Mamia Ori-otse Agbese

IMPLEMENTING ARTIFICIAL INTELLIGENCE ETHICS IN TRUSTWORTHY SYSTEMS DEVELOPMENT: EXTENDING ECCOLA TO COVER INFORMATION GOVERNANCE PRINCIPLES

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION TECHNOLOGY
2021

ABSTRACT

Agbese, Mamia Ori-otse
Implementing AI ethics in trustworthy system development: Extending ECCOLA to cover Information Governance principles
Jyväskylä: University of Jyväskylä, 2021, 100pp.
Information Systems, Master's Thesis
Supervisor: Abrahamsson, Pekka

This Master's thesis assesses how to extend a higher-level developmental method for trustworthy artificial intelligent systems, ECCOLA, by evaluating it with Information Governance principles. Artificial intelligent systems are ubiquitous, with their application prevalent in virtually all sectors. In addition, Artificial intelligent systems rely on data and information they collect from users for their development. These issues have prompted ethical concerns, especially as their usage crosses boundaries in sensitive areas such as health, transportation, and security, calling for better governance. As such, there is a need for developing ethical artificial intelligent systems with effective governance that users can trust with their information. Several guidelines exist to help facilitate these developments; however, very few transition into methods with virtually no method existing for higher-level development methods. ECCOLA is proposed as a solution in transitioning from guidelines to development methods at higher levels. The study extends ECCOLA by evaluating its ethical tenets with Information Governance principles (Generally Accepted Recordkeeping Principles, GARP®) as a governance framework to improve its robustness in line with ethical guidelines. This was accomplished by following the Design Science Research methodology approach using a conceptual framework based on ethical guidelines of the European Commission and content analysis. The findings reveal a vulnerability of the GARP® principles of Retention and Disposition in ECCOLA. A possible solution artifact has been developed, which remains to be tested.

Keywords: Artificial Intelligence, Ethics, Trustworthy Artificial Intelligence, Information Governance, ECCOLA, GARP®), Ethical Artificial Intelligence

FIGURES

TABLES

TABLE OF CONTENTS

1 INTRODUCTION

The progress of Artificial Intelligence (AI) has arguably made it one of the most promising technologies of the current decade, providing a wealth of opportunities (Thiebes, Lins & Sunyaev, 2020). Artificial intelligence comprises various technologies that produce intelligence associated with human intelligence (Leijnen, Aldewereld, Belkom, Bijvank & Ossewaarde, 2020). AI emulates human intelligence by "perceiving" their environment, acquiring data, interpreting the collected data, processing, or "reasoning" the information derived from the data to decide the best course of action to achieve complex goals (Rossi et al., 2019).

AI has enabled applications across commercial, creative, and scientific fields, creating an awareness of its large-scale impact. Advances in technology such as machine learning (ML) have completely revolutionized AI. ML techniques have improved autonomous and semi-autonomous Artificial Intelligent systems (AIS), increasingly employed in sensitive sectors such as health, transportation, and production (Jobin, Ienca & Vayena, 2019). Its offerings range from autonomous vehicles, virtual assistants, automated services to AI-assisted health services where AI systems make diagnosis more precise, enabling better prevention of diseases (European Commission [EU], 2020). Considering the powerful and transformative impact these systems have on users, many debates and concerns have ensued regarding the ethical development and use of these systems.

Some of these debates focus on the inscrutability of Artificial Intelligent systems that employ Machine Learning techniques, resulting in ethical and practical worries in various fields (Asatiani et al., 2021). ML techniques operate mindlessly with no conscious understanding of the broader context of their processes and cannot contemplate the ethics of their actions (Asatiani et al., 2021). An example is an accident from an autonomous uber vehicle resulting in the loss of life, fuelling fears for further scaling of AIS (Rassloff, 2020). Other ethical concerns stem from worries that AIS will lead to job loss for humans, propagate bias, undermining fairness, thwart accountability, and misused maliciously to perpetuate evil. (Jobin et al., 2019.)

Furthermore, AIS is different from traditional decentralized systems as their models revolve significantly around data (Kumar, Braud, Tarkoma & Hui, 2020). AIS depends on data for training, and some rely continuously on data for learning which is instrumental throughout their lifecycle. Data enables AIS decision-making processes, so it is crucial to understand how they handle data in a manner that is perceived as fair, aligned to human values, relevant to the problem to be solved, capability for explanations, reasoning, and decision making. (Rossi, 2018.)

