to draw a conclusion and express their confidence. This value is on purpose excluded from the intelligence standard scales as it forces the analyst to exactly rate his confidence, no matter if high or low, without using the above mentioned value as a solution to avoid liability issues. It is, however, interesting to notice that this value has been used twice by the participants, but this is not reported in the graph as the players soon after asked if they could re-rate the confidence. Another interesting observation regarding the confidence rating is that some players when requested to express their confidence were stating that it was unchanged with respect to the previous round. The facilitator, however, requested the players to explicitly rate the current confidence levels. The resulting rating was in general not equal to the previous one, suggesting that there had been a change of which the players were not conscious. This also suggests that the mechanisms to try to minimise the carryover effect were effective on the confidence rating.

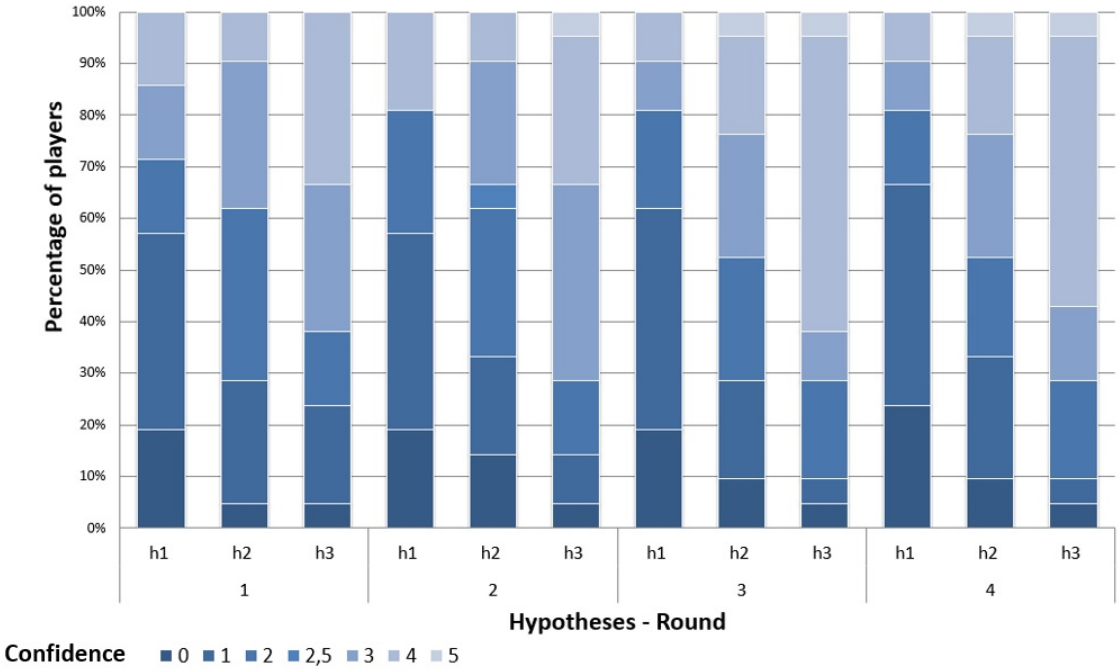


Fig. 5.13 Confidence ratings by hypothesis in the different rounds

5.5.6 Outcomes on card positions

At the end of each round a picture of the board has been taken. From the pictures the author has been able to record all the data regarding the single card assessment. Table 5.6 reports an example of the card positions in the different rounds played by one participant. It is important to underline that although this example shows only cards positioned in the points specified on the board (corners, mid of axes and center of the triangle) the players in general used the full spectrum of the possible positions, more specifically the axes displayed on the board (Figure 5.4). Table 5.6 reports only the final position for each round, while the card shuffling is not reported. This is because although the players have been allowed to shuffle cards during the game (G6), this G has been used only twice during the experiment run.

Table 5.6 shows how position variations between the rounds have been consistently observed. In fact, if we transform the card positions in beliefs expressed by numerical probabilities (see Section 5.7.4 for details on the transformation) we can easily see the impact of meta-information on the beliefs. An exemplification is displayed in the charts in Figure 5.14. Each one shows a player's belief change related to the presentations of a card (i.e. E, F, G and H). As previously explained, in fact, each card of the game has a different presentation in the different rounds. If we observe Figure 5.14a we can notice how the player's belief in R1 ($P(h_1) = 100\%$) did not change in R2, when only source quality is provided. On

	Position			
Card	Round 1 ($R1$)	Round 2 ($R2$)	Round 3 ($R3$)	Round ($R4$)
1	h_1	h_1	h_3	h_3
2	h_1 or h_3	h_1	h_1	h_3
3	h_2	h_1 or h_2	h_1 or h_2	h_1 or h_2
4	h_3	h_1 or h_3	h_3	h_1 or h_3
5	h_2	h_3	h_1	h_1
6	h_2	h_1 or h_3	h_1 or h_3	h_1 or h_3
7	h_1 or h_2 or h_3	h_1 or h_2 or h_3	h_1	h_1 or h_2 or h_3
8	h_3	h_1 or h_3	h_1 or h_3	h_1
9	h_3	h_1 or h_2 or h_3	h_1 or h_2 or h_3	h_3
10	h_3	h_3	h_3	h_3
11	h_3	h_3	h_3	h_3

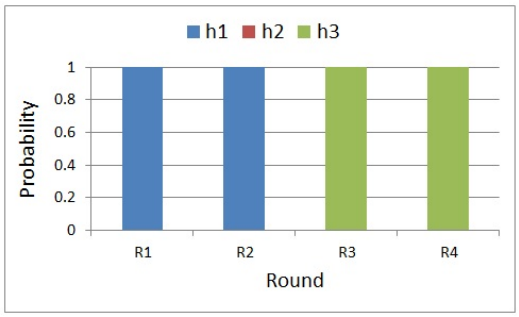
Table 5.6 Example of card positions collected for each player

the contrary, we can observe how it substantially changed in R3 ($P(h_3) = 100\%$), when meta-information on source type is provided. Figure 5.14b to Figure 5.14d similarly show belief variations associated to other three cards and their presentations (information plus potential meta-information). The four charts have been selected as they well depict how much a belief can change just as the result of meta-information. For example, it could pass from $P(h_3) = 100\%$ to $P(h_1) = 100\%$ (Figure 5.14c) or vice-versa. In general, these extreme changes are not observed for all the cards processed by all the players. In fact, in some cases the change in belief is weaker.

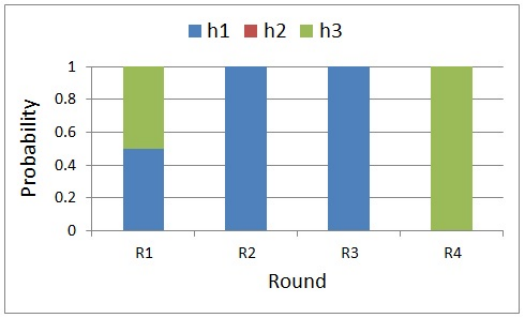
From a qualitative analysis it has been possible to observe the impact of source quality and source type on players' assessment by means of the change of the card positions on the board between the different rounds played by the same participant. However, a more in depth analysis is required to be able to quantify this impact and draw connections between those factors, the player SA and final confidence. Such an analysis requires a formalisation of belief assessment together with a proper encoding of the players' cards positions.

5.6 Discussion on validation

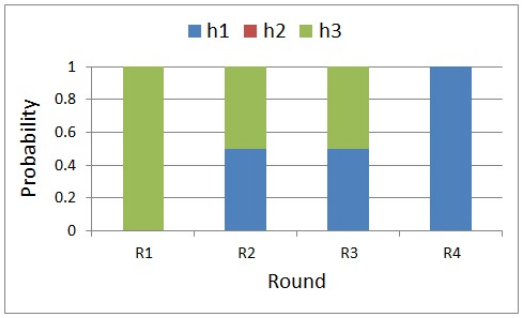
The purpose of the Reliability Game is to collect data regarding players' belief changes as a function of source factors, more specifically source type and quality. To gather such data each player is presented with a scenario and plays several rounds of the game. The only variation between rounds consists in his knowledge regarding source type and quality. The corresponding belief changes are captured though the variation of game items position (cards) and final confidence ratings. A qualitative analysis was performed on the data gathered



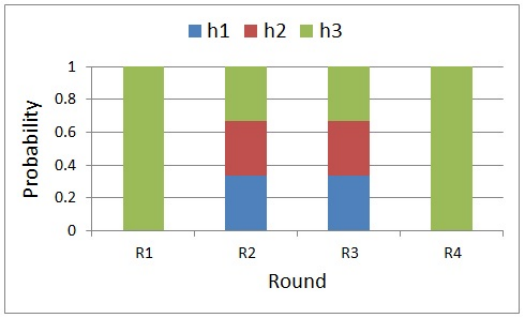
(a) Variation of beliefs between rounds for cards E



(b) Variation of beliefs between rounds for cards F



(c) Variation of beliefs between rounds for cards G



(d) Variation of beliefs between rounds for cards H

Fig. 5.14 Examples of belief variations due to card presentation.

through an experiment run with a non-digital version of the Reliability Game in order to evaluate the effectiveness of the game design and game mechanics. The variations of the players' belief assessments between the different rounds demonstrate that the proposed methodology effectively captures elements of source factors impact on SA. Moreover, the analysis allows assessing important aspects of the use of the rating scales, which might be relevant to the standardisation efforts in communication of uncertainty (e.g. confidence, reliability).

In addition to the collection of in-game data, a post-game data collection has been performed in the form of a feedback questionnaire. Although a specific player experience questionnaire was no used, the results collected show that the game is perceived both as engaging and relevant. Moreover, the game scope and game mechanics were easily understood.

To summarise, the analysis performed presents positive results for the validation activity along the four dimensions of psychological, process, structural and predictive validity. With respect to the predictive validity, moreover, we did not only perform the above mentioned qualitative analysis, that highlighted the ability to detect change and isolate the reason for such

change, but also a quantitative analysis (Section 5.7) that shows how the knowledge collected is relevant to the stated KE problem and how it can be used to train fusion algorithms.

5.7 The Reliability Game KEBN

5.7.1 Background

The next sections explain how the game results have been modelled with a BN and how learning techniques have been applied to extract patterns of reasoning from the available data. More specifically, the focus has been on understanding how the players accounted for the source factors and, therefore, implicitly understood source reliability. It is important to underline that the model described hereafter has not to be considered as an attempt to model the real cognitive process of the player, but rather as the development of a computational model able to provide results comparable to the one of the players [216]. An intuitive explanation of the stated problem is that we are trying to understand how humans fuse information provided by different sources and how much they discount (or ultimately discard) them based on meta-information on the source itself. The discounting factor, could be interpreted as the source reliability, which we are trying to estimate.

Feature	Specification	Value
Gender	Male	94%
	Female	6%
Age	Average	43.5 years
	Standard Dev.	11.3 years
Status	Military	50%
	Civilian	50%
Nationality	Canadian	3.1%
	Danish	9.4%
	France	3.1%
	German	15.6%
	Italian	46.9%
	Norwegian	3.1%
	Romanian	3.1%
	Turkish	3.1%
	United Kingdom	9.4%
	United States	3.1%

Table 5.7 Participants demographics and characteristics

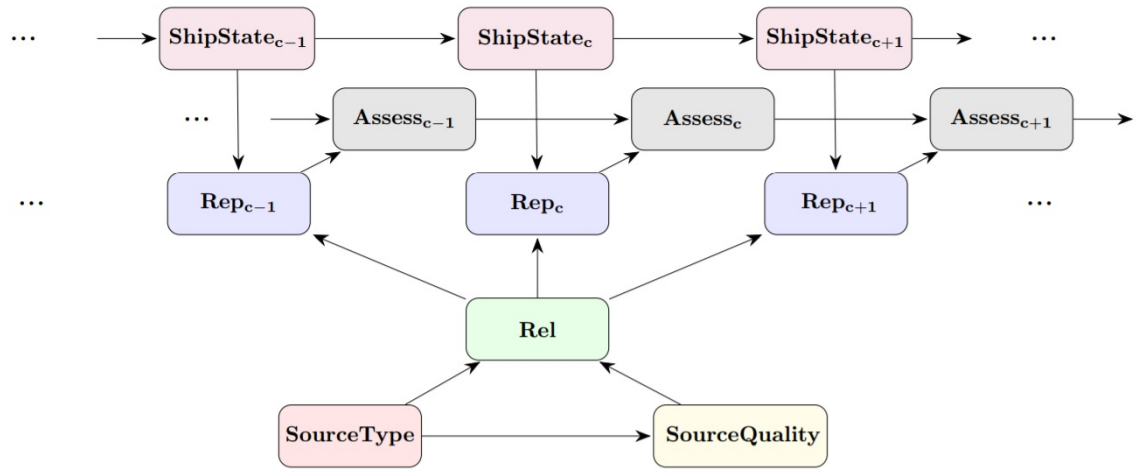


Fig. 5.15 The Reliability Game Bayesian network

The quantitative analysis described hereafter has been performed on a wider sample than the previous analysis. It comprised, in fact, data collected from thirty-two (32) players. Their characteristics and demographics are reported in Table 5.7.

5.7.2 Relevant Reliability Game BN variables

When modelling the Reliability Game BN (RGBN) (Figure 5.15) we attempt to create a computational model that obtains results equivalent to the ones outputted by the player in order to be able to deduce the way in which the participant accounted for source factors.

The relevant variables (or BN nodes) for this model are clustered into the following sets of variables:

ShipState: the set of variables $ShipSt_c$ with $c = 1 \ldots N$; N represents the number of total cards provided during each round (i.e. $N = 11$); each $ShipSt_c$ node is characterised by three states, h_s with $s = 1,2,3$ to indicate the three exhaustive and mutually exclusive hypotheses presented in the Reliability Game;

Assessment: the set of variables $Assess_c$ with $c = 1 \ldots N$, encoding the players beliefs on the three different hypotheses after a piece of information and possible meta-information is provided through cards; those variables present the same states as $ShipSt_c$, as it represent the players assessment of such variable;

Reporting: the set of variables corresponding to the different pieces of information on the situation to be assessed; if N represents the number of total cards provided during

Set of variables	Variable	Universe of disclosure
ShipState	$ShipSt_c$ with $c = 1,\ldots,11$	$\{h_1,h_2,h_3\}$
Assessment	$Assess_c$ with $c = 1,\ldots,11$	$\{h_1,h_2,h_3\}$
Reporting	Rep_c with $c = 1,\ldots,11$	$\{True,False\}$
Reliability	*Rel*	$\{True,False\}$
SourceFactors	*SourceType*	{*AIS*, *LRIT*, *Posit.Pred.*, *CSO*, *Rep.Proced.*, *Radio*, *Intelligence*, *SafetyAgency*, *PoLsCalls*, *PoLsRoutes*, *VTS*}
	SourceQuality	{1,2,3,4,5}

Table 5.8 Summary of the RGBN nodes and their states

each round, the reporting variables will be Rep_c where $c = 1\ldots N$; those are Boolean variables;

Rel: the *Rel* is a Boolean latent variable, because within the Reliability Game it is never mentioned explicitly to the player and no data on it is directly collected during the experiments;

SourceFactors: the set of variables including *SourceType* and *SourceQuality*; those are the two source factors that were selected to experiment with in a first instance; *SourceType* is a variable characterised by as many states as the different source types taken into consideration in the Reliability Game (Table 5.1); *SourceQuality* is a variable which states correspond to the different levels of the source quality rating scale proposed in the Reliability Game (i.e. one to five); it has to be noticed that the rating scale presents also the possibility to rate the source quality as *unknown*, but this value is not included as a possible variable state as it would translate into a uniform distribution on all the *SourceQuality* states.

The full universes of disclosure of the different variables of the RGBN can be seen in Table 5.8.

5.7.3 The RGBN structure

The next step has been to determine the ordering and dependencies of the RGBN variables.

To model the problem we started from the idea of an Hidden Markov Model (HMM) in which the hidden variable would be the state in which the ship is, that corresponds to one of

possible hypotheses proposed to the player. The ship state ($ShipSt_c$) is never observed by the player and is only known to the game designer, that selects one when the scenario and cards of the game are designed. What the player, instead, observes are the reports provided through the cards (Rep_c). To account for possible partial reliability of the sources of information proving the reports we need to introduce a *reliability structure* such as the one defined in Section 5.3. An important difference compared to this one is that the *SourceFactors* variables included in the RGBN, namely *SourceQuality* and *SourceType*, are not conditional independent. In fact, knowing the source type allows to estimate a generic source quality. This dependency has also been observed in the first analysis of the data collected through the experiment [55].

Finally, each $Assess_c$ variable depends not only on the current report (Rep_c), but also on $Assess_{c-1}$. This dependency has been introduced to account for the fact that, although it is requested to assess each piece of information separately, it is reasonable to assume that the new assessment is not independent from the previous one. In fact, an independence assumption might be too strong in this context as anchor effects might be observed between subsequent assessments.

5.7.4 Learning the RGBN reliability CPT

To learn the CPT of the *Rel*, which is the main scope of this modelling effort, the Netica software [160] has been employed as it is capable of handling machine learning with latent variables and uncertain evidence. This is the case for the data collected through the Reliability Game. More specifically, the players beliefs recorded as positions on the triangle of game board had to be translated in evidences on $Assess_c$. For example, positioning a card in the lower corner of the triangle indicates that the specific piece of information provided by that card is pointing towards the hypothesis h_1 only, therefore, the associated probabilities would be:

$$p(Assess = h_1 | Rep_c = True) = 1$$
$$p(Assess = h_2 | Rep_c = True) = 0$$
$$p(Assess = h_3 | Rep_c = True) = 0$$

Positioning the card in the middle of the axis between h_1 and h_2 would be translated into the following probabilities:

$$p(Assess = h_1|Rep_c = True) = 0.5$$
$$p(Assess = h_2|Rep_c = True) = 0.5$$
$$p(Assess = h_3|Rep_c = True) = 0$$

From the above mentioned examples, we can easily observe how evidence in this case should be regarded as *uncertain* and more specifically as *likelihood* evidence.

In fact, it is important to distinguish the different types of evidence, as well as their interpretation. More in details, evidence can be *certain* (or *hard*) or *uncertain*. Confusion exists in the literature regarding the different kinds of *uncertain* evidence. Therefore, we will refer to the terminology proposed in [153], that distinguishes between *likelihood* evidence (or *virtual evidence*), *fixed probabilistic evidence* (or *soft* evidence) and *not-fixed probabilistic* evidence. The *likelihood* evidence is an evidence specified through a likelihood ratio and is interpreted as evidence *with* uncertainty. Evidence in which the uncertainty is generated by the *unreliability* of an information source is a prototypical example of *likelihood evidence*. On the contrary, *probabilistic evidence* refers to *uncertain evidence* which is specified through a probability distribution and defines a constraint on specific variables after the propagation within the network. The two terms *fixed* and *not-fixed* are introduced to highlight two possible mechanisms of update of the posterior probability distributions when subsequent evidence is provided. For further details the reader is referred to [153].

An Expectation-Maximisation (EM) algorithm has been employed to perform the parameter learning. BN parameter learning with latent variables and small sample sizes can be a difficult task. Therefore, in the example proposed in this work to show the potentiality of K2AGs, we introduced some prior knowledge in the BN, by setting the initial parameters of the variables based on domain knowledge elicited from experts [120].