Thus, AI ethics deals with the moral behavior of humans in the design, usage, and behavior of machines (Müller, 2020). According to Jain, Luthra, Sharma, & Fatima (2020), ethical issues of AI apply moral values that focus on the various sociotechnical discrepancies or issues generated from the construction and function of AIS. Such problems are becoming more evident as AI technology transcends systems development beyond technological and engineering boundaries to sociotechnical boundaries (Winby & Mohram, 2018). Moreover, given the sensitivity of data, its increasing value, and the role AI plays as generators and accessors of data (Sætra, 2021), there is an urgent need to implement ethical values into the design of AI systems to enable trust (Vakkuri & Abrahamsson 2018).

Establishing trust between humans and AIS requires answers to ethical questions such as why a particular decision was made over another? At what point does the AIS succeed?; When does it fail?; When can humans trust the AIS?; And when can humans correct an error with an AIS? (Wickramasinghe, Marino, Grandio, & Manic, 2020). To this end, researchers, governments, and organizations have deliberated and produced many frameworks and guidelines. Some of these organizations are - High-Level Expert Group on Artificial Intelligence [[AI HLEG], Expert Group on AI in the society of the Organization for Economic Co-operation and Development (OECD), Advisory Council on the Ethical Use of Artificial Intelligence and Data in Singapore, and the Initiative for Ethically aligned design (EAD) for autonomous and intelligent systems by IEEE to help provide ethical guiding principles for the development of trustworthy AIS that users can trust. (Jobin et al., 2019.)

Trustworthy artificial intelligent systems (TAIS) or Trustworthy Artificial Intelligence (TAI) are geared towards strengthening human trust in Artificial Intelligent systems (Wickramasinghe et al., 2020). Trustworthiness is based on the idea that trust is a fundamental foundation for the economy and sustainable development of the society in AIS development (Thiebes et al., 2020). The concept of trust is explained as having different dimensions and interpretations. It starts with an initial trust where individuals have little or no prior experience with the other party, which can then develop into a knowledge-based trust where the individual generates enough information on the other party to make predictions. In addition, trust in a technology influences or directly impacts users trusting intentions to engage in a trust-related behavior such as sharing personal information, using a system for its functionalities, or the information it provides. (Thiebes et al., 2020.) Therefore, TAIS aims to enable humans and so-

ciety to design, develop, and use AIS without any sense of foreboding, fear, or doubt (Wickramasinghe et al., 2020).

However, these guidelines are yet to effectively transition into trustworthy developed AIS as developers struggle to implement theoretical approaches into the development processes (Vakkuri, Kemel, Kultanen & Abrahamsson, 2020). Limited literature also exists in academics and practice for proven methods that translate principles and guidelines in developing ethical and trustworthy AIS due to them being considered challenging (Mittelstadt, 2019). Existing methods, such as the Ethical framework for designing autonomous intelligent systems (Leikas et al., 2019), are more focused on design at a higher level than development (Vakkurri et al., 2020). Consequently, there seem to be no methods currently that focus on higher-level development surrounding ethical AIS. (Vakkuri et al., 2020.)

As a result, the ECCOLA method has been developed and proposed as a possible solution to creating higher-level development of trustworthy AIS (Vakkuri et al., 2020). ECCOLA is a tool for developers and product owners that seeks to implement AI ethics practically at a higher development decision level. ECCOLA is developed to help bridge the gap between research and practice to create trustworthy AIS with a human-centered approach that requires human actors to be the clear focus with methodologies designed to reflect this. (Vakkuri et al., 2020.) In addition, it aligns with The EU ex-Ante approach of aligning AIS with ethical guidelines where there is a need to determine beforehand whether AIS meets the guidelines given the sensitivity of their interaction with society (Leijnen et al., 2020). AIS that conforms to the recommended guidelines that methods and tools to be developed should allow for guidelines to be integrated during the development process to provide an understanding of what constitutes these guidelines (Leijnen et al., 2020).

While ECCOLA is a potentially powerful tool, one of its weaknesses is that it is relatively new and has not been subjected to numerous rigorous analyses to improve its robustness and widespread adoption. Thus, tools to help develop ECCOLA's robustness are needed. Hamon, Junklewitz & Sanchez (2020) explain that for robustness to be attained, method models need to be subjected to rigorous evaluations to benchmark areas that have not been taken into consideration or fully exploited. In addition, a lack of robustness in a method can lead to duplicated efforts with little practical benefits slowing the pace of research (Taschuk & Wilson, 2017). Hence, for method models like ECCOLA to fully explore its full potential and attain certification status, a need exists for further evaluation (Hamon et al., 2020).