Finally, a K-fold stratified cross-validation approach has been adopted in order to determine the performances of the learning exercise. This re-sampling procedure can be used to evaluate machine learning models when a limited data sample is available. The approach consists in generating K non-overlapping folds from the original data sample, which should be previously shuffled. On turn one fold is assigned to the test data-set, while the remaining ones will constitute the training set. The different models generated can then be evaluated and the most accurate can be retained. For additional details the reader is referred to [99]. The value of K can vary depending on the issue at hand, but often a $K = 10$ is used. In this case a stratified approach has been adopted, which consisted in selecting a K value equal to the number of players and each fold contains the observations from a player. This choice

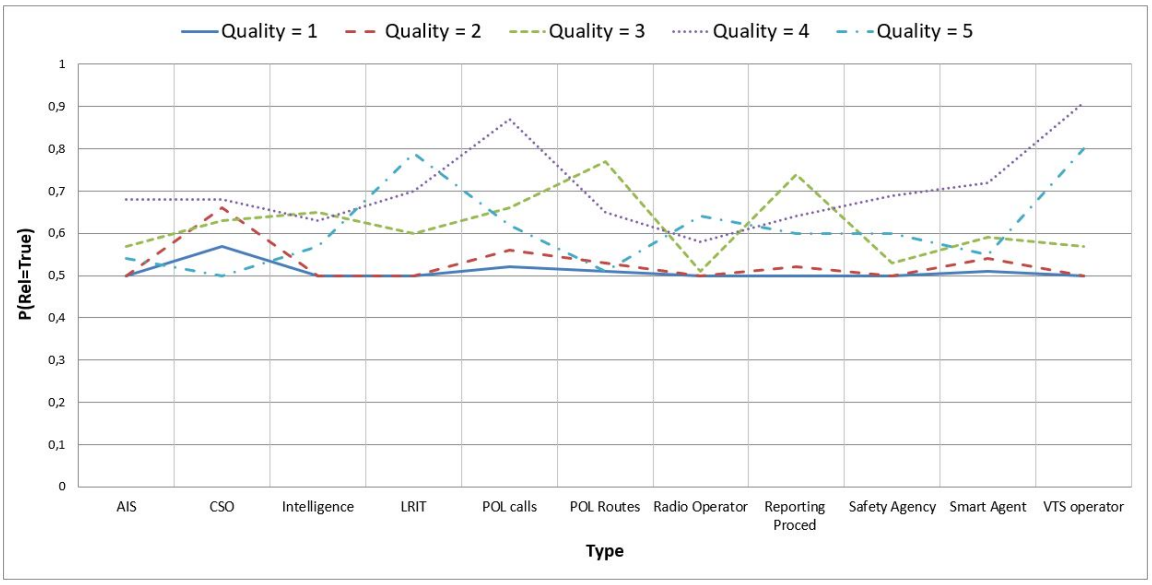


Fig. 5.16 Example of results for the *Rel* CPT learning

is motivated by the fact that in the RGBN each $Assess_m$ is dependent on $Assess_{m-1}$, with $m = 2, \ldots, 11$.

5.8 Results and discussion

The performances observed for the proposed algorithm in terms of time have shown good results, however, the resulting accuracy of the models is not particularly high. In fact, in the case of the best model obtained we observe 122 EM learning iterations and log loss values for the $Assess_i$ variables between 0.14 and 1.2. Despite the accuracy results, we can still see that the machine learning exercise yields promising results.

Figure 5.16 and Table 5.9 show an example of the results obtained for the *Rel* CPT in one of the computational models obtained. From this example, we can observe how the outcomes are overall coherent. Let us consider three probabilities relative to different sources with same quality excerpted from Table 5.9:

$$p(Rel = True|Type = LRIT, Quality = 5) = 0.79$$
$$p(Rel = True|Type = \text{POL Calls}, Quality = 5) = 0.62$$
$$p(Rel = True|Type = \text{VTS}, Quality = 5) = 0.80$$

	Quality = 1		Quality = 2		Quality = 3		Quality = 4		Quality = 5	
Type	Rel = True	Rel = False	Rel = True	Rel = False	Rel = True	Rel = False	Rel = True	Rel = False	Rel = True	Rel = False
AIS	0.5	0.5	0.5	0.2	0.57	0.43	0.68	0.32	0.54	0.46
CSO	0.57	0.43	0.66	0.34	0.63	0.37	0.68	0.32	0.5	0.5
Intelligence	0.5	0.5	0.5	0.5	0.65	0.35	0.63	0.37	0.57	0.43
LRIT	0.5	0.5	0.5	0.5	0.6	0.41	0.7	0.3	0.79	0.21
PoLs Calls	0.52	0.48	0.56	0.44	0.66	0.34	0.87	0.13	0.62	0.38
PoLs Routes	0.51	0.49	0.53	0.47	0.77	0.23	0.65	0.35	0.51	0.49
Radio	0.5	0.5	0.5	0.5	0.51	0.49	0.58	0.42	0.64	0.37
Rep. Proced.	0.5	0.5	0.52	0.48	0.74	0.26	0.64	0.36	0.6	0.4
Safety Agency	0.5	0.5	0.5	0.5	0.53	0.47	0.69	0.31	0.6	0.4
Posit. Pred.	0.51	0.49	0.54	0.46	0.59	0.41	0.72	0.28	0.55	0.45
VTS	0.5	0.5	0.5	0.5	0.57	0.43	0.91	0.09	0.8	0.2

Table 5.9 Results example for $p(Rel = True|Type = t, Quality = q)$ learning

Those values are aligned with the intuitive understanding of reliability. In fact, VTS operators and LRIT are generally considered more reliable than algorithms, such as the ones providing POLs, which are considered still in early development phases. Moreover, from Figure 5.16 we can see how for most source at low values of $Quality$ the probability of the source being reliable is around fifty percent (50%), corresponding to a full uncertainty. In the figure the probability of being reliable appears to be bounded between fifty (50%) and ninety (90%) percent. As can be observed in Figure 5.17, the same occurs for all the folds results. This is interesting as it corresponds to the fact that some sources of high quality are considered highly reliable, but never fully reliable, which would correspond to hundred percent (100%) probability. The degree of reliability is confined between a high value of certainty and the complete uncertainty. On the contrary, we are not observing values below fifty percent (50%), that would correspond to the negative part of the continuum between being certain of the reliability of a source (i.e. 100%) and being certain of the unreliability of a source (i.e. 0%).

It can also be observed how for some sources the increase in quality corresponds to a slight increase in the probability of being reliable (e.g. the smart agent), while for others (e.g. the VTS operator) there is a substantial increase. This is in line with the understanding that quality is not the only source factor that impacts source reliability assessment. In fact, there are several other factors, such as *vulnerability to manipulation* or *history of past use* that play an important role. Those variables can be considered all aggregated here in the variable $Type$. With respect to the above mentioned sources this corresponds to some verbal comments

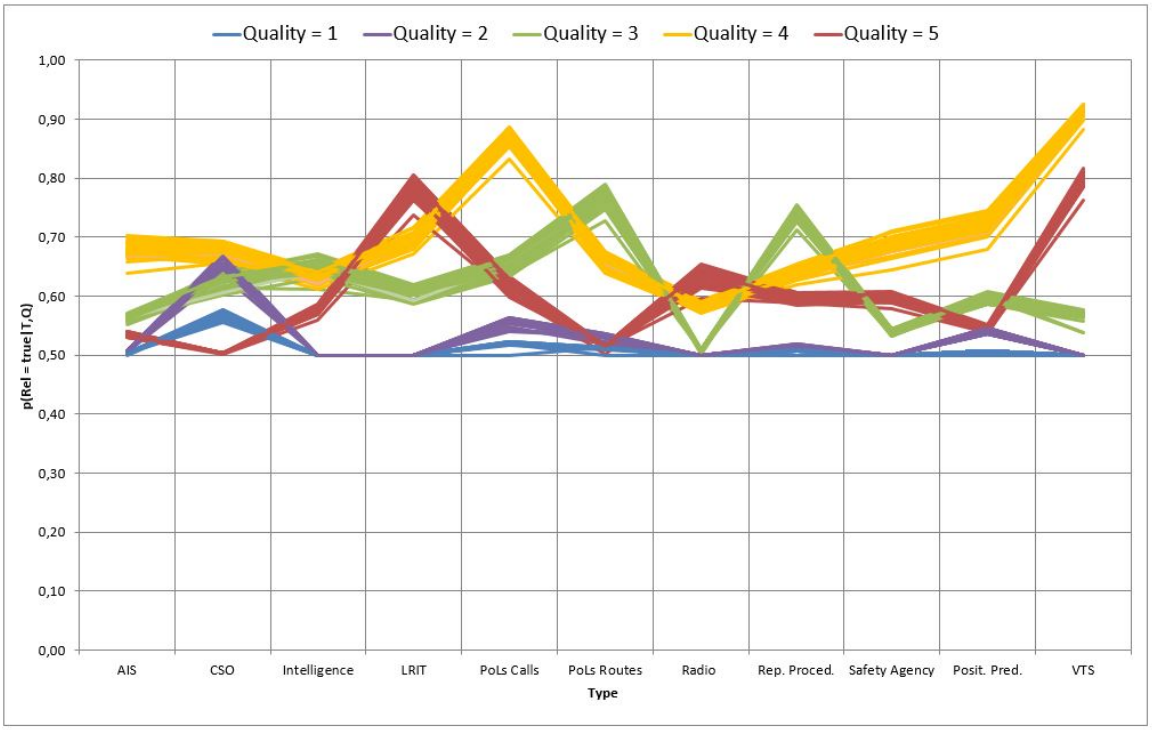


Fig. 5.17 Results for the *Rel* CPT learning

provided by the players that asserted to trust less new algorithms (or smart agents) as these are tools they are not used to work with. However, counter-intuitively for most sources we obtained:

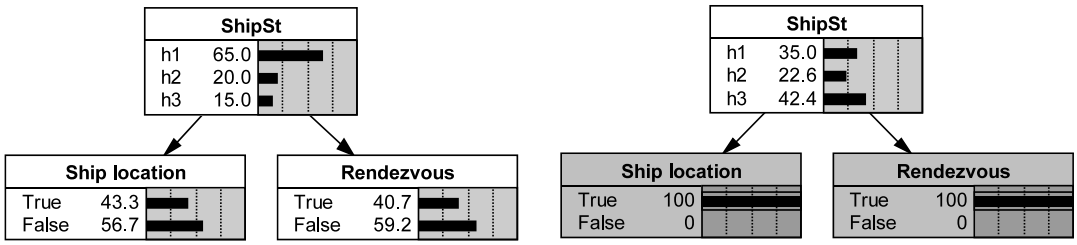
$$p(Rel = True|Type = t, Quality = 5) \neq$$
$$\max p(Rel = True|Type = t, Quality = q)$$

The explanation is possibly connected to three aspects, namely the small sample size, a non-adequate conditioning of the problem and the possible impact of source factors other than the ones included in the experiment. The first aspect relates not only to the issue of the availability of expert to play the game, but also, in the case of analog games, to the necessity of physically reach those experts and the need for an expert facilitator. To overcome this problem digital games appear to be a suitable solution. The second aspect, instead, is related to the fact that at the time of experiment design the use of the collected data for KEBN was not foreseen. Therefore, the data collection and analysis plan [27] was not tailored on such activity. This will be overcome in future studies on the topic, through a more effective design of the experiment. Further experiments will be conducted in order to understand in which

proportion the model accuracy has been affected by the different aspects mentioned above and to investigate other source factors. Nevertheless, the results reported indicate that beside being an engaging and easy elicitation method from the expert perspective, K2AG should be regarded as useful KA method, able to collect relevant data to be used in the design of algorithms.

5.9 Example network

In this section we show how the results of the learning of the reliability latent variable could be used in practice. To this end, we can refer to a simple example in which we want to estimate for a vessel of interest the *ShipSt* with $\Omega = \{h_1, h_2, h_3\}$ and we have two different reports related to the situation. More specifically, one report is on the location of the ship, which is reported by the AIS system and one refers to a rendezvous involving the vessel of interest. This second report is provided by a VTS operator. Both the reports tend to induce a belief change towards the fact that something is happening to the ship (i.e. h_2 or h_3). Figure 5.18a implements a simple BN to reason about this problem, in the more traditional approach, which does not account for the information source factors. In this BN the evidence provided by the reports is entered directly as hard evidence on the variables corresponding to the ship location and the ship rendezvous (Figure 5.18b). Figure 5.19, on the contrary, shows the implementation of the reasoning on the same issue, accounting for source factors. In this case evidence is not directly entered on the two situation variables (i.e. *Ship location* and *Rendezvous*), but on the reporting variables (Figure 5.20). Moreover, we can further enter evidence on the source. Like in many real life cases, in this example we do not know the source quality or source reliability, but we know the source type. Therefore, the evidence is entered on the *SourceType* variable. From Figure 5.18b and Figure 5.20 it is possible to clearly see how the result of the reasoning varies if the source is considered or not. In fact, in the latter case, we see how the probability that nothing is happening, once evidence is provided, varies from $p(ShipSt = h_1) = 0.65$ to $p(ShipSt = h_1) = 0.35$. Differently, in the case in which the source reliability is considered we see a variation from $p(ShipSt = h_1) = 0.65$ to $p(ShipSt = h_1) = 0.54$. This is because the sources are not fully reliable (i.e. $p(Rel = True|Type = AIS) = 0.57$ and $p(Rel = True|Type = VTS) = 0.87$). Therefore, we observe a mitigation effect on the belief change that the evidence can induce.



(a) Example network.

(b) Example network with evidence provided by the reports entered on the variables.

Fig. 5.18 Example BN implementing the traditional reasoning without source reliability accounting.

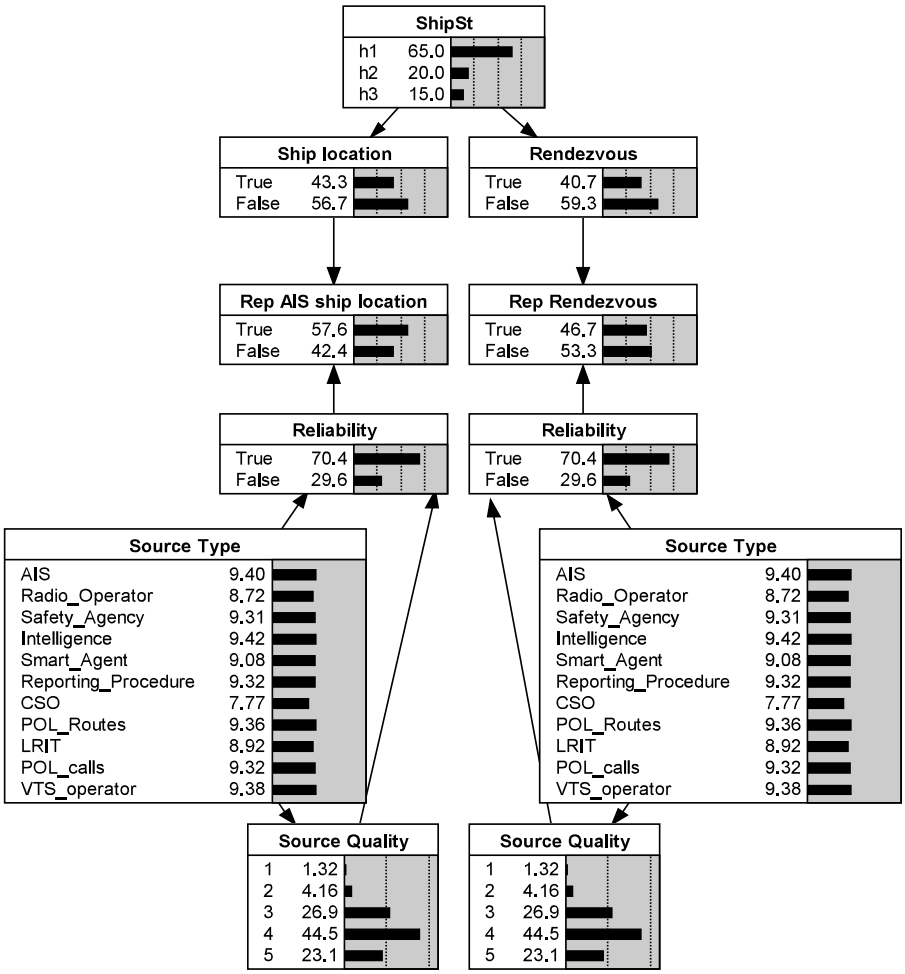


Fig. 5.19 Example BN implementing the reasoning with source reliability accounting, without evidence provided.

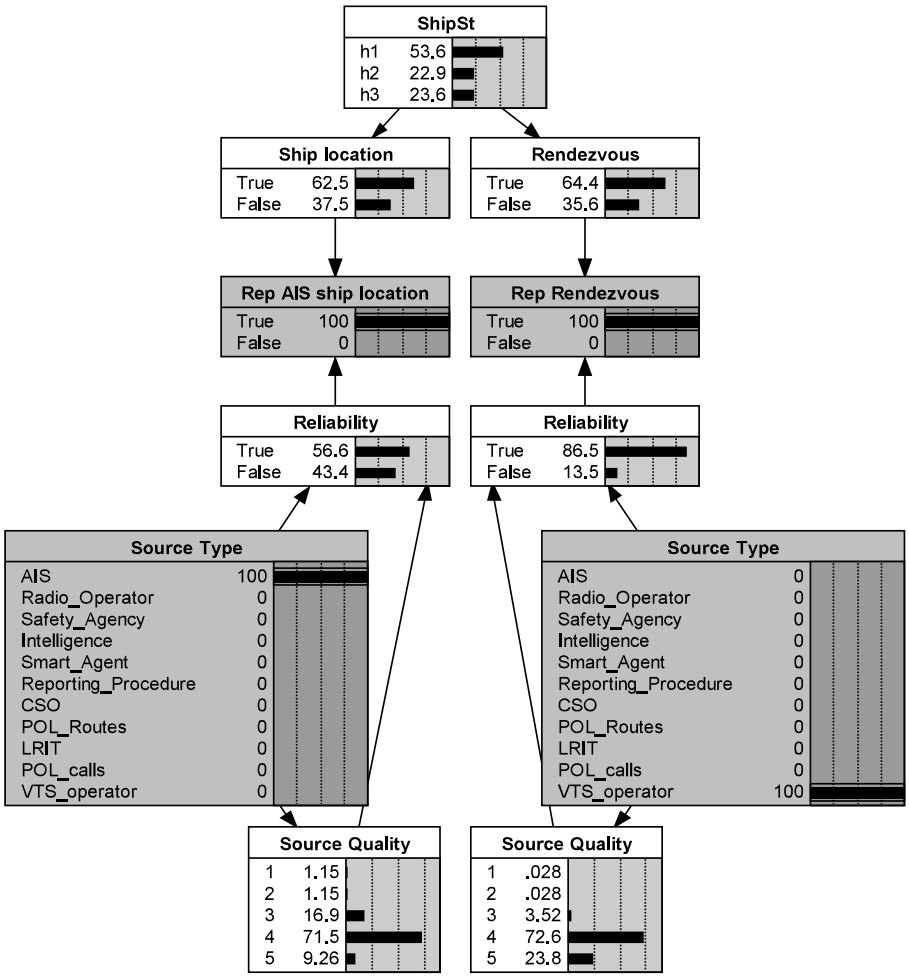


Fig. 5.20 Example BN implementing the reasoning with source reliability accounting and evidence provided.

Chapter 6

The Reliability Game and other Human Factors methods

6.1 Motivation

An integral part of the intelligent system design process is the testing and evaluation (T&E) phase. The formal assessment of the mental state of SAW is a complex task. It appears that SAW assessment in testing and evaluation is often either overlooked by adopting a technology-focused approach or only partially addressed through the use of specific human factors methods. Testing and evaluation of intelligent systems in general and maritime surveillance systems specifically should account both for the system components enabling Situational Awareness and the human element [54]. Reconciling the technology and human factors perspectives for T&E of systems in support of SAW, it is suggested that the following elements should be analysed in a holistic approach: (i) operational picture quality evaluation; (ii) interphase evaluation; (iii) SA evaluation; (iv) workload evaluation and (v) SAW evaluation.

Although all those elements are correlated, they still need to be tested independently in order to evaluate if a system is enabling an adequate (or enhanced) SAW level. In fact, if SAW is an end-state, then its assessment alone is not sufficient to inform T&E. Therefore, it is important to explore all the building blocks that lead to that state.

The SA evaluation is an element that has received attention only very recently. In fact, researchers have started mentioning the importance to explore the SA processes (e.g. [76]) and highlighted how different persons might reach the same level of SAW, but trough different reasoning paths. Exploring the process that leads to SAW could give important cues to understand which are the system elements that might have caused a low quality of SAW.

Therefore, researchers have started looking at methods to explore SA (e.g. eye tracking, verbal protocols [218], scenario manipulation [194]). K2AGs explore the SA process and consequent SAW, therefore, could be regarded as potential tools to complement the other methods used in T&E. Therefore, in this chapter we lay the foundation for a future in depth analysis on the use of such games for the evaluation of SA and SAW for T&E.

To this end after a brief survey of some human factors (HF) methods adopted in the context of Situational Awareness assessment in Section 6.2 and the presentation of relevant results of psychology and social science research on source factors in Section 6.3, the Reliability Game method is outlined as a human factors method in Section 6.4. Section 6.5 summarises a qualitative comparison with other HF methods and highlights novelties and similarities. Finally, some discussions and way ahead are presented in Section 6.6.