According to Eitel-Porter, (2021), ethical principles alone are insufficient for the development and deployment of trustworthy AIS and further requires strong governance controls which manages processes and creates associated audit that enforces principles. Currently, in literature, the topic of AI governance is widely unexplored (Wirtz, Weyerer & Sturm 2020), with very few studies on AI governance and regulatory issues (Wang & Siau, 2018). This may be attributed to AI governance being a global issue and not a one size fits all rec-

ommendation but a coherent framework where different practices can vary in respect of contextual and cultural particulars (Wang & Siau, 2018). Most of the studies on AI governance are centered on algorithms and data and virtually none on AI Information governance (Wirtz et al., 2020). Since ECCOLA is relatively new, critical evaluation and analysis of its trustworthy components with IG principles can help highlight any perceived vulnerabilities for correction and improvement, which can lead to a more robust method to add to the burgeoning IG AI governance body of knowledge. Moreover, when IG principles are added to AIS development methods like ECCOLA, the governance will make it easier to scale AIS and reduce associated risks (Eitel-Porter, 2021). Therefore, this study aims to extend ECCOLA by evaluating its trustworthy components with Information Governance (IG)principles.

.

1.1 The rationale of the Research

The research on ECCOLA is still in its early stages, thus making it fit for further development and making it a robust method in line with ethical guidelines for developing trustworthy AIS. In addition, academic research on IG practices with trustworthy development methods is virtually non-existent in information systems. Thus, this research provides an opportunity to be part of the process of creating new knowledge and lessons learned as it can contribute to theory and practice.

In terms of theory contribution, the research can contribute a comprehensive view about areas of vulnerabilities within ECCOLA by applying theoretical perspectives from the AI ethics and IG research field to explain how these vulnerabilities are corrected. This new knowledge can create and add to the existing body of knowledge in making trustworthy development methods more robust for broader adoption and increase our overall understanding of ECCOLA.

For practical contribution, a robust ECCOLA method can successfully help practitioners develop trustworthy AIS more systematically to develop AIS that is less likely to fail. In addition, a robust and successful ECCOLA with practitioners within small organizations can lead to adaptation in larger organizations both within Finland and possibly the wider economy.

1.1.1 Research Objective

The objective of this study is to explore how to improve the ECCOLA method. As such, the goal seeks to fill these gaps by aiming to evaluate ECCOLA's tenets with Information Governance principles in a bid to make the method more robust hence the main research question is:

1.1.2 Method for Review of Literature

I conducted the literature review using Google Scholar, IEEE Xplore, Journal of the Association of Information System (AIS), and Information Systems Journal because of their relevance and coverage. I searched for the literature with keywords "Trustworthy artificial intelligent systems," "artificial intelligence," "ethical artificial intelligence," "artificially intelligent systems," "Information Governance." Other literature used includes "Ethics guidelines for Trustworthy AI" by European Commission (2019), Generally acceptable Record-Keeping Principles (GARP®) by ARMA and the IEEE guideline for Ethically aligned designs as references.

I conducted the research methodology part of the study by using the Design Science Research Methodology (DSRM) by Peffers, Tuunanen, Rothenberger, and Chatterjee (2007). To evaluate ECCOLA, a problem-centered initiation by way of a conceptual framework by Tolvanen (2020) as the possible entry point of research follows through the nominal process sequence.

1.1.3 Thesis Structure

The following chapter presents the literature review for this study that discusses and defines the key concepts from reviewing previous literature related to the research subject. After the literature review, the requirement for the evaluation is reviewed, followed by examining the empirical findings and a discussion that connects the empirical results to the theoretical background. In the final chapter, the study is concluded with an answer to the research question, a discussion on the limitations of the study, and possible propositions for future research opportunities.

1.2 Research Methodology

This section discusses the research design and techniques used in the study. It also discusses the rationale for the research methods chosen and how the research results are analyzed.

1.2.1 Design Science Research Methodology

The research methodology for this study follows the design science research methodology (DSRM) approach by Peffers et al. (2007) to create a design to modify the existing method, ECCOLA. Peffers et al. (2007) define design science

(DS) as a methodology that helps create and evaluate IT artifacts intended to solve identified organizational problems. Hevner, March, Park, and Ram (2004) explain that the DSRM helps extend human and organizational boundaries to create new and innovative artifacts. However, designing artifacts can be challenging due to the complexities of creative advances in fields with limited theory (Hevner et al., 2004).

According to Peffers, Rothenberger, Tuunanen, and Vaezi (2012), artifact types suitable for design science include models, framework, instantiation, and conceptual methods (non-algorithmic) actionable instructions. Hanid (2014) describes methods as steps or guidelines used to perform a task. He explains further that methods are built on underlying constructs or language and represent the solution space. In addition, methods can be attached to specific models wherein the steps take part of the model as input and can translate from one model to another during problem-solving. (Hanid, 2014.) The DS method is employed in this study to help achieve the goal of extending ECCOLA.

1.2.2 Design Sequence

A" problem-centered initiation" of the design science research method is employed as the possible entry point of research to achieve this goal. It follows through the nominal process sequence as illustrated in figure 1.