6.2 Situational Awareness assessment methods

A literary review by Stanton et al. [212] highlighted the existence of several human factors methods dedicated to the assessment of SAW. Most techniques concentrate on the assessment of individual SAW through measurement approaches for example by looking at physiological aspects, performance aspects, embedded tasks, subjective ratings and questionnaires [212, 75]. Less emphasis has been put on distributed or team SAW techniques [212]. Following [212] the individual SAW assessment techniques can be categorised as:

1. SAW requirements analysis techniques;
2. freeze probe technique;
3. real-time probe technique;
4. self-rating techniques;
5. observer-rating techniques.

SAW requirement analysis techniques, which might be based on interviews with Subject Matter Experts, questionnaires and goal-directed task analysis [73], aim at understanding which are the elements that contribute to SAW with respect to a specific task or environment. In freeze probe techniques (e.g. SACRI [104], SAGAT [75] and SALSA [101]) a task and/or scenario is simulated and participants have to respond to SAW related queries administered during a freeze of the simulation. Real-time probe techniques (e.g. SASHA [110] and SPAM [71]), differently from the previous ones, administer the SAW queries without freezing the simulation, while in the self-rating techniques (e.g. CARS [144], MARS [143], SARS [239],

SART [220] and C-SAS [62]) the participants are requested, generally post-trial, to self-rate dimensions related to SAW. For example, in SART the dimensions include familiarity, complexity of situation, information quality, information quantity and concentration of attention. In observer-rating techniques (e.g. SABARS [143]), contrary to the ones previously mentioned, there is an appropriate subject matter expert who rates the participants SAW, while observing them performing a specific task.

The aim of the above-mentioned methods (with the exception of SAW requirements techniques) is to measure the level of SAW, often with the primary scope of assessing specific operational systems and/or innovative technologies and designs. Besides serving as key performance indicators of the effectiveness of novel technologies, those techniques and measurements allow also to investigate [75]:

1. the nature of SAW;
2. factors affecting SAW;
3. the strategies and processes adopted to acquire SAW.

The last two points are the objectives of the Reliability Game method, which is going to be detailed in the following sections. In fact, the Reliability Game method aims at characterising the impact of factors related to sources of information (e.g. source type and source quality) on the Situational Assessment process and final SAW.

6.3 Source factors impact on human assessment

Aspects related to the impact of source factors on human assessments have been the subject of social science and the experimental psychology for decades. Although the results of those studies cannot be directly incorporated within the modelling paradigms used in the context of information fusion, they served as basis for the interpretation and analysis of the data gathered through the Reliability Game method.

Persuasion literature reports on the mechanisms that determine the effectiveness of sources of information perceived as credible, attractive, similar or powerful [23, 116, 186, 214]. Research has shown the complexity and dynamic nature of the processes taking place with respect to source factor impact on attitude change: there is not a linear mapping between a message provided by a more attractive or expert source and a higher degree of persuasion (or attitude change in the expected direction). Therefore, studies have been focusing on complementary aspects, such as how persuasive sources might affect both primary levels of cognition (e.g. source serving as peripheral cue, source influencing the direction of thoughts

or source influencing the amount of thoughts) and secondary or metacognition levels (e.g. thought confidence) [23]. It has to be underlined that most of the conducted research explores attitude change, as it is assumed to serve as key mediation construct with respect to other targets of change, such as emotions, behaviors and beliefs [23].

In the contemporary theories on attitude formation and update, such as Dual-Processing theories (e.g. [171, 36]) and Dual-System theories (e.g. [77]), researchers postulate that several factors, including source factors, can affect attitudes through processes such as:

1. acting as peripheral cue or heuristics [172];
2. acting as issue-relevant argument [126];
3. impacting the amount of processing taking place [59, 102, 180];
4. biasing the nature of thoughts [37];
5. impacting structural properties of thoughts (e.g. thought confidence) [24].

With regard to the factors that might be used as cues a relevant role is played by attractiveness of the source [100] and credibility of the source [24], often referred to as source reliability [53].

6.4 The Reliability Game features

6.4.1 The Reliability Game method

As explained in the previous chapter, during each round the cards need to be positioned on a game board and the selected position reflects the weight of belief that the information in a card provides toward some subsets of the mutually exclusive and collectively exhaustive hypotheses.

6.4.2 Training and time

The game session has been designed to last around thirty to forty minutes, which is quite fast. The facilitator first introduces the participant to the game scope, rules and scenario. The game does not foresee a real pilot run. On the other hand the facilitator guides the player when positioning the first card, making the initial brief easy and short.

6.4.3 Domain of application and example

The method has been developed in the context of Maritime Situational Awareness, thus with the specific objective of assessing the impact of source factors on human Maritime Situational Assessment and resulting SAW. It has been played with twenty-one players, subject matter experts of Maritime Situational Awareness.

Although it has been developed with respect to the maritime domain, it is important to underline that the game could be easily tailored to other domains (e.g. air traffic control, medicine, emergency and disaster recovery). Moreover, the method shows its potential to assess the impact by other factors related to uncertainty and information quality, such as trueness and precision.

6.5 The Reliability Game and the other methods in the context of Situational Awareness

Differently from the other Human Factor methods available in the context of Situational Awareness assessment (see Section 6.2), the Reliability Game main focus is not on SAW, but rather on the Situational Assessment process. Therefore, it does not provide a measure of SAW (contrary to SAGAT, SART, SALSA) or a set of SAW requirements (e.g. SAW requirement Analysis [73]). The Reliability Game objective is to evaluate which and how much information and source factor impact human beliefs, which are assumed as basic constructs that build up SAW. Given that the game is not measuring SAW, the correct estimation of the true hypotheses by the participant is of secondary importance. In fact, the analysis is concentrating mainly on the extent and direction of belief changes induced by the above-mentioned factors.

The Reliability Game is a simulation technique (each round corresponds to a simulation), however the simulation is not performed in a high-fidelity simulator. In fact, the participant is just presented with a scenario map, briefed on a scenario story and presented with incoming information reported on cards. The design choice was driven by attempt to reduce the impact of information visualisation and system familiarity, while focusing on the information processing. Although the complexity of the element design can be regarded as medium, as it might require the support of subject matter experts to the scenario and cards design, the effort and required resources are less than for other common HF methods, such as SAGAT or SALSA. This has the advantage of making the Reliability Game a quick, easy to apply and low-cost approach (such as SART and C-SAS). Moreover, the method is characterized by a low training and facilitation complexity.

The Reliability Game presents both elements of freeze probe techniques (e.g. SAGAT, SACRI and SALSA) and self-rating techniques (e.g. CARS, MARS, SARS, SART and C-SAS). In fact, like in freeze probing techniques the simulation is frozen during query administration to the participant. The method has only a three item predefined set of queries that require a self-rating from the participant, that is:

1. request to position the card on the board at each freeze;
2. request to rate the source quality at each freeze (only in specific simulations);
3. request to rate the confidence in the different hypotheses at the end of each simulation.

Table 6.1 reports the main elements of a qualitative comparison between the Reliability Game and the other HF methods described in Section 6.2. The first column lists the comparison criteria, which are instantiated for the Reliability Game in the second column. The comparison criteria have all a binary outcome (✓= equal, • = not equal), with the exception of the element design complexity (✓= equal, • = not equal – Reliability Game lower, × = not equal – Reliability Game higher). This table highlights how the Reliability Game shares with other HF methods important elements such as the type of technique (number of HF methods with equal value $n_{\checkmark}$=8), simplicity of query design ($n_{\checkmark}$=5), simplicity of facilitation ($n_{\checkmark}$=5) and the simplicity of query administration and data collection ($n_{\checkmark}$=8). Moreover, the method presents a low cost ($n_{\checkmark}$=6), low execution time ($n_{\checkmark}$=5) and low training time ($n_{\checkmark}$=9).

The comparative analysis showed that the Reliability Game presents many common elements with CARS, MARS and SART. On the other hand, it highlights some of the innovative aspects of this method, namely its main focus and the dimensions analysed. Those aspects are intrinsically linked to the innovative scope of the method that is to guide the design of reasoners and algorithms to be used in support systems. An additional innovative feature of the Reliability Game is the data collection technique. Although many techniques as previously mentioned present a low level of complexity with respect to the data gathering, to the best of author's knowledge, positioning the card on the game board is an original way of answering a SAW related query and directly record the participant belief, while minimising the intrusiveness of the procedure. In fact, positioning the card is actually supporting the assessment process, instead of interrupting it to answer to the query. Moreover, the gamified approach creates an engaging context, as showed by the participant feedback. This has a direct impact on participants' message processing mechanism. In fact, it has been demonstrated that engagement enhances the information elaboration motivation [13, 213], leading to a more in-depth consideration of the message content and to minor reliance on cues. Moreover, the feedback from the participants, show that the game is not only perceived as engaging,

but also as realistic, relevant with respect to operational needs and effective in the elicitation component.

6.6 Discussion

Although researchers have acknowledged the usefulness of the Situational Assessment construct in providing guidance and valuable information for system design, it has only partially received attention. It is desirable that further research is conducted on methods to evaluate Situational Assessment, in order to further understand its potential with respect to a comprehensive system T&E.

This section presents a comparison of the Reliability Game and other human factors methods available in the context of Situational Awareness assessment.

The comparative analysis between the Reliability Game and other thirteen HF methods available in the context of Situational Awareness assessment, shows that although the former shares many common elements with some of the latter (e.g. CARS, MARS and SART) it also presents some unique features. In fact, the Reliability Game does not provide a measure of Situational Awareness, but rather an evaluation of which factors might influence human beliefs and overall SAW. The gamified approach introduces an engaging component in the setup and the specific design of the method allows the collection of data expressing second-order uncertainty. The data collected provide useful insight into the Situational Assessment aspects.

Although this analysis has been performed specifically for the Reliability Game, most of the aspects extend to the K2AG framework in general. In fact, another game, called the Variety Game, is a specific adaptation of the Reliability Game to be used in the T&E of the systems in support to maritime SAW. More specifically it was designed as part of a broader set of T&E experiments of the prototype developed under the EC H2020 Big Data Analytics for Time Critical Mobility Forecasting (datAcron) project [262]. The results obtained in the experiment have not been fully analysed at the time of writing this thesis and will be part of future work.

Criteria	Method												
	Reliability Game	SAW Req. Anal.	SACRI	SAGAT	SALSA	SASHA	SPAM	CARS	MARS	SARS	SART	C-SAS	SABARS	SA-SWORD
Technique	Freeze probe / Self-rating	•	✓	✓	✓	•	•	✓	✓	✓	✓	✓	•	•
Main focus	Situational Assessment	•	•	•	•	•	•	•	•	•	•	•	•	•
Dimensions analysed	Source Factors impact	•	•	•	•	•	•	•	•	•	•	•	•	•
Set-up	Game	•	•	•	•	•	•	•	•	•	•	•	•	•
Query design	Simple	•	•	•	•	•	•	✓	✓	•	✓	✓	•	✓
Element design (e.g. scenario, cards)	Medium	×	•	•	•	•	•	✓	✓	•	•	•	✓	•
Execution time	Low	•	•	•	•	•	✓	•	•	✓	✓	✓	•	✓
Cost	Low	•	•	•	•	•	•	✓	✓	•	✓	✓	✓	✓
Training time	Low	•	✓	✓	✓	•	•	✓	✓	✓	✓	✓	✓	✓
Facilitation complexity	Simple	•	✓	•	•	•	•	✓	✓	✓	✓	•	•	•
Query administration complexity	Simple	•	✓	✓	✓	•	•	✓	✓	✓	✓	•	•	✓
Data Collection complexity	Simple	•	✓	✓	✓	•	•	✓	✓	✓	✓	•	•	✓
Possible Mathematical Modelling	Direct	•	•	•	•	•	•	•	•	•	•	•	•	•
Domain of application	Multiple	✓	•	✓	•	✓	✓	✓	✓	•	✓	✓	•	•

Table 6.1 Comparison of the Reliability Game method and other HF methods in the context of Situational Awareness assessment

Chapter 7

Case study: MARISA Game

7.1 Motivation

7.1.1 EC H2020 MARISA Project

The Multi-Source Dynamic Bayesian Network for Behavioural Analysis Service [52, 9, 10] is one of the innovative fusion services developed in the EC H2020 Maritime Integrated Surveillance Awareness (MARISA) project is a border and external security project. The project main goal is to increase end-users operators SAW through the provision of an integrated toolkit of information fusion services. Those services span from low level data fusion services, able to efficiently fuse ship tracks, up to high level information fusion services. The high level fusion services aim at supporting the understanding of the current situation, for example through behavioural analysis services, and the prediction of future situations through predictive analysis services. The MARISA toolkit is envisioned as a complement to the European Common Information-Sharing Environment (CISE) [78], currently under joint development by the European Commission and the Member States. This infrastructure will integrate surveillance systems and networks, allowing for national and international seamless data and information sharing. The legacy systems of each Member State should connect to the infrastructure through national CISE nodes. The MARISA toolkit is developed as the potential fusion engine that would fuse incoming information provided through CISE or other national systems. The project is organised in a two-phase iterative design approach, each culminating in the validation activities of five operational trials run by the MARISA end-user partners. More specifically, the trials are: (i) the Northern Sea Trial, run by the Netherlands Coast Guard; (ii) the Iberian Trial, run jointly by the Guardia Civil and Portuguese Navy; (iii) the Ionian Trial, run jointly by the Italian Navy and Hellenic Ministry of Defence; (iv) the

Aegean Trial, run by the Hellenic Ministry of Defence; (v) the Bonifacio Trial, supported by the French Navy.

The MARISA toolkit is configured for the needs of each trial using different data fusion services, networking set-ups and fusion services.

7.1.2 Multi-Source Dynamic Bayesian Network for Behavioural Analysis

The MARISA Multi-Source Dynamic Bayesian Network (MSBN) for Behavioural Analysis Service is a probabilistic based vessel behavioural analysis tool. While the reader is referred to Section 2.5 for basic notion on dynamic Bayesian networks, this section reports specifically on the service design. The MSDBN presents a layered hierarchical structure, proposing an easy yet powerful mechanism to define a multi-source Bayesian network accounting for source reliability. In fact, the network structure is composed by two main layers, namely the Situation Layer and a Reporting Layer.

In the Situation Layer a situation of interest is modelled at different levels of abstraction. As described in [10], following [258] a situation is defined as "an external semantic interpretation of sensor data" and following [80] the semantic statement can be *True* of *False*. Depending on the level of abstraction, the situations can be classified either as *elementary situations* or as *abstract situations*. The first ones (lower hierarchical levels) are the ones for which the existence probability is mapped deterministically, while the later (higher hierarchical levels) are the ones for which the existence probability depends on other situations.

The MSDBN is able to manage evidences received by different kind of information sources. The network structure has been extended, in fact, to include a Reporting Layer beside the Situation one. This layer allows an individual consideration of reliability degrees that might differ from one source to another. Evidence received by the MSDBN is not directly entered on the situation variables, as this is not a direct evidence on the state of the variable, but rather a report by some source on the variable state. Therefore, evidence is entered through the Reporting Layer, which accounts for the reliability of the source of information. In the MSDBN the behaviour of a reliable source is modelled as a truthful source, while the unreliable source is modelled as a randomiser [20]. The randomiser corresponds to a source that provides reports that are equally likely and uncorrelated with the true state of the world.

Figure 7.1 shows a small portion of one time slice of the MSDBN, with the different hierarchical levels and layers. In this figure we can notice how the reporting layer is composed by several *reliability structures* (see Section 5.3). Evidence is entered in the Rep_i and $Source_i$

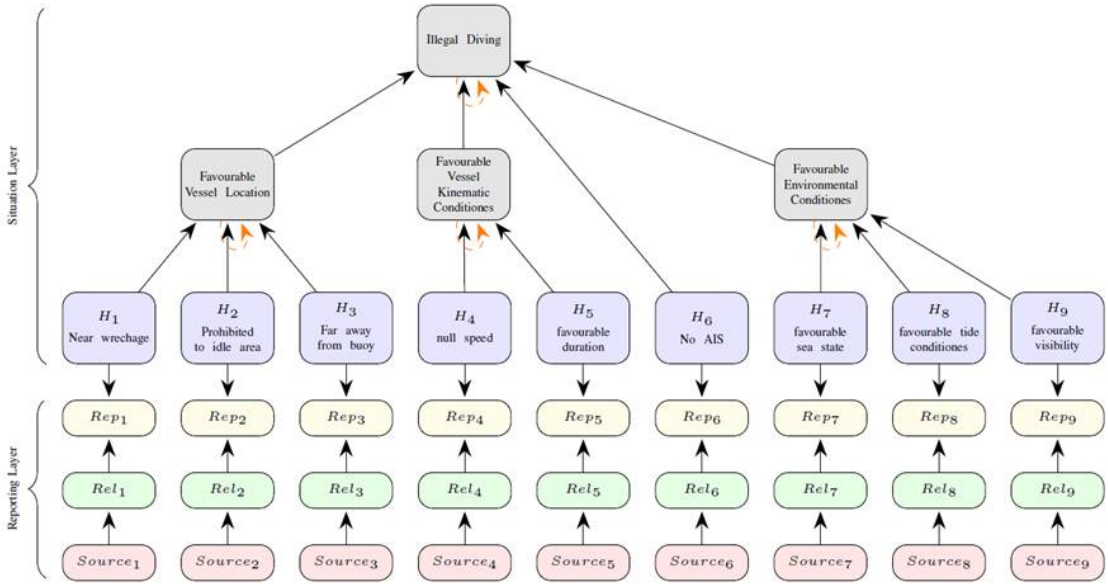


Fig. 7.1 Portion of the MARISA MSDBN Behavioral Analysis structure (reproduced from [10])

nodes and propagated through the network, allowing to take into consideration both the variety and the potential difference of reliability of sources. Currently the only source factor modelled is *source type*. This modelling choice was dictated by the fact that some limitations on the current CISE data model do not allow to properly share source quality related information [53]. However, if such information would became available the MSDBN could be easily extended to account for additional source factors.

7.1.3 Initial validation activities

A first prototype was developed addressing only one of the different use cases identified by the MARISA end-users: illegal diving, illegal-unregulated-unreported fishing, illegal immigration, smuggling of goods and piracy. The initial prototype was successfully deployed in the first Northern Sea Trail (see Figure 7.2).

During this trial the Netherlands Coastguard organised *MARISA Alert*, a dedicated training exercise with three relevant operational scenario, one of which was illegal diving. The exercise involved three ships of the Netherlands Coastguard, namely the watch ship *Guardian*, the patrol ship *Visarend* and the support ship *Terschelling*. The MARISA Toolkit was connected to a live feed of the Coastal Surveillance System, while the ships simulated anomalous behaviours. From the Netherlands Coastguard back-up operations facility, where MARISA

(a) Netherlands Coastguard back-up operations facility

(b) Netherlands Coast Guard ships

Fig. 7.2 First MARISA Northern Sea Trial

Toolkit was situated, it has been possible to observe how the MSDBN service generated an accurate and early alert to the operator, corresponding to the simulated suspicious event.

7.2 The knowledge engineering problem statement

In the first design phase, as previously mentioned, the MSDBN prototype has been tailored only to one use case presented by the Netherlands Coast Guard, namely illegal diving. This is a problem often encountered in the Northern Sea. In fact, shipwrecks of the Golden Age, which are protected by national legislation, are an appealing target for treasure hunters. After the network structure was developed and validated with experts, the MSDBN CPTs have been elicited using a traditional questionnaire approach, for the $t = 0$ time slice, while for $t > 0$ the CPTs are defined following the algorithm proposed in [80].

It has to be noticed that the dimension of the initial MSDBN prototype was smaller than the final one. Therefore, the number of questions was relatively contained. Moreover, a colour coding was adopted in order to facilitate the intuitive understanding of the state of the MSDBN parameters included in the questions (e.g. variable name is written either in blue or red, if the variable state is *True* or *False* respectively). Nevertheless, a considerable effort has been required by the domain expert due to the complexity of the task.

During the second design phase, instead, an innovative KA method, based on the K2AG framework has been adopted. The MARISA (MARItime Surveillance knowledge Acquisition) Game is specifically designed to provide an engaging environment to increase the quality of the information elicitation and to facilitate open discussions, which are very valuable in system design. The need for a different KA approach draws from the one of providing to the experts an easier tool for the elicitation of CPTs of the final MSDBN, which where considerably more than in the first phase of the project. More in details, the

Fig. 7.3 MARISA Game board

final MSDBN service includes one sub-network for each of the illegal activities selected as relevant use-cases by the end-users for the trials in which the service has been validated (see Section 7.1.3). The structure of those sub-networks has been defined on the bases of relevant literature on maritime anomaly detection (e.g. [139, 33]), news regarding trends of such illegal activities (e.g. illegal immigration in Italy) and discussions with experts. Beside the complexity of the task, another important aspect that played a role in the decision of applying K2AG techniques was to facilitate the user in the task execution by providing context and support in the question interpretation. In fact, the questions presented in the first phase KA were not always easily understood.

The next section will provide details on the MARISA Game design, on the two experiments run and the results obtained.

7.3 MARISA Game

7.3.1 World design

The MARISA Game is a one round multi-player game, set in a maritime scenario. Each participant plays the role of a junior navy officer assigned to one of the duty locations on one of the islands of a fictitious archipelago, called the MARISA Islands.

The six MARISA Islands are depicted on the game board (Figure 7.3). This archipelago is very peculiar as all the five bigger islands suffer of one and only one illegal activity (i.e. piracy, IUU fishing, illegal diving, smuggling of goods and illegal immigration). This peculiarity makes those islands a perfect duty location to train junior officers, as they need to focus only on one specific activity and gain domain knowledge on it. Those islands are named after the illegal activity taking place on them (e.g. Piracy Island, IUU fishing Island, Illegal diving Island, Smuggling of goods Island and Illegal Immigration Island), while the smaller central island is the Reliability Island, where the main Command Centre for the area is located. On this island the junior officers are trained on the assessment of the reliability of the sources of information providing reports to the Command Centre.

The players need to gain a sound domain knowledge in order to proceed in their career. The level of expertise of a player is represented by the number of knowledge tokens that the player collects. The first player that collects all the knowledge tokens is the winner.

7.3.2 Content design

At the start of the game session the players are presented with an informed consent form to be signed, explaining the aims of the experiment and the way in which the data will be treated. In addition a short pre-game personal information form is provided. Then the players are introduced by a facilitator to the game core and to the game mechanics. As in the Reliability Game, in this pre-game brief the scenario, rules and different game elements (e.g. game board, cards and query track sheets) are presented to the player. After the game, instead, the players are requested to fill in the K2AGQ post-game questionnaire.

For each island there is a corresponding knowledge card deck. The knowledge cards provide knowledge constructs in the form of portions of the MARISA MSDBN sub-networks of the different illegal activities. Two different kinds of knowledge cards have been designed and used. More specifically, the cards that contain the knowledge constructs of the MSDBN Situational Layer present such knowledge in a graphical form, which recalls the graphical representations of Bayesian networks (Figure 7.4). For the Reporting Layer, instead, the knowledge structures are presented in natural language form (Figure 7.5). Those knowledge structures are included only in the Reliability Island card deck. Similarly to the first phase KA, a color coding is adopted as a visual cue to support experts assessment (i.e. green for a *True* state and red for a *False* state). The query variable, which is always presented in the upper part of the graphical knowledge structure, is white as the players need to state their belief that the state of a certain parameter might be *True* or *False*, based on state of the other ones. Overall, the parameters refer to the ship characteristics, ship kinematic conditions,

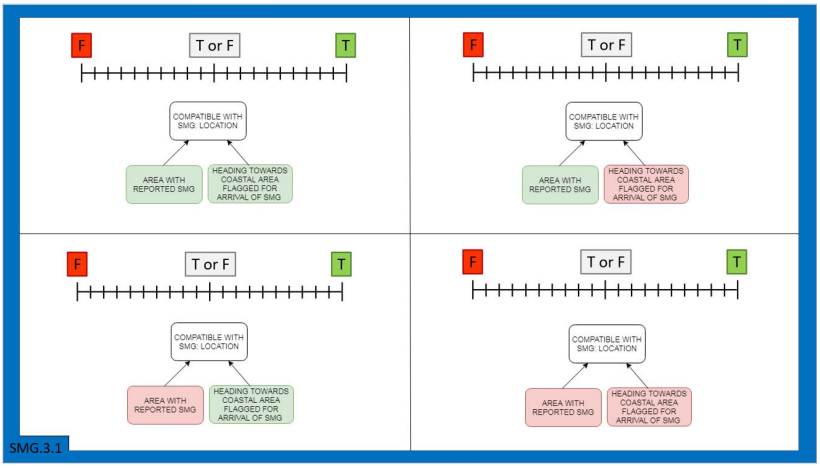


Fig. 7.4 Example of knowledge card from the Smuggling of goods Island decks

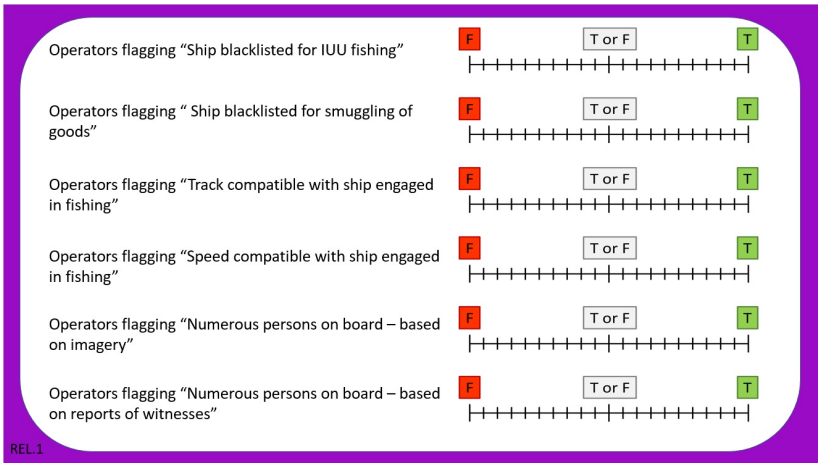


Fig. 7.5 Example of knowledge card from the Reliability Island deck

geographical factors and environmental conditions that are compatible with the analysed illegal activities and the source type reporting information.

In addition to the knowledge card decks there is a commendation card deck that contains cards (e.g. Figure 7.6) to request at any time additional knowledge cards or to move with the helicopter to an island of choice.

An additional color coding is used to help players refer to a specific illegal activity. The color code is repeated on the knowledge cards of a specific deck and on the respective query track sheet.

The data gathering area, differently than in the Reliability Game, is fully included in the card area containing the knowledge structure. It can be noticed that in the case of the MARISA Game the data gathering area is a segment between the two possible hypotheses

Fig. 7.6 Example of MARISA Game commendation card

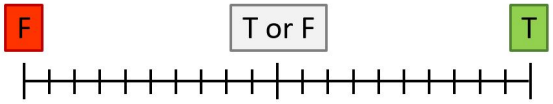


Fig. 7.7 MARISA Game data gathering method

on the query variable state (see Figure 7.7), that are $H_1 = False$ or $H_2 = True$ (see Section 4.3.3 for additional details).

Additional information has been collected from the players discussions, which has been captured in the facilitator notes. In fact, in the MARISA Game the facilitator has also the role of data collector.

7.3.3 System design

In order to collect knowledge tokens, players need to proceed in the game by moving on each island path. This is achieved by throwing a dice and move the pawn of a number of corresponding tiles. This path tiles have a colors that correspond to the following actions or areas:

- Green: no action associated;
- Beige: allows to look at the knowledge cards as per result of another dice;
- Purple: allows to pick one card from the commendation card deck;
- Red: helicopter landing area tile and start tile;
- Blue: harbors to move between islands.

Variable	Description	Frame
M	Message conveyed by a *knowledge card*	$\{M_1^1, \ldots, M_n^d \ldots, M_{n_{max}}^{d_{max}}\}$
KS	Knowledge structure conveyed by a message	$\{KS_{1,1}^1, \ldots, KS_{s,c}^d, \ldots, KS_{s_{max},c_{max}}^{d_{max}}\}$
D	Variable with dependency from Q state	$\{True, False\}$
Q	Query Variable state (i.e. hypothesis)	$\{H_1, H_2\} = \{True, False\}$

Table 7.1 MARISA Game state

Islands are connected via Sea Lines of Communication (SLOCs) that are represented on the board game. Those SLOCs can be travelled with the ships that are available at the anchorages of the islands ports (harbor tiles). The SLOCs connecting the bigger islands are two-ways, while the ones connecting those islands to the Reliability Island are one-way.

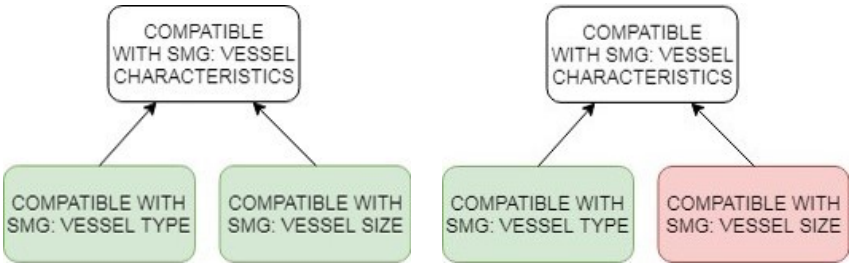
As previously mentioned, each island has a corresponding knowledge card deck with portions of MARISA MSDBN sub-model, which we will denote as a knowledge structures $KS_{s,c}^d$, where $d = 1, \ldots, d_{max}$ represents the number of card decks, $s = 1, \ldots, s_{max}$ the knowledge structures in the card deck d and $c = 1, \ldots, c_{max}$ the possible presentations of the knowledge structure s for deck d. Each knowledge card presents a message M_n^d, where $n = 1 \ldots n_{max}$ are the cards in a card deck d. A message can contain one or more knowledge structures. Those are composed by a query variable (Q) and other variables (D) that have a direct dependency with Q as per MSDBN sub-model. More specifically, for the MSDBN sub-models considered in this game the D variables are all parents of the query variable. Each knowledge structure is presented to the player in all its configurations, that are all the possible combinations of D states.

Table 7.1 and Table 7.2 summarise the MARISA Game state and view respectively.

Figure 7.8 shows an example of four different configurations of the same knowledge structures included in a knowledge card of the Smuggling of Good (SMG) Island card deck.

In the MARISA Game, the knowledge cards contain not only the message, but also the data gathering area. The player needs to provide the belief on the query variable (white node of the network) state. The belief is recorded on the data gathering segment by the players. The probabilities contained in the CPT are obtained through the mapping proposed in Table 4.4. The other nodes have a green color to indicate that their state is *True* or red to indicate that their current state is *False*. Figure 7.9 presents beliefs that could be recorded for the example knowledge structure in Figure 7.8 and the resulting CPT for the query variable.

Players fill in the knowledge card and then put it aside. They are allowed to go back and look at them when filling in the next ones. As previously mentioned, to obtain knowledge cards the player need to reach a beige tile on the map and trow a ten facet dice. If we denote



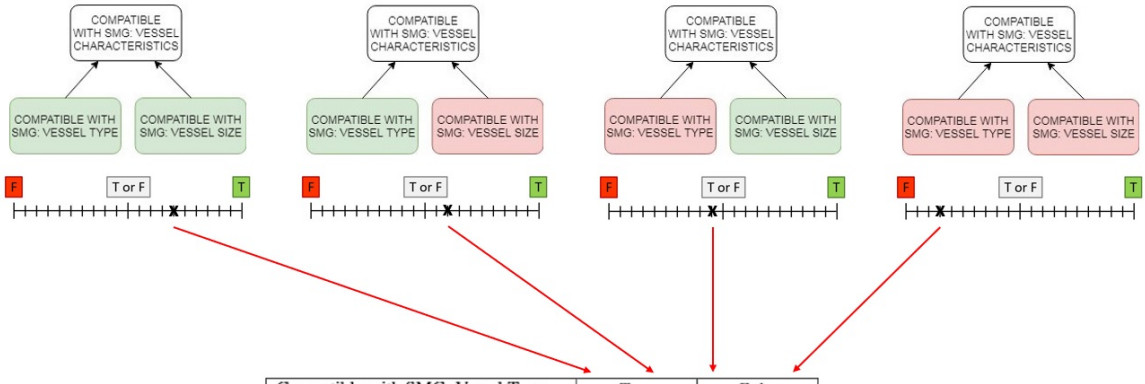
(a) The knowledge structure with both the parent variables with a state *True*

(b) The knowledge structure with the first the parent variable with a state *True* and the second one with a state *False*

(c) The knowledge structure with the first the parent variable with a state *False* and the second one with a state *True*

(d) The knowledge structure with both the parent variables with a state *False*

Fig. 7.8 Examples of different configurations of the same knowledge structure



Compatible with SMG: Vessel Type	*True*		*False*	
Compatible with SMG: Vessel Size	*True*	*False*	*True*	*False*
True	0.70	0.60	0.45	0.15
False	0.30	0.40	0.55	0.85

Fig. 7.9 Example of the CPT of the query variable *Compatible with SMG: vessel characteristics*

Variable	Description	View
M	Message conveyed by a *knowledge card*	Provided[1]
KS	Knowledge structure conveyed by a message	Provided
D	Variable with dependency from Q state	Provided
Q	Query Variable state (i.e. hypothesis)	Assessed[2]

[1] Provided = item value provided to the player;
[2] Assessed = player has to assess the item and record the assessment.

Table 7.2 MARISA Game view

the number obtained through the dice as N, for $N \leq s_{max}$ the player receives a knowledge card presenting some configurations of the N^{th} knowledge structure. If the participant already received all the knowledge cards corresponding to the considered knowledge structure, he or she will not receive an additional knowledge card. For $N > s_{max}$ then the player can choose the knowledge card to receive. It has to be mentioned that the game mechanics to obtain the knowledge cards have been slightly modified on the fly in both experiments in order to make the game quicker to accommodate operational contingencies that limited the expert time availability (Section 7.5).

As the players fill in the cards they advance in their career and this is tracked on their career status sheets, on which the knowledge card already received can be marked (Figure 7.10). Moreover, the full structure of the MSDBN sub-models for the different illegal activities are depicted. When players complete a knowledge card deck they earn the corresponding knowledge tokens and are entitled to receive a card from the commendation card deck.

The player who first obtains all the promotions (i.e. plays all the card decks) is the winner. The game continues until all the players have finished all the cards.

To summarise the following game mechanics have been identified for the MARISA Game:

1. the assessment of hypotheses relative to maritime anomalies;
2. the use of cards to communicate messages to the player;
3. the investigation component;
4. the rating of the player beliefs related to the knowledge constructs provided through cards;
5. the collection of knowledge tokens.

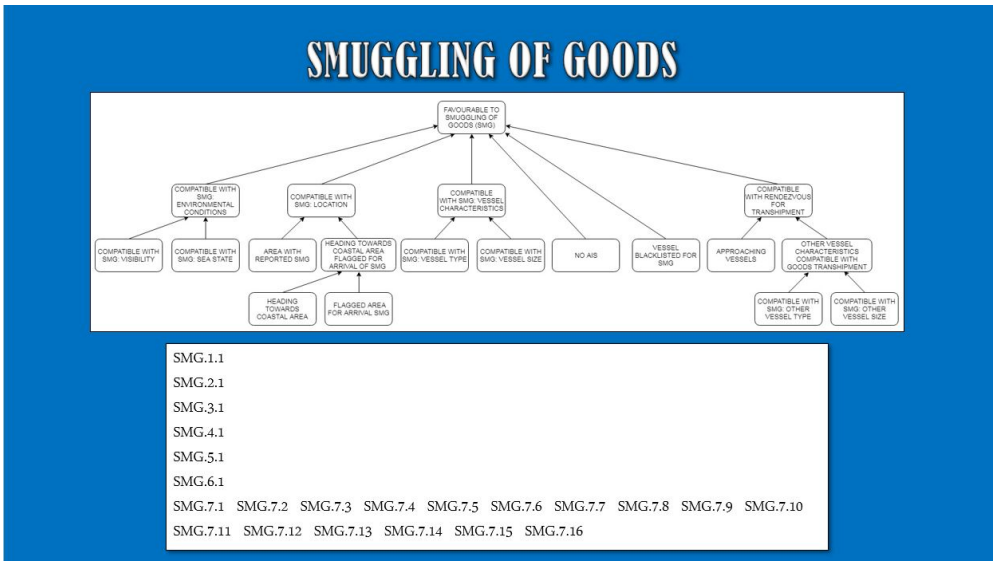


Fig. 7.10 Example of MARISA Game carrier status recording sheet

It can be noticed that the first four mechanics are the same as the ones of the Reliability Game.

Figure 7.11 presents a simple diagram of the flow of the game. As previously mentioned at any time during the game players can use their commendation cards. This process has been omitted in this figure for clarity.

7.4 Knowledge acquisition experiments

The K2AG approach has been successfully applied to the MARISA second phase in which one game session has been played with the participation of the Spanish Guardia Civil representatives and one with the representatives of the Italian Navy. This allowed to adapt the MSDBN design to the maritime illegal activity patterns observable in the Iberian area and the Ionian Sea area respectively.

The main objectives of the experiments can be summarised as follows:

1. validation of the MSDBN structure;
2. collection of the players belief to be transformed in conditional probabilities to be included in the MSDBN;
3. evaluation of the MARISA Game.

The first experiment (EXP1) has been run at the Guardia Civil premises in Madrid with four experts, namely three maritime law enforcement experts from the Guardia Civil and

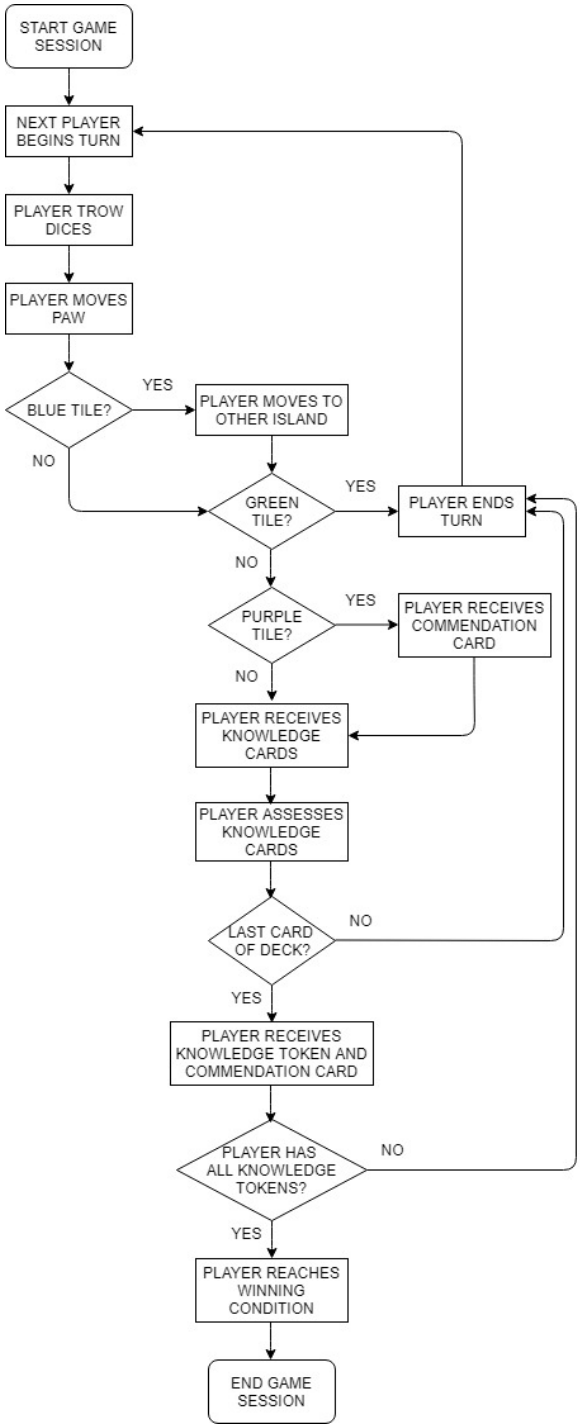


Fig. 7.11 Diagram of a MARISA Game session

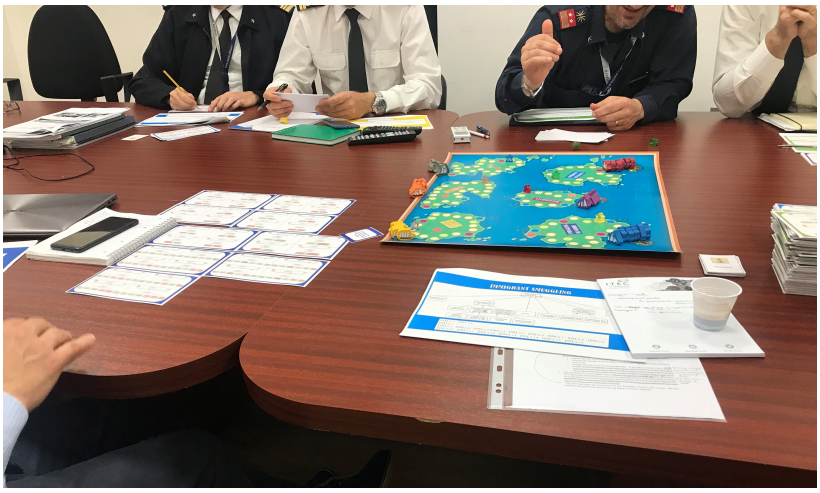


Fig. 7.12 MARISA Game taking place at the Italian Navy premises

Feature	Specification	EXP1	EXP2
Gender	Male	75%	100%
	Female	25%	0%
Age	Average	34.3 years	49.8 years
	Standard Dev.	5.4 years	16.3 years
Status	Law Enforcement / military	75%	80%
	Civilian	25%	20%
Nationality	Italian	0%	100%
	Spanish	100%	0%

Table 7.3 Participants demographics and characteristics

one engineer from the company leading the MARISA Iberian Trial. A second experiment (EXP2) has been run at the Italian Navy premises in Rome with five participants (Figure 7.12), namely four military experts and one engineer from the company leading the MARISA Ionian Trial. Table 7.3 provides additional information on the participants characteristics and demographics.

7.5 Results and discussion

From the analysis of the game results it has been possible to deduce that all three objectives discussed in Section 7.4 have been met. In fact, the game allowed collecting information regarding characteristic behaviours of illegal activities, validating the MSDBN structure, and the data for the conditional probabilities of the MSDBN. More specifically, the data collected in the form of beliefs in the two experiments has been transformed into subjective probabilities, following the mapping presented in Table 4.4. Those subjective probabilities

have been used as conditional probabilities to populate the CPTs of the MSDBN, as explained in the example in Section 7.3.3.

The game overall was well perceived and after a short explanation of the rules, players started playing confidently, without major support from the facilitator, whose main role becomes to distribute the game items. This allowed the facilitator to focus on the discussion and collection of verbal inputs. During those discussions qualitative (i.e. structure) and quantitative aspects related to the MSDBN have been analysed in depth.

Although the mechanics selected for the MARISA Game did not prove inadequate, due to time constraints dictate by operational duties of the persons participating to the experiments, it was agreed at the beginning of the experiment sessions to increase the number of knowledge cards provided to the player at each turn and to modify the winning condition. More specifically it was agreed that the winner is the players collecting more knowledge tokens in a certain amount of time. Due to this change the players where also invited to focus on the Islands corresponding to the illegal activity on which they had more expertise.

As discussed in Section 7.3.2, there are two different types of knowledge card layouts. One of the verbal feedback received by most of the players is that the graphical presentation of the knowledge structure in the form of a naive BN, is very useful if there are not many variables (D) in addition to the query variable. The average response was that the graphical presentation is convenient if $D \leq 4$, while for $D > 4$ the presentation in natural language form appears more adequate. However, those are preliminary observations and further investigation should be performed.

As mentioned, the use of the game allowed to verify and then validate the MSDBN structure. In fact, during the play test phase some wrong design assumption where highlighted by a player. These assumptions have been removed in the final MSDBN structure, which has been presented in the two experiments. The presentation of knowledge cards acted as an inject, naturally stimulating discussions and substituting the questions of structured interviews approaches. From these discussions during the game sessions it appeared that the proposed MSDBN structure is appropriate.

Differently than the traditional questionnaire approach, the MARISA Game, both from the verbal feedback and the feedback questionnaire, appears to be perceived as a more engaging and stimulating method to elicit the CPTs for Bayesian networks.

The results of the K2AGQ show that the overall attitude (Figure A.1) towards the game is positive. The game was perceived as an adequate experimentation tool (Figure A.6). Although, players did not have the initial feeling that it would be easy, they felt that the facilitation approach of the game, the structure and its content made them confident that the experiment would support the stated goal (Figure A.3). However, from the feedback it

appears that it was not very clear for all the players how the game relates to the stated goal. That might be due to the fact that most of the players were not familiar with the MSDBN Service concepts nor the MARISA project. Therefore, the introduction brief providing context to the participants should have been more detailed and focused on this aspect. The answers suggest that players did not have a preference of using games over other KA methods. However, none of them was involved in the first phase KA session. The responses regarding flow (Figure A.5) show that the participants were concentrated and occupied with the game, but as expected the MARISA Game is not a high-engagement game that leads to a full immersion, forgetting about time and surroundings. An overall positive feedback has been provided with respect to: (i) the usability of the game as a system (Figure A.4), (ii) the facilitation process (Figure A.4), (iii) the sensory and imaginative immersion (Figure A.8) and (iv) the players' satisfaction (Figure A.7).

Finally, an interesting result was observed through the feedback related to workload (Figure A.9). In fact, it can be observed how the players felt that the task to accomplish was complex and demanding in terms of mental activity. However, they stated that at the same time the work to achieve the level of performance obtained was moderated. This might be related to the fact that the game is considered easy (Figure A.2). This result is in contrast to the one received for the KA session using the questionnaire approach and positively support the concept of using analytical games for KA.

Moreover, free form written feedback highlights the positive attitude of the operators towards the use of traditional game methods (i.e. board games) for the purposes beyond entertainment (i.e. "Board games are a fantastic classic way to have fun! Adopting them to obtain feedback and motivate people is a great idea!").

One player stated that the game was unfolding too quickly to appreciate the relevance and ability of such approach to support the stated goal. However, he also admitted to be severely sleep deprived due to work related activities.

An interesting observation relates to consistency of answers from the player. In fact, some players requested to the facilitator some duplicated knowledge cards, probably because they forgot to mark them on the recording sheet. By comparing the players' answers it has been possible to observe how some players provide fairly consistent answers when exposed to the same stimulus, up to a player that provided the same answers in all the three duplicated knowledge cards. Instead, another player proved to be inconsistent when providing the answers to the duplicated cards. Although this is just a preliminary observation, these results might suggest that similar strategies should be investigated further in future studies to determine if they could be employed to better characterise the players' profile and, therefore, the validity of the answers provided.

The results of both experiments wield in general a positive outcome of the game evaluation as they allowed to collect the planned data and to positively engage the experts. In addition to the evaluation of the MARISA Game from a "fit for purpose" perspective a positive evaluation of the game has been obtain with regard to the predictive validity (i.e. collected knowledge utility). In fact, the MSDBN designed with the support of a K2AG has been successfully tested and evaluated in two of the MARISA project operational trials, namely the Northern Sea trial and the Iberian trial. As can be noticed, the first one is not one of the areas for which the MSDBN was specifically tailored by playing the game with local experts. However, the system evaluation suggests that the knowledge collected could be representative also of other geographical regions. This will be subject to further investigations. Further, the MSDBN service will be evaluated in an additional operational trial taking place in the Ionian Sea in the near future.

Chapter 8

Conclusions

In this thesis we have introduced the concept of analytical games. Analytical games are platform independent simulation games, designed not for entertainment, which main purpose is research on decision-making, either in its complete dynamic cycle or a portion of it (i.e. Situational Assessment and Situational Awareness steps). Situational Assessment is one of the main cognitive tasks, which intelligent systems are supporting with higher degrees of automation. An important step for the design of those systems is the knowledge engineering phase of knowledge acquisition.

In this doctoral work we discuss how analytical games could support the knowledge engineering task in general and the knowledge acquisition task specifically. More in details, this work introduces the Knowledge Acquisition Analytical Game (K2AG), which is an innovative game framework for the design of analytical games that aim at collecting data for the design of such systems. The framework introduced in this doctoral work, was born as a generalisation of the Reliability Game, which on turn was inspired by the Risk Game. This thesis reports on the main elements of that characterise K2AG, namely the scenario, the knowledge cards and the Geometric Belief Data Gathering (GBDG) method. The GBDG is an innovative approach to belief gathering, which takes advantage of geometrical features of simple shapes (e.g. a triangle) to easily collect players' beliefs. Those beliefs once collected can be mapped to subjective probabilities or masses and used for algorithm design purposes. Although, K2AGs might use different means of conveying information to the players and to collect data, the use of knowledge cards appears to be effective and efficient.

Furthermore, this work describes the K2AG design life cycle, which extends simulation design frameworks, previously proposed in literature. More specifically, van der Zee's simulation gaming design framework has been extended in order to account for the fact that the design cycle steps should be modified to include the different kind of models that characterise the design of simulation games and simulations in general. In fact, it should

include a conceptual model (platform independent), a design model (platform independent) and one or more implementation models (platform dependent). In addition, the processes that lead from one model to the other have been mapped to design phases of analytical wargaming. Finally, the validation approach for K2AG is discussed. More specifically, based on a literature review the set of validation criteria is identified and the corresponding methods are proposed. Important steps of the validation activity are the evaluation of the game as a tool which is fit for purpose with respect to the research objectives and the evaluation of the player experience. To this end, a player experience questionnaire, which builds on other questionnaires available in literature, has been specifically developed for K2AGs. However, due to time constraints an in-depth validation of it has not been possible at the time of writing and will be part of future research activities.

Two instantiations of the K2AG framework, namely the Reliability Game and the MARISA Game, have been designed and analysed in details to show the potentialities of such approach.

The Reliability Game is a K2AG designed to characterize the impact of source factors, such as source quality and source type, on human situational assessment process. A qualitative analysis has been performed on the data gathered through an experiment run with the Reliability Game in order to evaluate the effectiveness of the game design and game mechanics. The variations of the players' belief assessments between the different rounds demonstrate that the proposed methodology effectively captures elements of source factors impact on SA. Moreover, the analysis allows assessing important aspects of the use of the rating scales, which might be relevant to the standardization efforts in communication of uncertainty (e.g. confidence, reliability). In addition to the collection of in-game data, a post-game data collection has been performed in the form of a feedback questionnaire. The results show that the game is perceived as engaging, but also as realistic, relevant with respect to operational needs and effective in the elicitation component. Moreover, the game scope and game mechanics were easily understood. In the quantitative analysis, instead, a Bayesian Network has been built and trained with the beliefs collected from the players, with the aim of deriving the network parameters for a latent variable (in the specific case the source reliability). The results reported indicate that beside being an engaging and easy elicitation method from the expert perspective, analytical games should be regarded as useful KA methods, able to collect relevant data to be used in the design of algorithms. In fact, the parameter learning showed promising results. In fact, the results obtained through a machine learning approach for the latent variable encoding source reliability are overall coherent with the intuitive understanding of source reliability and the players verbal feedback. Relevant results can also be observed with respect to the interpretation of the source reliability concept.

The above-mentioned analyses allowed validating the game along the validity dimensions identified for K2AGs, namely psychological, structural, procedural and predictive validity.

A comparative analysis between the Reliability Game and the other thirteen HF methods used in the context of SAW assessment has been performed, which shows that although the former shares many common elements with some of the latter it also presents some unique features. The analysis highlighted the simplicity of the approach in terms of query design, facilitation complexity, query administration and data collection. Moreover, it shows how execution time, cost and training time are low. Moreover, differently than the other human-factors methods analysed, the Reliability Game concentrates on the SA process and not merely on SAW as its end-state. Despite these positive aspects, it has to be highlighted the complexity of the game scenario design. In fact, domain knowledge is required in order to provide a simple, yet realistic scenario, able to stimulate correctly the expert reasoning. To mitigate the risk of not appropriate scenarios, experts should be involved during the design phase and should validate the assumptions made by the game designer.

The MARISA Game is another K2AG designed during this doctoral work as an alternative to traditional questionnaire approaches for the elicitation of a considerable amount of conditional probabilities to be encoded in Bayesian Networks. More specifically, the game has been used in order to perform the knowledge acquisition for a multi-source dynamic Bayesian network for behavioral analysis within the maritime context Additionally, it allowed validating the network structure with experts and the game approach.

From the analysis of the game results it has been possible to deduce that the objectives defined for the MARISA Game have been met. In fact, the game allowed collecting information regarding characteristic behaviours of illegal activities, validating the network structure, and the data for the conditional probabilities of the network.

The reported results indicate that beside being an engaging and easy elicitation method from the expert perspective, K2AG should be regarded as useful KA methods, able to collect relevant data to be used in the design of algorithms. For example, the RGBN parameter learning described in this work shows promising results. However, the model accuracy obtained is not very high, possibly due to the effect of additional source factors or issues such as a not ideal conditioning of the problem and the relatively small sample obtained from the experiment. Therefore, future work will not only concentrate on the development of specific K2AGs, but also on the development of best practices for a more rigorous experiment design with such games. The method described has been designed to specifically investigate one latent variable, however, it could be extended to include more than one. Therefore, such aspect, together with the extensions of the hypotheses set presented to the player (i.e. data collection with different geometrical shapes), will be subject to further investigation. The use

of the GBDG method as belief gathering technique will not only be further investigated with respect to the number of hypotheses that can be handled, but also from the perspective of its capability to collect data to be modeled in different uncertainty frameworks. To this end, additional analysis will be performed on the data collected through their modelling into the Evidential framework.

References

[1] Model, https://www.acm-sigsim-mskr.org/glossary.htm#m retrieved on 11 January 2019.

[2] Serious game, https://www.igi-global.com/dictionary/serious-games/26549 retrieved on 24 April 2018.

[3] Confidence. https://en.oxforddictionaries.com/definition/confidence.

[4] Energy performance certificate. https://en.wikipedia.org/wiki/Energy_Performance_Certificate.

[5] Reliability. https://en.oxforddictionaries.com/definition/reliability.

[6] Abt, C. (1970). *Serious Games*. The Viking Press, New York.

[7] Adams, E. and Rollings, A. (2003). *Andrew Rollings and Ernest Adams on Game Design*. New Riders.

[8] Aebischer, D. (2018). *Bayesian Networks for Descriptive Analytics in Military Equipment Applications*. CRC Press, Taylor & Francis Group.

[9] Anneken, M., de Rosa, F., Jousselme, A.-L., and Robert, S. (2018). Modelling dynamic Bayesian networks to identify suspicious behaviour. In *Proceedings of the Maritime Big Data Workshop 2018*, number CMRE-CP-2018-002. NATO-STO Centre for Maritime research and Experimentation.

[10] Anneken, M., de Rosa, F., Kröker, A., Jousselme, A.-L., Robert, S., and Beyerer, J. (2019). Detecting illegal diving and other suspicious activities in the north sea: Tale of a successful trial. In *Proceedings of the 20th Interational Radar Symposium*.

[11] Arsene, O., Dumitrache, I., and Mihu, I. (2011). Medicine expert system dynamic Bayesian network and ontology based. *Expert Systems with Applications*, 38(12):15253–15261.

[12] Baños, R., Botella, C., Alcañiz, M., Liaño, V., Guerrero, B., and Rey, B. (2004). Immersion and emotion: their impact on the sense of presence. *CyberPsychology & Behavior*, 7(6):734–741.

[13] Bakker, A. B. (1999). Persuasive communication about AIDS prevention: Need for cognition determines the impact of message format. *AIDS Education and Prevention*, 11(2):150—162.

[14] Balci, O. (1994). Validation, verification, and testing techniques throughout the life cycle of a simulation study. *Annals of Operational Research*, 53:121–173.

[15] Barrington, L., O'Malley, D., Turnbull, D., and Lanckriet, G. (2009). User-centered design of a social game to tag music. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pages 7–10, New York, NY, USA. ACM.

[16] Berlo, D. K. (1960). *The process of communication: an introduction to theory and practice*. Holt, Rinehart and Winston, New York.

[17] Birnbaum, M. H. and Mellers, B. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology*, 45:792–804.

[18] Birnbaum, M. H. and Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise and the judge's point of view. *Journal of Personality and Social Psychology*, 37:48–74.

[19] Boose, J. H. (1989). A survey of knowledge acquisition techniques and tools. *Knowledge Acquisition*, 1(1):3–37.

[20] Bovens, L. and Hartmann, S. (2003). *Bayesian epistemology*. Oxford University Press.

[21] Brandt, E. and Messeter, J. (2004). Facilitating collaboration through design games. In *8th Conference on Participatory Design*, Toronto, ON.

[22] Brathwaite, B. and Schreiber, I. (2008). *Challenges for Game Designers*. Charles River Media.

[23] Briñol, P. and Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology*, 20:49–96.

[24] Briñol, P., Petty, R. E., and Tormala, Z. L. (2004). The self-validation of cognitive responses to advertisements. *Journal of Consumer Research*, (30):559—573.

[25] Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., and Pidruzny, J. N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, 45(4):624–634.

[26] Brooks, J. L. (2012). Counterbalancing for serial order carryover effects in experimental condition orders. *Psychological Methods*, 17(4):600–614.

[27] Burns, S., editor (2015). *War Gamers' Handbook: A Guide for Professional War Gamers*. U.S. Naval War College Newport, United States.

[28] Cacioppo, J. T. and Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42:116–131.

[29] Calderón, A. and Ruiz, M. (2015). A systematic literature review on serious games evaluation: An application to software project management. *Computers & Education*, 87(C):396–422.

[30] Calvillo-Gámez, E. H., Cairns, P., and Cox, A. L. (2010). Assessing the core elements of the gaming experience. In Bernhaupt, R., editor, *Evaluating user experience in games*, pages 47–71. Springer, London, UK.

[31] Cambria, E., Rajagopal, D., Kwok, K., and Sepulveda, J. (2015). Gecka: Game engine for commonsense knowledge acquisition. In *Twenty-eight Florida Artificial Intelligence Research Society Conference*.

[32] Cambria, E., Xia, Y., and Hussain, A. (2012). Affective common sense knowledge acquisition for sentiment analysis. In *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, page 3580–3585.

[33] Camossi, E. (2013). *A Reasoned Survey of Anomaly Detection Methods for Early Maritime Domain Awareness*. Technical Report JRC80902. European Commission - Joint Research Centre.

[34] Carvalho, M. B., Bellotti, F., Berta, R., De Gloria, A., Sedano, C. I., Baalsrud Hauge, J., Hu, J., and Rauterberg, M. (2015). An activity theory-based model for serious games analysis and conceptual design. *Computers and Education*, 87:166–181.

[35] Cetinkaya, D. and Verbraeck, A. (2011). Metamodeling and model transformations in modeling and simulation. In *Proceedings of the 2011 Winter Simulation Conference (WSC)*, pages 3043–3053.

[36] Chaiken, S., Liberman, A., and Eagly, A. H. (1989). Heuristic and systematic processing within and beyond the persuasion context. In Uleman, J. S. and Bargh, J. A., editors, *Unintended thought*, page 212—252. Guilford Press, New York.

[37] Chaiken, S. and Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology*, 66(3):460—473.

[38] Chin, J. P., Diehl, V. A., and Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of SIGCHI 1988*, pages 213–218, New York. ACM/SIGCHI.

[39] Chklovski, T. (2003). Learner: A system for acquiring commonsense knowledge by analogy. In *Proceedings of the 2nd International Conference on Knowledge Capture*, K-CAP '03, pages 4–12, New York, NY, USA. ACM.

[40] Christensen, J. (1985). The nature of systems development. In *Human Factors Engineering: Engineering Summer Conferences*, Ann Arbor. University of Michigan.

[41] Cobb, B. R. and Shenoy, P. P. (2006). On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–330.

[42] Coleridge, S. T. (1817). *Biographia Literaria*. Rest Fenner, London.

[43] Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popovic, Z., and players, F. (2010). Predicting protein structures with a multiplayer online game. *Nature*, 466:756–760.

[44] Corner, A. and Hahn, U. (2009). Evaluating science arguments: Evidence, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied*, 15(3):199–212.

[45] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley Sons, Inc, second edition.

[46] Crookall, D., Martin, A., Saunders, D., and Coote, A. (1986). Human and computer involvement in simulation. *Simulation & Gaming*, 17:345–375.

[47] Cuzzolin, F. (2014). *Visions of a generalized probability theory*. Lambert Academic Publishers, Saarbrücken, Germany.

[48] Dabrowski, J. J. and de Villiers, J. P. (2015). Maritime piracy situation modelling with dynamic Bayesian networks. *Information Fusion*, 23:116–130.

[49] Darlington, K. (2017). Explainable AI systems: Understanding the decisions of the machines. https://www.bbvaopenmind.com/en/explainable-ai-systems-understanding-the-decisions-of-the-machines/#.WkEZgAJbI-A. twitter retrieved on 10 November 2018.

[50] Das, B. (2004). Generating conditional probabilities for Bayesian networks: Easing the knowledge acquisition problem. *ArXiv*, cs.AI/0411034.

[51] de Campos, C. P. and Ji, Q. (2008). Improving Bayesian network parameter learning using constraints. In *19th International Conference on Pattern Recognition*, pages 1–4.

[52] de Rosa, F., Ben Abdallah, N., Jousselme, A.-L., and Anneken, M. (2018a). Source quality handling in fusion systems: a Bayesian perspective. In *Proceedings of the Maritime Big Data Workshop 2018*, number CMRE-CP-2018-002. NATO-STO Centre for Maritime research and Experimentation.

[53] de Rosa, F. and Jousselme, A.-L. (2018). *Critical review of uncertainty communication standards in support to Maritime Situational Awareness*. Number CMRE-FR-2018-010.

[54] de Rosa, F. and Jousselme, A.-L. (2020). Towards a coherent assessment of Situational Awareness to support system design in the maritime context. In Ahram, T., editor, *Advances in Artificial Intelligence, Software and Systems Engineering*, volume 965, pages 409–420. Springer.

[55] de Rosa, F., Jousselme, A.-L., and De Gloria, A. (2018b). A reliability game for source factors and situational awareness experimentation. *International Journal of Serious Games*, 5(2):45–64.

[56] De Vellis, R. F. (2003). *Scale Development: Theory and Applications*. Sage Publications.

[57] de Waal, A., Koen, H., de Villiers, P., Roodt, H., Moorosi, N., and Pavlin, G. (2016). Construction and evaluation of Bayesian networks with expert-defined latent variables. In *Proceedings of the 19th Int. Conference on Information Fusion*, pages 1158–1165.

[58] Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Artificial Intelligence*, 93(1):1–27.

[59] DeBono, K. G. and Harnish, R. J. (1988). Source expertise, source attractiveness, and the processing of persuasive information: A functional approach. *Journal of Personality and Social Psychology*, (55):541—546.

[60] Dempster, A. (1967). Upper and lower probabilities induced by multivalued mapping. *The Annals of Mathematical Statistics*, 38:325–339.

[61] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

[62] Dennehy, K. (1997). Cranfield - Situation Awareness Scale user manual. Technical Report. College of Aeronautics, Cranfield University, Bedford.

[63] Development, Concepts and Doctrine Centre, UK Ministry of Defence Shrivenham (2013). *Red Teaming Guide*. 2nd edition edition.

[64] Development, Concepts and Doctrine Centre, UK Ministry of Defence Shrivenham (2017). *Wargaming Handbook*.

[65] Digioia, G., Foglietta, C., Oliva, G., Panzieri, S., and Setola, R. (2013). Moving from measuring to understanding: Situation Awareness in homeland security. In *Effective Surveillance for Homeland Security: Balancing Technology and Social Issues*, chapter 10, pages 229–256. CRC Press.

[66] Djaouti, D., Alvarez, J., and Jesse, J.-P. (2011a). Classifying serious games: the G/P/S model. *Handbook of Research on Improving Learning and Motivation through Educational Games: Multidisciplinary Approaches*, 1:118–136.

[67] Djaouti, D., Alvarez, J., Jessel, J.-P., and Rampnoux, O. (2011b). Origins of serious games. In Ma, M., Oikonomou, A., and Jain, L. C., editors, *Serious Games and Edutainment Applications*. Springer, London.

[68] Duda, R. and Shortliffe, E. (1983). Expert systems research. *Science*, 202(4594):261–268.

[69] Duke, R. (1980). A paradigm for game design. *Simulation & Gaming*, 11(3):364–377.

[70] Duke, R. and Geurts, J. (2004). *Policy games for strategic management*. Dutch University Press, Amsterdam, The Netherlands.

[71] Durso, F. T., Hackworth, C. A., Truitt, T., Crutchfield, J., and Manning, C. A. (1998). Situation Awareness as a predictor of performance in en route air traffic controllers. *Air Traffic Quarterly*, (6):1—20.

[72] Ellington, H. I., Addinall, E., and Percival, F. (1982). *A handbook of game design*. Kogan Page, London, England.

[73] Endsley, M. R. (1993). A survey of Situation Awareness requirements in air-to-air combat fighters. *The International Journal of Aviation Psychology*, 3(2):157–168.

[74] Endsley, R. M. (1987). The application of human factors to the development of expert systems for advanced cockpits. In *Human Factors Society 31st Annual Meeting*, pages 1388–139. Human Factor Society, Santa Monica, CA.

[75] Endsley, R. M. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64.

[76] Endsley, R. M. (1996). Automation and Situation Awareness. In *Human factors in transportation. Automation and human performance: Theory and applications*, pages 163–181. Hillsdale, New Jersey USA: Lawrence Erlbaum Associates.

[77] Epstein, S., Pacini, R., Denes-Raj, V., and Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology*, (71):390—405.

[78] European Commission (2014). Communication from the commission to the european parliament and the council better situational awareness by enhanced cooperation across maritime surveillance authorities: next steps within the common information sharing environment for the eu maritime domain com/2014/0451 final. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celexretrieved on 18/10/2017.

[79] Fenton, N. E., Neil, M., and Lagnado, D. A. (2013). A General Structure for Legal Arguments About Evidence Using Bayesian Networks. *Cognitive Science*, 37(1):61–102.

[80] Fischer, Y. (2016). *Wissensbasierte probabilistische Modellierung für die Situationsanalyse am Beispiel der maritimen Überwachung*. PhD thesis, Karlsruhe Institute of Technology.

[81] Friedman, R. (1987). Route analysis of credibility and hearsay. *Yale Law journal*, 96:667–742.

[82] Fu, F.-L., Su, R.-C., and Yu, S.-C. (2009). EGameFlow: A scale to measure learners' enjoyment of e-learning games. *Computer Education*, 52(1):101–112.

[83] Fumarola, M., van Staalduinen, J.-P., and Verbraeck, A. (2011). A ten-step design method for simulation games in logistics management. *Journal of Computing and Information Science in Engineering*, 12(1):150–158.

[84] Gage, J. (1995). A method for measuring subjective probability. *MD Computing*, 12:172–177.

[85] Gavrilova, T. and Andreeva, T. (2012). Knowledge elicitation techniques in a knowledge management context. *Journal of Knowledge Management*, 16(4):523–537.

[86] Geurts, J. L., Duke, R. D., and Vermeulen, P. A. (2007). Policy gaming for strategy and change. *Long Range Planning*, 40(6):535 – 558.

[87] Goldman, A. I. (1999). *Knowledge in a social world*. Oxford University Press.

[88] Graafland, M. and Schijven, M. P. (2013). A serious game to improve Situation Awareness in laparoscopic surgery. In Schouten, B., Fedtke, S., Bekker, T., Schijven, M., and Gekker, A., editors, *Games for Health*, pages 173–182. Springer Fachmedien Wiesbaden.

[89] Greenblat, C. S. and Duke, R. (1975). *Gaming-simulation : rationale, design, and applications*. Sage Publications, New York.

[90] Greenblat, C. S. and Duke, R. D. (1981). *Principles and practice of gaming simulation*. Sage Publications, London.

[91] Habraken, N. J. and Gross, M. D. (1988). Concept design games. *Design Studies*, 9(3):150–158.

[92] Haenni, R. and Hartmann, S. (2006). Modelling partially reliable information sources: a general approach based on Dempster-Shafer theory. *Information Fusion*, 7(4):361–379.

[93] Hahn, U., Oaksfor, M., and Harris, A. J. L. (2013). Testimony and argumentation: A bayesian perspective. In Zenker, F., editor, *Bayesian Argumentation*. Springer Library.

[94] Hall, D. L. and Jordan, J. M. (2010). *Human-centered Information Fusion*. Artech House, Boston.

[95] Hänninen, M., Valdez Banda, O. A., and Kujala, P. (2014). Bayesian network model of maritime safety management. *Expert Systems with Applications*, 41(17):7837–7846.

[96] Harmon, P. and King, D. (1985). *Expert systems*. John Wiley & Sons, Inc, New York, NY.

[97] Harper, B. D. and Norman, K. L. (1993). Improving user satisfaction: The Questionnaire for User Interaction Satisfaction version 5.5. In *Proceedings of the 1st Annual Mid-Atlantic Human Factors Conference*, pages 224–228, Virginia Beach, VA.

[98] Hartmann, T., Wirth, W., Schramm, H., Klimmt, C., Vorderer, P., Gysbers, A., Böcking, S., Ravaja, N., Laari, J., Saari, T., Gouveia, F., and Sacau, A. (2016). The spatial presence experience scale (SPES): A short self-report measure for diverse media settings. *Journal of Media Psychology*, 28:1–15.

[99] Hastie, T., Tibshirani, R., and Friedman, J. (2017). *The elements of statistical learning: data mining, inference and prediction*. Springer.

[100] Haugtvedt, C. P., Petty, R. E., and Cacioppo, J. T. (1992). Need for Cognition and advertising: Understanding the role of personality variables in consumer behavior. *Journal of Consumer Psychology*, (1):239—260.

[101] Hauss, Y., Gauss, B., and Eyferth, K. (2001). SALSA - a new approach to measure Situational Awareness in air traffic control. focusing attention on aviation safety. In *11th International Symposium on Aviation Psychology*, Columbus.

[102] Heesacker, M. H., Petty, R. E., and Cacioppo, J. T. (1983). Field dependence and attitude change: Source credibility can alter persuasion by affecting message-relevant thinking. *Journal of Personality*, (51):653—666.

[103] Herdagdelen, A. and Baroni, M. (2010). The concept game: better commonsense knowledge extraction by combining text mining and game with a purpose. In *Proceedings of the 2010 Commonsense Knowledge Symposium (AAAI CKS)*.

[104] Hogg, D., Folleso, K., Strand-Volden, F., and Torralba, B. (1995). Development of a Situation Awareness measure to evaluate advanced alarm systems in nuclear power plant control room. *Ergonomics*, 38(11):2394–2413.

[105] Hoppenbrouwers, S., Schotten, B., and Lucas, P. (2010). Towards games for knowledge acquisition and modeling. *International Journal of Gaming and Computer-Mediated Simulations*, 2(4):48–66.

[106] Human Performance Research Group (1986). *NASA Task Load Index*. NASA Ames Research Center.

[107] IJsselsteijn, W. A., de Kort, Y. A. W., and Poels, K. (2007). *Game Experience Questionnaire*. Technische Universiteit Eindhoven.

[108] International Organization for Standardization (2011). *Accuracy (trueness and precision) of measurement methods and results — Part 1: Introduction and basic principles*. Number ISO 15725.

[109] International Standard Organisation (2014). Systems and software engineering – systems and software quality requirements and evaluation (SQuaRE) – system and software quality models.

[110] Jeannott, E., Kelly, C., and Thompson, D. (2003). *The Development of Situation Awareness Measures in ATM Systems*. EATMP.

[111] Joint Systems Analysis (JSA) Group, Methods and Approaches for Warfighting Experimentation Action Group 12 (AG-12) (2006). Guide for understanding and implementing defence experimentation (GUIDEx).

[112] Jones, K. (1998). Simulations: Reading for action. *Simulation & Gaming*, 29:326–327.

[113] Jousselme, A.-L., Pallotta, G., and Locke, J. (2018). Risk game: Capturing impact of information quality on human belief assessment and decision making. *International Journal of Serious Games*, 5(4):23–44.

[114] Kass, R. A. (2006). *The Logic of Warfighting Experiments*. CCRP Press, Washington, DC.

[115] Keller, J. M. (1987). Development and use of the ARCS model of instructional design journal of instructional development. *Journal of instructional development*, 10(2).

[116] Kelman, H. C. and Hovland, C. I. (1953). Reinstatement of the communicator in delayed measurement of opinion change. *Journal of Abnormal and Social Psychology*, (48):327—335.

[117] Kemp-Benedict, E. (2008). *Elicitation techniques for Bayesian network models*. Number Working Paper WP-US-0804. Stockholm, Sweden: Stockholm Environment Institute.

[118] Khenissi, M. A., Essalmi, F., and Jemni, M. (2015). Comparison between serious games and learning version of existing games. *Procedia - Social and Behavioral Sciences*, 191:487–494.

[119] Klabbers, J. H. G. (2009). *The Magic Circle: Principles of Gaming & Simulation*. Sense Publishers, third and revised edition.

[120] Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

[121] Kondreddi, S. K., Triantafillou, P., and Weikum, G. (2013). Human computing games for knowledge acquisition. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, CIKM '13, pages 2513–2516, New York, NY, USA. ACM.

[122] Korb, K. B. and Nicholson, A. E. (2010). *Bayesian Artificial Intelligence*. Chapman & Hall, 2nd edition.

[123] Kosmadoudi, Z., Lim, T., Ritchie, J., Louchart, S., Liu, Y., and Sung, R. (2013). Engineering design using game-enhanced cad: the potential to augment the user experience with game elements. *Computer-Aided Design*, 45:777–795.

[124] Kriz, W. C. (2003). Creating effective learning environments and learning organizations through gaming simulation design. *Simulation & Gaming*, 34(4):495–511.

[125] Kriz, W. C. (2017). Types of gaming simulation applications. *Simulation & Gaming*, 48(1):3–7.

[126] Kruglanski, A. W. and Thompson, E. P. (1999). Persuasion by a single route: A view from the unimodel. *Psychological Inquiry*, (10):83—110.

[127] Krüger, M., Ziegler, J., and Heller, K. (2012). A generic Bayesian network for identification and assessment of objects in maritime surveillance. In *Proceedings of the 15th International Conference on Information Fusion (FUSION)*, pages 2309–2316.

[128] Kuo, Y.-l., Lee, J.-C., Chiang, K.-Y., Wang, R., Shen, E., Chan, C.-W., and Hsu, J. Y.-J. (2009). Community-based game design: Experiments on social games for commonsense data collection. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pages 15–22, New York, NY, USA. ACM.

[129] Kurapati, S., Kourounioti, I., Lukosch, H., Tavasszy, L., and Verbraeck, A. (2018). Fostering sustainable transportation operations through corridor management: A simulation gaming approach. *Sustainability*, 10(2):364–377.

[130] Laamarti, F., Eid, M., and Saddik, A. E. (2014). An overview of serious games. *International Journal of Computer Games Technology*, 2014.

[131] Lacy, L., Randolph, W., Harris, B., Youngblood, S., Sheehan, J., Might, R., and Metz, M. (2001). Developing a consensus perspective on conceptual models for simulation systems. In *Proceedings of the 2001 Spring Simulation Interoperability Workshop*.

[132] Lagnado, D. A., Fenton, N. E., and Neil, M. (2013). Legal idioms: a framework for evidential reasoning. *Argument & Computation*, 4(1):46–63.

[133] Lasswell, H. (1948). *The Structure and Function of Communication in Society*. The Communication of Ideas. Harper and Brothers, New York.

[134] Law, E., von Ahn, L., Dannenberg, R., and Crawford, M. (2007). Tagatune: A game for music and sound annotation. In *International Conference on Music Information Retrieval*, page 361–364.

[135] Lukosch, H., Groen, D., Kurapati, S., Klemke, R., and Verbraeck, A. (2016a). The role of awareness for complex planning task performance: A microgaming study. *International Journal of Game-Based Learning*, 6(2):15–28.

[136] Lukosch, H., Kurapati, S., Groen, D., and Verbraeck, A. (2016b). Microgames for situated learning: A case study in interdependent planning. *Simulation & Gaming*, 47(3):1–22.

[137] Lukosch, H. K., Bekebrede, G., Kurapati, S., and Lukosch, S. G. (2018). A scientific foundation of simulation games for the analysis and design of complex systems. *Simulation & Gaming*, 49(3):279–314.

[138] Ma, H., Chandrasekar, R., Quirk, C., and Gupta, A. (2009). Improving search engines using human computation games. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2009, page 746–747.

[139] Margarit, G. and Nunes, A. (2012). *NEREIDS D.440.2 - Simulation element definition for anomal analysis*. Technical Report.

[140] Markotschi, T. and Völker, J. (2010). Guess What?! Human intelligence for mining linked data. In *Proceedings of the Workshop on Knowledge Injection into and Extraction from Linked Data (KIELD) at the International Conference on Knowledge Engineering and Knowledge Management (EKAW)*.

[141] Marshev, V. and Popov, A. (1983). Element of a theory of gaming. In Ståhl, I., editor, *Operational Gaming*, Frontiers of Operational Research and Applied Systems Analysis. Pergamon Press.

[142] Marusich, L. R., Jonathan, Z., Bakdash, J. Z., Onal, E., Yu, M. S., Schaffer, J., O'Donovan, J., Höllerer, T., Buchler, N., and Gonzalez, C. (2016). Effects of information availability on command-and-control decision making: Performance, trust, and Situation Awareness. *Human Factors*, 58(2):301 – 321.

[143] Matthews, M. D., Pleban, R. J., Endsley, M. R., and Strater, L. D. (2000). Measures of infantry Situation Awareness for a virtual MOUT environment. In *Human Performance, Situation Awareness and Automation Conference (HPSAA II)*, Daytona.

[144] McGuinness, B. and Foy, L. (2000). A subjective measure of SA: the Crew Awareness Rating Scale (CARS). In *Human Performance, Situational Awareness and Automation Conference*, Savannah.

[145] McVay, J. C. and Kane, M. J. (2009). Conducting the train of thought: Working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology, learning, Memory, and Cognition*, 35(1):196–204.

[146] Meyer, M. A. and Booker, J. M. (2001). *Eliciting and Analyzing Expert Judgment: A Practical Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

[147] Michael, D. and Chen, S. (2006). *Serious games: Games that educate, train, and inform*. Thomson Course Technology PTR, Boston,MA.

[148] Millar, A. (2012). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 101(2):343–352.

[149] Moneta, G. B. and Csikszentmihalyi, M. (1996). The effect of perceived challenges and skills on the quality of subjective experience. *Journal of Personality*, 64(2):275–310.

[150] Moneta, G. B. and Csikszentmihalyi, M. (1999). Models of concentration in natural environments: A comparative approach based on streams of experiential data. *Social Behavior and Personality: An international journal*, 27:603–638.

[151] Morrison, D., Marchand-Maillet, S., and Bruno, E. (2010). Tagcaptcha: Annotating images with captchas. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 1557–1558, New York, NY, USA. ACM.

[152] Mouafo, S. R. T., Vaton, S., Courant, J.-L., and Gosselin, S. (2016). A tutorial on the EM algorithm for Bayesian networks: application to self-diagnosis of GPON-FTTH networks. In *IWCMC 2016 : 12th International Wireless Communications - Mobile Computing Conference*, pages 369 – 376.

[153] Mrad, A., Delcroix, V., Piechowiak, S., Leicester, P., and Abid, M. (2015). An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence. *Applied Intelligence*, 43(4):802–824.

[154] Murphy, K. P. (2002). Dynamic Bayesian networks: Representation, inference and learning.

[155] Nacke, L., Drachen, A., and Goebel, S. (2010). Methods for evaluating gameplay experience in a serious gaming context. *International Journal of Computer Science in Sport*, 9(2 / Special Issue).

[156] Nacke, L., Drachen, A., Kuikkaniemi, K., Niesenhaus, J., Korhonen, H., Hoogen, van den, W., Poels, K., IJsselsteijn, W., and Kort, de, Y. (2009). Playability and player experience research. In Atkins, B. and Kennedy, H., editors, *Breaking new ground : innovation in games, play, practice and theory*. DiGRA.

[157] Nacke, L. E., Bateman, C., and Mandryk, R. L. (2014). BrainHex: A neurobiological gamer typology survey. *Entertainment Computing*, 5(1):55–62.

[158] Nance, R. E. (1994). The conical methodology and the evolution of simulation model development. *Annals of Operations Research*, 53(1):1–45.

[159] Nemeth, C. P. (2004). *Human Factors Methods for Design: Making Systems Human-Centered*. CRC Press, Boca Raton.

[160] Netica (2018). Norsys software corp. https://www.norsys.com/index.html.

[161] Norman, K. L. (2013). GEQ (Game Engagement/Experience Questionnaire): A review of two papers. *Interacting with Computers*, 25(4):278–283.

[162] Osborne, J. W. and Blanchard, M. R. (2010). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology*, 1.

[163] Perla, P. (1990). *The Art of Wargaming: A Guide to Professionals and Hobbiests*. Naval institute Press, Annapolis, Md.

[164] Perla, P. and McGrady, E. (2011). Why Wargaming Works. *Naval War College Review*, 64(3).

[165] Peters, V., Vissers, G., and Heijne, G. (1998). The validity of games. *Simulation & Gaming*, 29(1):20–30.

[166] Peters, V. and Westelaken, M. (2014). *Simulation Games — A concise introduction to game design*. Samenspraak Advies, Nijmegen, The Nethernalds.

[167] Petri, G., Gresse von Wangenheim, C., and Borgatto, A. F. (2016). *MEEGA+:An evolution of a Model for the Evaluation of Educational Games*. INCoD/GQS.05.2018.E. INCoD.

[168] Petri, G., Gresse von Wangenheim, C., and Borgatto, A. F. (2019). MEEGA+, systematic model to evaluate educational games. In Lee, N., editor, *Encyclopedia of Computer Graphics and Games*. Springer.

[169] Petri, G. and von Wangenheim, C. G. (2016). How to evaluate educational games: a systematic review. *Journal of Universal Computers Science*, 22(7):992–1021.

[170] Petri, G., von Wangenheim, C. G., and Borgatto, A. F. (2017). A large-scale evaluation of a model for the evaluation of games for teaching software engineering. In *Proceedings of the 39th International Conference on Software Engineering: Software Engineering and Education Track*, ICSE-SEET '17, pages 180–189, Piscataway, NJ, USA. IEEE Press.

[171] Petty, R. E. and Cacioppo, J. T. (1984). Source factors and the elaboration likelihood model of persuasion. *Advances in Consumer Research*, (11):668—672.

[172] Petty, R. E., Cacioppo, J. T., and Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, (41):847—855.

[173] Pew, R. (1985). Human skills and their utilization. In *Human Factors Engineering: Engineering Summer Conferences*. University of Michigan, Ann Arbor.

[174] Pichon, F., Dubois, D., and Denoeux, T. (2012). Relevance and truthfulness in information correction and fusion. *International Journal of Approximate Reasoning*, 53:159–175.

[175] Pichon, F., Mercier, D., Lefèvre, E., and Delmotte, F. (2016). Proposition and learning of some belief function contextual correction mechanisms. *International Journal of Approximate Reasoning*, 72:4–42.

[176] Pilato, G., Augello, A., Missikoff, M., and Taglino, F. (2012). Integration of ontologies and Bayesian networks for maritime situation awareness. In *Proceedings of the IEEE Sixth International Conference on Semantic Computing (ICSC)*, pages 170–177.

[177] Poels, K., de Kort, Y., and IJsselsteijn, W. (2007). *D3.3 : Game Experience Questionnaire: development of a self-report measure to assess the psychological impact of digital games*. Technische Universiteit Eindhoven.

[178] Polski, M. M. (2019). Back to basics—research design for the operational level of war. *Naval War College Review*, 72(3).

[179] Price, P. (1998). Effects of a relative-frequency elicitation question on likelihood judgement accuracy: The case of external correspondence. *Organizational Behavior and Human Decision Processes*, 76:277–297.

[180] Priester, J. R. and Petty, R. E. (1995). Source attributions and persuasion: Perceived honesty as a determinant of message scrutiny. *Personality and Social Psychology Bulletin*, 21(6):637—654.

[181] Raser, J. R. (1969). Simulation and society: An exploration of scientific gaming. In *Methodology in the Behavioral Sciences Series*. Allyn and Bacon, Boston.

[182] Rein, K. and Biermann, J. (2013). Your high-level information is my low-level data. a new look at terminology for multi-level fusion. In *16th International Conference on Information Fusion*, Istanbul, Turkey.

[183] Ren, Y., Bayrak, A. E., and Papalambros, P. Y. (2016). Ecoracer: game-based optimal electric vehicle design and driver control using human players. *Journal of Mechanical Design*, 138(6).

[184] Renooij, S. (2001). Probability elicitation for belief networks: issues to consider. *The Knowledge Engineering Review*, 16:255–269.

[185] Renooij, S. and Witteman, C. (1999). Probability elicitation for belief networks: issues to consider. *International Journal of Approximate Reasoning*, 22:169–194.

[186] Rhine, R. J. and Severance, L. J. (1970). Ego-involvement, discrepancy, source credibility, and attitude change. *Journal of Personality and Social Psychology*, 16(2):175—190.

[187] Robinson, S. (2008a). Conceptual modelling for simulation Part I: definition and requirements. *Journal of the Operational Research Society*, 59(3):278–290.

[188] Robinson, S. (2008b). Conceptual modelling for simulation Part II: a framework for conceptual modelling. *Journal of the Operational Research Society*, 59(3):291–304.

[189] Rogova, G. L. and Nimier, V. (2004). Reliability in information fusion: literature survey. In *7th Int. Conference on Information Fusion*, pages 1158–1165.

[190] Roungas, B. (2016). A model-driven framework for educational game design. *International Journal of Serious Games*, 3(3):19–37.

[191] Rubel, R. C. (2006). The epistemology of war gaming. *Naval War College Review*, 59(2).

[192] Ruhnke, V. (2018). Wargames and systems thinking. In *Connections UK*.

[193] Sargut, G. and McGrath, R. G. (2011). Learning to live with complexity. *Harvard Business Review*, 89(9):68–76.

[194] Sarter, N. and Woods, D. (1991). Situation awareness: a critical but ill-defined phenomenon. *International Journal of Aviation Psychology*, 1:45–57.

[195] Savi, R., Wangenheim, C. G. V., and Borgatto, A. F. (2011). A model for the evaluation of educational games for teaching software engineering. In *2011 25th Brazilian Symposium on Software Engineering*, pages 194–203.

[196] Sawaragi, T., Fujii, K., Horiguchi, Y., and Nakanishi, H. (2016). Analysis of team Situation Awareness using serious game and constructive model-based simulation. In *13th IFAC Symposium on Analysis, Design, and Evaluation of Human-Machine Systems*, volume 49, pages 537–542. Elsevier.

[197] Sawyer, B. and Rejeski, D. (2002). *Serious Games: Improving Public Policy Through Game-based Learning and Simulation*. Woodrow Wilson International Center for Scholars.

[198] Schmalhofer, F. (2001). Expert systems in cognitive science. In *International Encyclopedia of the Social & Behavioral Sciences*.

[199] Schramm, W. L. (1954). *The process and effects of mass communication*. University of Illinois Press.

[200] Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., Van de Velde, W., and Wielinga, B. (2000). *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press, Cambridge, MA.

[201] SCS Technical Committee on Model Credibility (1979). Terminology for model credibility. *Simulation*, 32(3):103–104.

[202] Seixas, F. L., Zadrozny, B., Laks, J., Conci, A., and Muchaluat Saade, D. C. (2014). A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment. *Computers in Biology and Medicine*, 51:140–158.

[203] Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company, Boston.

[204] Shafer, G. (1976). *A Mathematical Theory of Evidence*. University Press.

[205] Shang, Y. (2001). Expert systems. In *The Electrical Engineering Handbook*.

[206] Shannon, C. E. and Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.

[207] Shubik, M. (1983). Gaming: a state-of-the-art survey. In Ståhl, I., editor, *Operational Gaming*, Frontiers of Operational Research and Applied Systems Analysis. Pergamon Press, Oxford.

[208] Siorpaes, K. and Hepp, M. (2008). Games with a purpose for the semantic web. *IEEE Intelligent Systems*, 23(3):50–60.

[209] Smets, P. (1990). Constructing the pignistic probability function in a context of uncertainty. *Uncertainty in Artificial Intelligence*, 5(3):29–39.

[210] Smets, P. and Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, 66:191–234.

[211] Spetzler, C. S. and Stael von Holstein, C. A. (1975). Probability encoding in decision analysis. *Management Science*, 22:340–358.

[212] Stanton, N. A., Salmon, P. M., Walker, G. H., Baber, C., and Jenkins, D. P. (2006). *Human Factors Methods: A Practical Guide for Engineering And Design*. Ashgate Publishing Company, Brookfield.

[213] Stephan, J. and Brockner, J. (2007). Spaced out in cyberspace?: Evaluations of computer-based information. *Journal of Applied Social Psychology*, (37):210—226.

[214] Sternthal, B., Dholakia, R., and Leavitt, C. (1978). The persuasive effect of source credibility: Tests of cognitive response. *Journal of Consumer Research*, 4(4):252—260.

[215] Stikeleather, J. (2012). Big data's human component. https://hbr.org/2012/09/big-datas-human-component retrieved on 10 October 2017.

[216] Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1-2):161–197.

[217] Ståhl, I. (1983). What is operational gaming? In Ståhl, I., editor, *Operational Gaming*, Frontiers of Operational Research and Applied Systems Analysis. Pergamon Press, Oxford.

[218] Sullivan, C. and Blackman, H. (1991). Insights into pilot situation awareness using verbal protocol analysis. In *Human Factors Society 35th Annual Meeting*, page 57–61. Human Factors Society.

[219] Sweetser, P. and Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. *Computers in Entertainment*, 3(3):3–3.

[220] Taylor, R. M. (1990). Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design. In *Situational Awareness in Aerospace Operations*, page 3/1—3/17. Army Research Institute for the Behavioural and Social Sciences, Neuilly Sur Seine.

[221] Teach, R. D. (1990). Designing business simulations. In Gentry, J. W., editor, *Guide to business gaming and experiential learning*. Nichols/GP, East Brunswick, NJ/London.

[222] Thaler, S., S. E. and Siorpaes, K. (2011). Spotthelink: a game for ontology alignment. In *6th Conference for Professional Knowledge Management*, WM2011.

[223] Thavikulwat, P. (2004). The architecture of computerized business gaming simulations. *Simulation & Gaming*, 35(2):242–269.

[224] United States Army War College (2015). *Strategic Wargaming Series Handbook*.

[225] Van den Broeck, G., Mohan, K., Choi, A., Darwiche, A., and Pearl, J. (2015). Efficient algorithms for Bayesian network parameter learning from incomplete data. In *31st Conference on Uncertainty in Artificial Intelligence*, pages 161–170.

[226] van den Hoogen, J. and Meijer, S. (1980). Gaming and simulation for railway innovation: a case study of the dutch railway system. *Simulation & Gaming*, 46(5):489–511.

[227] van der Gaag, L., Renooij, S., Witteman, C., Aleman, B., and Taal, B. (1999). How to elicit many probabilities. In *Proceedings of the Fifteen conference on Uncertainty in Artificial Intelligence*, pages 647–654.

[228] van der Zee, D., Holkenborg, B., and Robinson, S. (2012). Conceptual modeling for simulation-based serious gaming. *Decision Support Systems*, 54(1):33–45.

[229] van Harmelen, F., Lifschitz, V., and Porter, B. W., editors (2007). *Handbook of Knowledge Representation*. Foundations of Artificial Intelligence. Elsevier, 3 edition.

[230] Vermillion, S., Malak, R., Smallman, R., Becker, B., Sferra, M., and Fields, S. (2017). An investigation on using serious gaming to study human decision-making in engineering contexts. *Design Science*, 3(E15).

[231] Vila-Francés, J., Sanchís, J., Soria-Olivas, E., Serrano, A. J., Martínez-Sober, M., Bonanad, C., and Ventura, S. (2013). Expert system for predicting unstable angina based on Bayesian networks. *Expert Systems with Applications*, 40(12):5004–5010.

[232] Vissers, G., Heyne, G., Peters, V., and Guerts, J. (2001). The validity of games. *Quality & Quantity*, 35:129–145.

[233] von Ahn, L. (2006). Games with a purpose. *Computer*, 39(6):92–94.

[234] von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *CHI Conference on Human Factors in Computing Systems*, page 319–326.

[235] von Ahn, L., Ginosar, S., Kedia, M., Liu, R., and Blum, M. (2006a). Improving accessibility of the web with a computer game. In *CHI Conference on Human Factors in Computing Systems*, page 79–82.

[236] von Ahn, L., Liu, R., and Blum, M. (2006b). Peekaboom: A game for locating objects in images. In *CHI Conference on Human Factors in Computing Systems*, page 55–64.

[237] von Neumann, J. and Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press, Princeton, NJ, US.

[238] Vorderer, P., Wirth, W., Ribeiro Gouveia, F., Biocca, F., Saari, T., Jäncke, L., Böcking, S., Schramm, H., Gysbers, A., Hartmann, T., Klimmt, C., Laarni, J., Ravaja, N., Sacau, A., Baumgartner, T., and Jäncke, P. (2004). Development of the MEC spatial presence questionnaire (MEC-SPQ). Technical Report MEC (IST-2001-37661).

[239] Waag, W. L. and Houck, M. R. (1994). Tools for assessing Situational Awareness in an operational fighter environment. *Aviation, Space and Environmental Medicine*, 65(5):A13—A19.

[240] Wagner, G. (2018). Information and process modeling for simulation – Part I. *Journal of Simulation Engineering*, 1.

[241] Wallsten, T.S., B. D. and Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgements. *Management Science*, 39:176–190.

[242] Wang, H., Dahs, D., and Druzdel, M. (2002). A method for evaluating elicitation schemes for probabilistic models. *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics*, 32:38–43.

[243] Wang, Y., Baciu, G., Yao, Y., Kinsner, W., Chan, K. C. C., Zhang, B., Hameroff, S. R., Zhong, N., Huang, C.-R., Goertzel, B., Miao, D., Sugawara, K., Wang, G., You, J., Zhang, D., and Zhu, H. (2012). Perspectives on cognitive informatics and cognitive computing. *International Journal of Cognitive Informatics and Natural Intelligence*, 4:1–24.

[244] Wang, Y., Kinsner, W., Anderson, J. A., Zhang, D., Yao, Y., Sheu, P., Tsai, J., Pedrycz, W., Latombe, J.-C., Zadeh, L. A., Patel, D., and Chan, C. (2009). A doctrine of Cognitive Informatics (CI). *Fundamenta Informaticae - Cognitive Informatics, Cognitive Computing, and Their Denotational Mathematical Foundations*, 90(3):203–228.

[245] Whitcomb, K. M., Onkal, D., Benson, P. G., and Curley, S. P. (1993). An evaluation of the reliability of probability judgements across response modes and over time. *Journal of Behavioral Decision Making*, 6:283–296.

[246] Wiebe, E. N., Lamb, A., Hardy, M., and Sharek, D. (2014). Measuring engagement in video game-based environments. *Computers in Human Behavior*, 32(C):123–132.

[247] Wiemeyer, J., Nacke, L., Moser, C., and 'Floyd' Mueller, F. (2016). Player experience. In Dörner, R., Göbel, S., Effelsberg, W., and Wiemeyer, J., editors, *Serious Games: Foundations, Concepts and Practice*, pages 243–271. Springer International Publishing, Cham.

[248] Wiener, E. L. (1988). Cockpit automation. In *Human Factors in Aviation (Cognition and Perception)*, chapter 13, pages 433–461.

[249] Wilkinson, P. (2016). A brief history of serious games. In *Entertainment computing and serious games*, pages 17–41. Springer, Cham.

[250] Wilson, D. W., Jenkins, J., Twyman, N., Jensen, M., Valacich, J., Dunbar, N., Wilson, S., Miller, C., Adame, B., Lee, Y.-H., Burgoon, J., and Nunamaker, J. F. (2016). Serious games: An evaluation framework and case study. In *49th Hawaii International Conference on System Sciences (HICSS)*.

[251] Winkler, R. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, 62:776–800.

[252] Wirth, W., Hartmann, T., Böcking, S., Vorderer, P., Klimmt, C., Schramm, H., Saari, T., Laarni, J., Ravaja, N., Gouveia, F. R., Biocca, F., Sacau, A., Jäncke, L., Baumgartner, T., and Jäncke, P. (2007). A process model of the formation of spatial presence experiences. *Media Psychology*, 9(3):493–525.

[253] Wisse, B., Van Gosliga, S., Van Elst, N., and Barros, A. (2008). Relieving the elicitation burden of Bayesian belief networks. In *Proceedings of the Sixth UAI Bayesian Modelling Applications Workshop*.

[254] Witteman, C. and Renooij, S. (2003). Evaluation of a verbal-numerical probability scale. *International Journal of Approximate Reasoning*, 33:117–131.

[255] Wright, G. and Ayton, P. (1988). Decision time, subjective probability, and task difficulty. *Memory and Cognition*, 16:176–185.

[256] Yan, J. and Yu, S.-Y. (2009). Magic bullet: a dual-purpose computer game. In *ACM SIGKDD Workshop on Human Computation*, page 32–33.

[257] Yankova, K. (2015). *The Influence of Information Order Effects and Trait Professional Skepticism on Auditors' Belief Revisions*. Gabler Verlag.

[258] Ye, J., Dobson, S., and McKeever, S. (2012). Situation identification techniques in pervasive computing: a review. *Pervasive and Mobile Computing*, 8(1):36–66.

[259] Zeigler, B. P. (1976). *Theory of Modeling and Simulation*. Wiley, New York.

[260] Zhou, Y., Fenton, N., and Neil, M. (2014). Bayesian network approach to multinomial parameter learning using data and expert judgments. *International Journal of Approximate Reasoning*, 55(4):1252–1268.

[261] Zhou, Y., Hospedales, T. M., and Fenton, N. (2016). When and where to transfer for Bayes net parameter learning. *Expert systems with applications*, 55:361–373.

[262] Zocholl, M., Iphar, C., Dréo, R., Camossi, E., de Rosa, F., Jousselme, A.-L., and Ray, C. (2019). User centric assessment of maritime Situation Awareness solutions. In *OCEANS 2019 MTS/IEEE*.

Appendix A

MARISA Game K2AGQ results

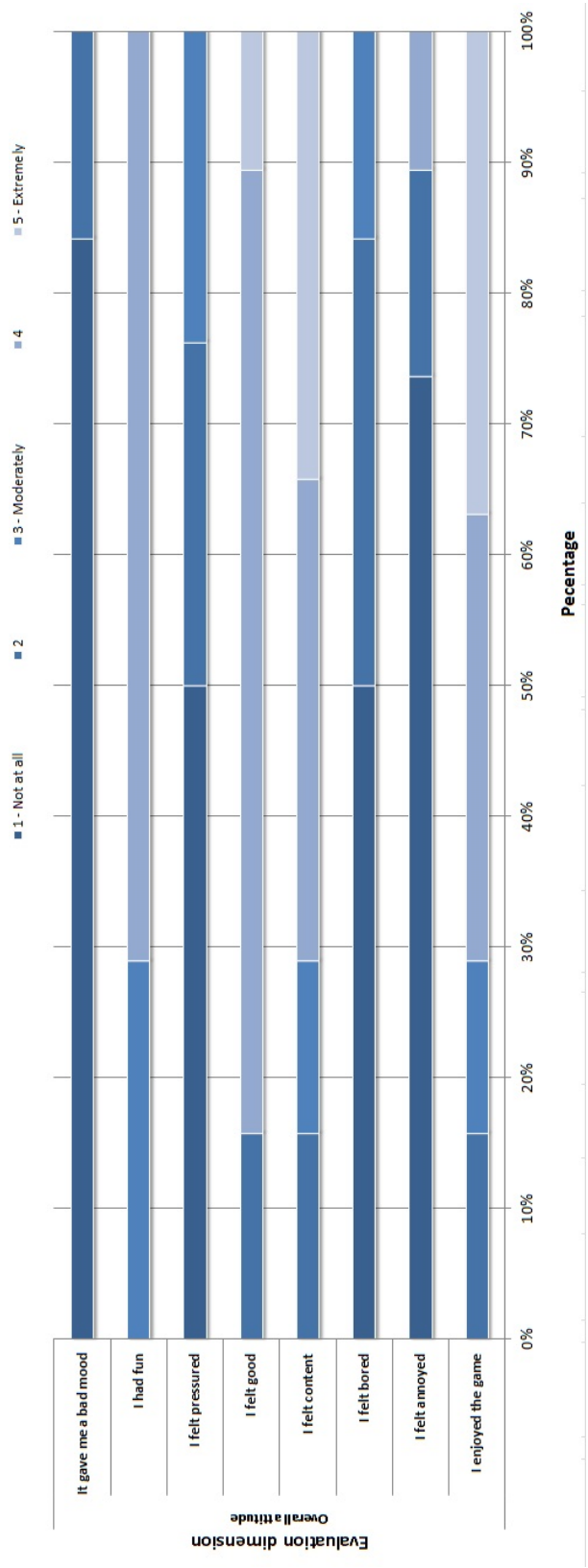


Fig. A.1 MARISA K2AGQ - attitude

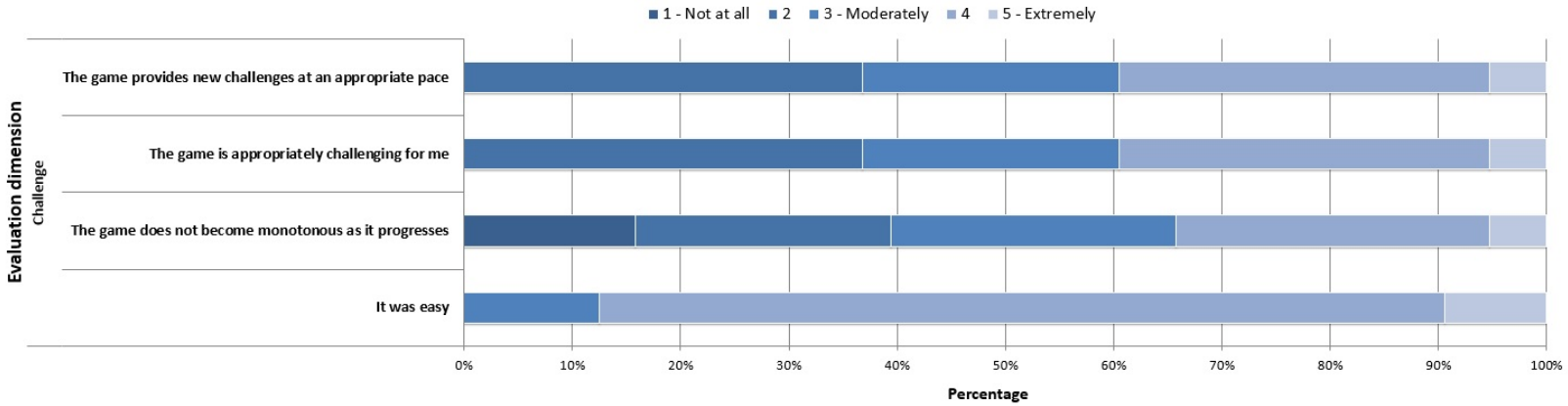


Fig. A.2 MARISA K2AGQ - challenge

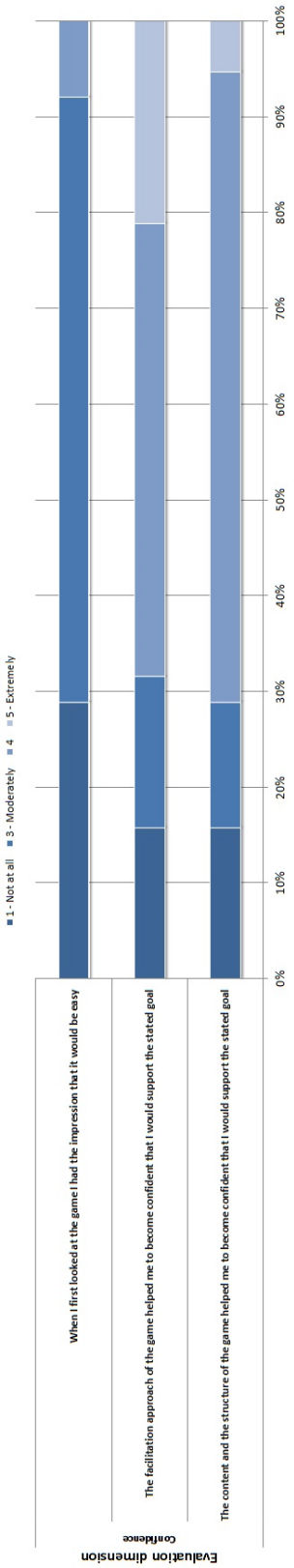


Fig. A.3 MARISA K2AGQ - confidence

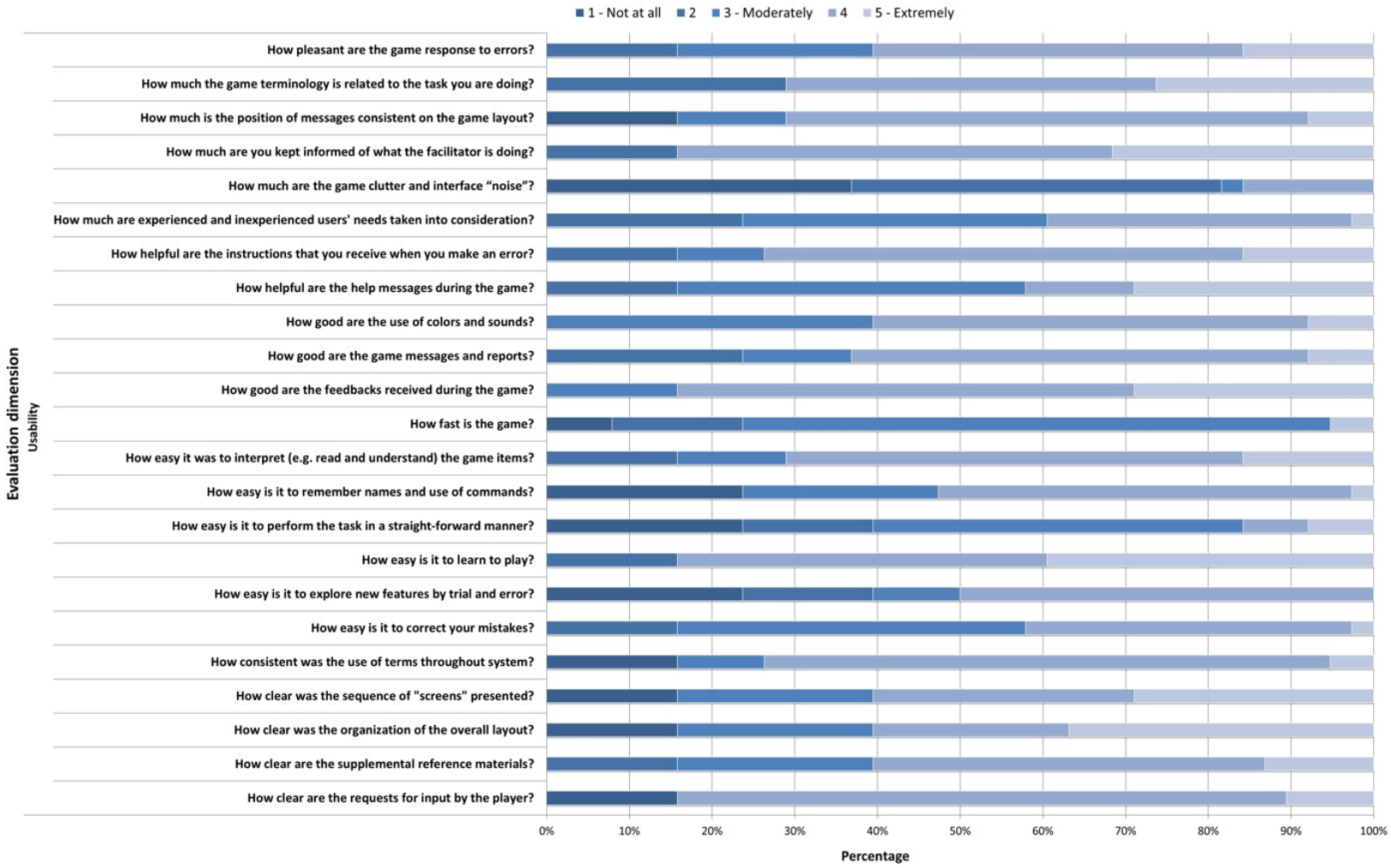


Fig. A.4 MARISA K2AGQ - usability

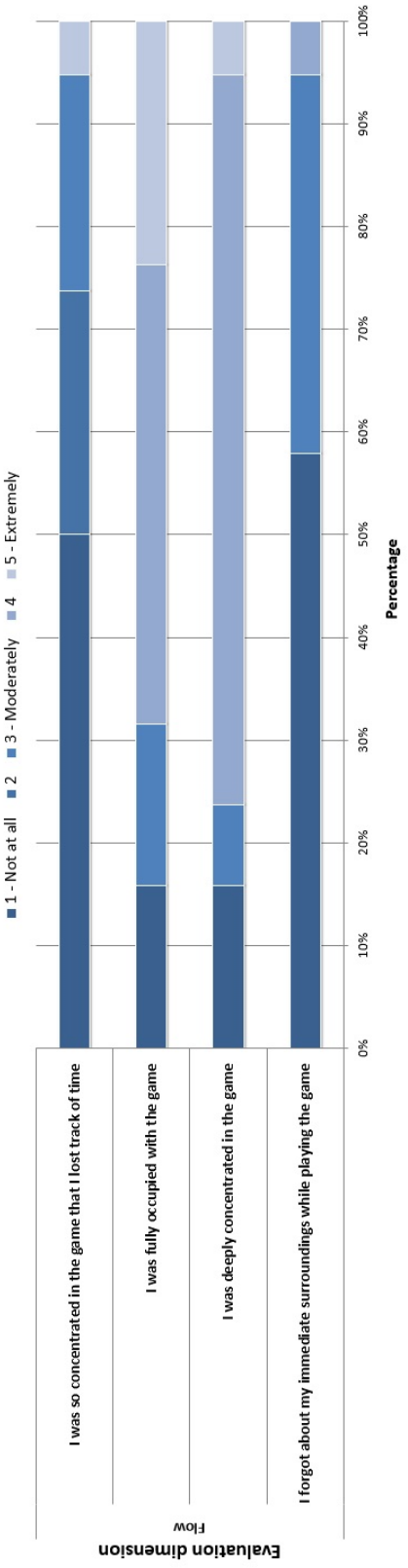


Fig. A.5 MARISA K2AGQ - flow

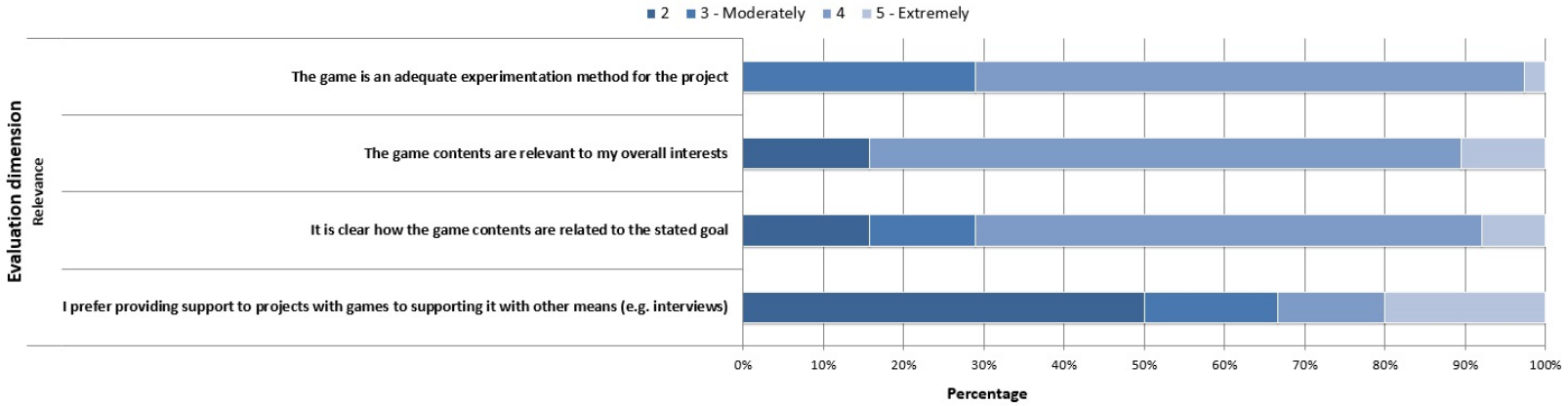


Fig. A.6 MARISA K2AGQ - relevance

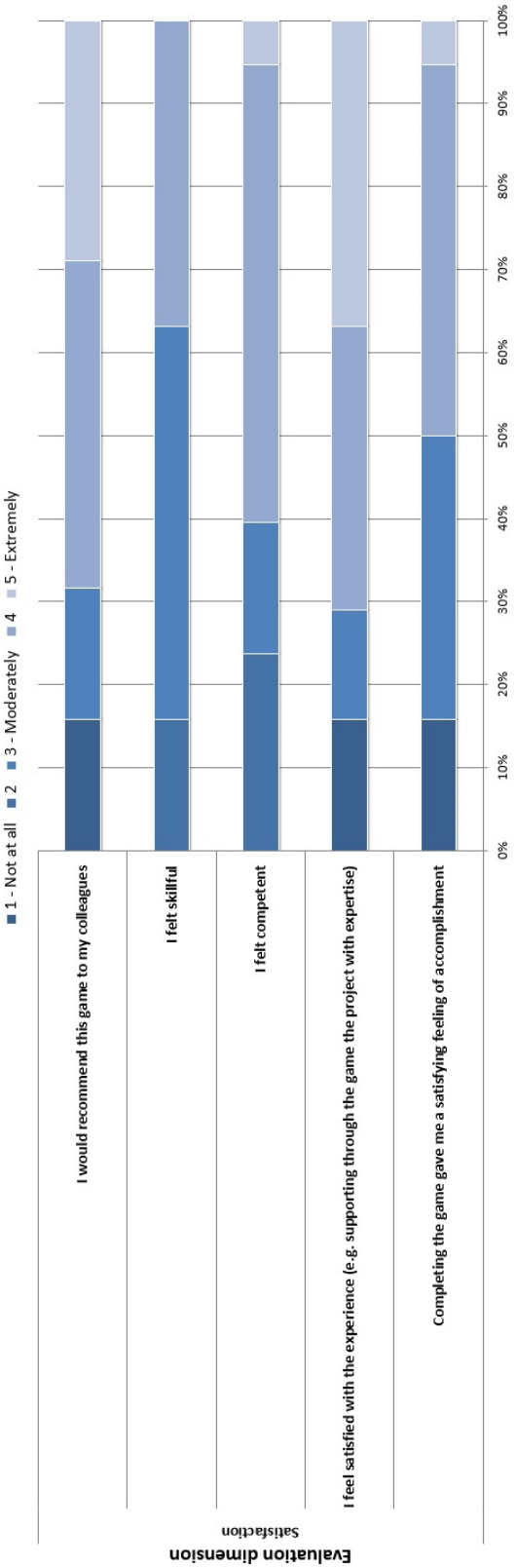


Fig. A.7 MARISA K2AGQ - satisfaction

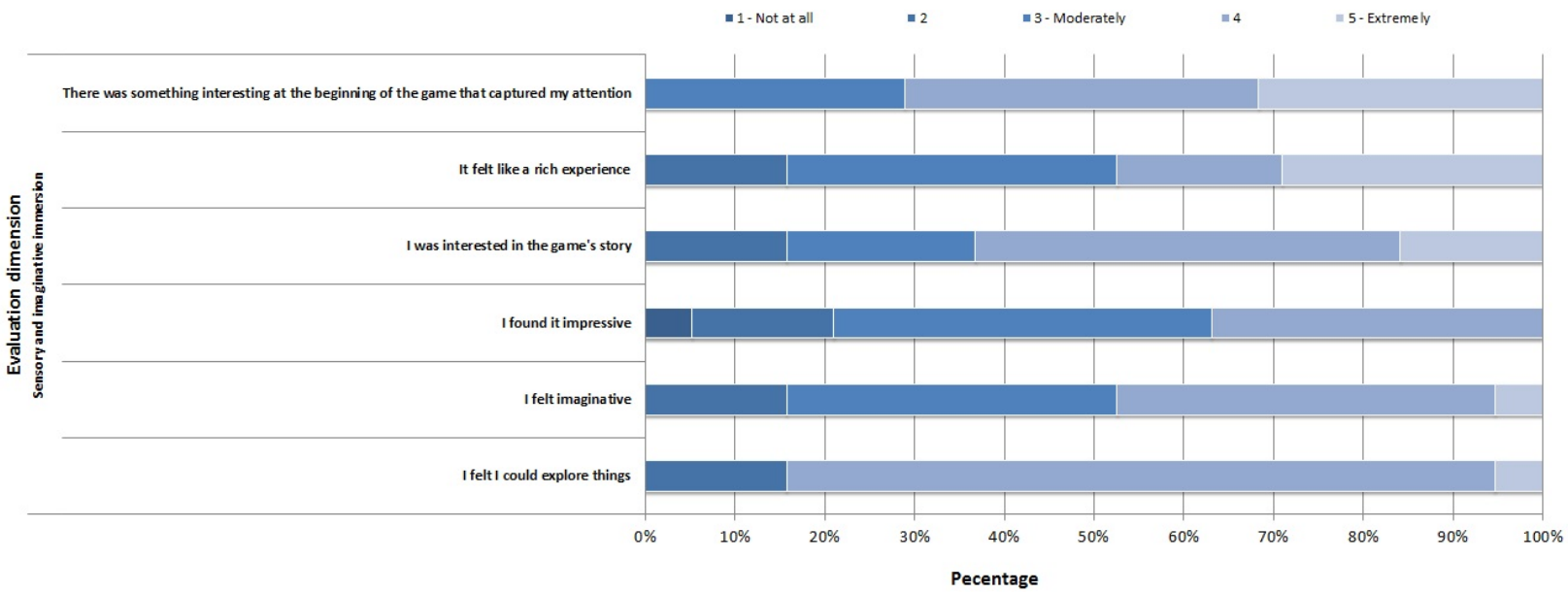


Fig. A.8 MARISA K2AGQ - sensory and imaginative immersion

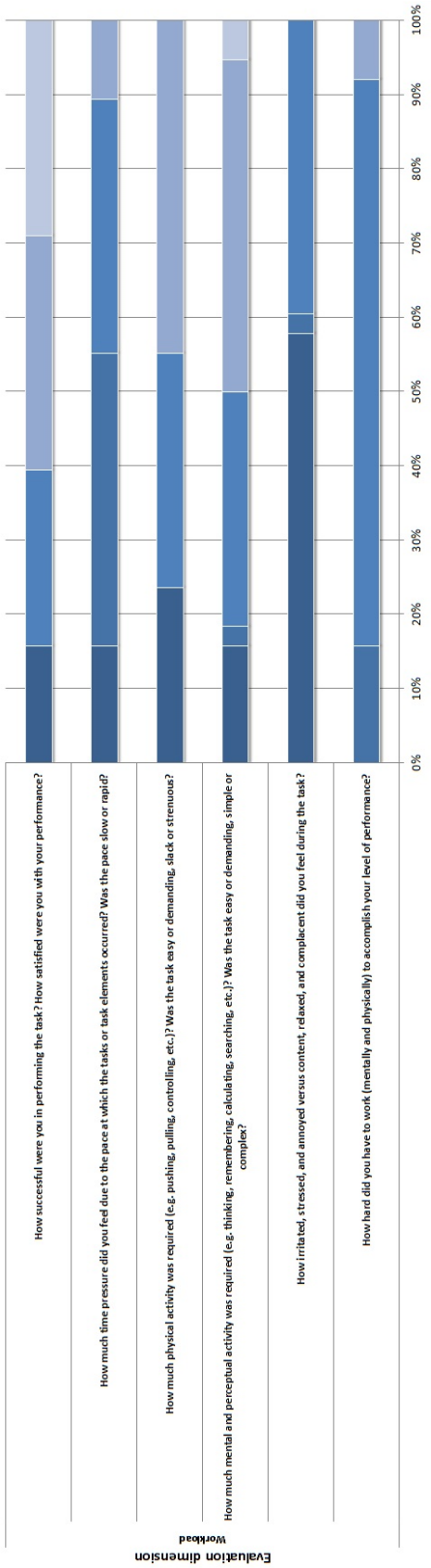


Fig. A.9 MARISA K2AGQ - workload