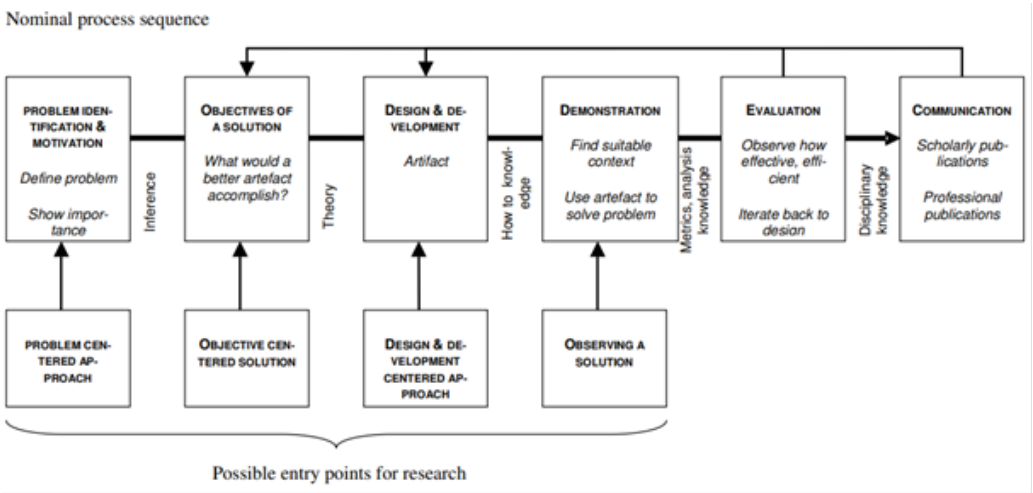

Figure 1 Design Science Research Process model (Peffers et al., 2007)

1.2.3 Problem Identification and Motivation

The research starts by identifying the problem and motivation, which is the first step in the Design Science Research Model. According to Peffers et al. (2012),

the Design Research process might start with an existing practical solution that is yet to be rigorously developed or documented, which is the case with ECCOLA (Vakkuri et al. 2020). In attaining this goal, A review of ECCOLA is conducted using a conceptual framework by Tolvanen (2019) based on the ethical guidelines of the AI HLEG (European Commission, 2019) to serve as the problem-centered approach.

ECCOLA's trustworthy tenets were vigorously reviewed with ethical and trustworthy principles from the guidelines for trustworthy AIS (European Commission, 2019) using the conceptual framework. This review provided a deeper understanding and helped to establish what has been done from what needs to be done, the context of the problem, the discovery of essential variables relevant to the topic, synthesis of the materials to gain a new perspective, and identification of new ideas and practice, (Onwuegbuzie & Frels, 2016).

The outcome of the review revealed a vulnerability in ECCOLA associated with accountability via governance frameworks. Effective governance practices in an ethical method help instill confidence for developers and users as it pertains to information management. Thus, the need to determine ECCOLA's accountability via a governance framework was established. Information governance was chosen due to the sensitivity of development methods with information assets.

1.2.4 The objective of a solution

The second step defines objectives for a solution. Based on the outcome of the evaluation carried out in step one, an extensive literature review was conducted to help determine and collect IG practices and guidelines. The examination revealed the Principles®, or GARP®, by ARMA (2009) to be the most accepted standard worldwide that provides a critical high-level framework of good practices for IG.

Based on this, each IG principle in the GARP® was critically analyzed with each tenet of ECCOLA using content analysis. A content analysis helped provide a guide from the GARP® IG practices used to identify similar practices in ECCOLA to determine which of the ECCOLA cards had IG practices and which did not.

The outcome from this process produced an innovative idea of a heatmap. The heatmap highlights areas in ECCOLA where IG (GARP®) practices can be incorporated, further incorporated, and areas that already reflect these practices. Hence, the objective of the solution in incorporating IG practices can improve the accountability of ECCOLA via an Information governance framework, thereby extending the method.

1.2.5 Design and Development Stage

In this stage, based on the findings of the heatmap, which identified the GARP® principles of retention and disposition to be the least incorporated

ARMA IG principles in ECCOLA, a possible solution was identified within the solution space. The outcome of a new artifact in the form of a card [#21] is designed. Card #21 embodies the practices of retention and disposition, which explains instances where the efficacy of the 22nd card can be determined by examining how it resolves the problems we identified in stage one.

1.2.6 Demonstration and Evaluation

While the study does not include a practical demonstration of the new artifact now, the latest artifact from this study was presented to the thesis supervisor, professor Pekka Abrahamsson and the ECCOLA team at the university to provide the necessary evaluation needed for this stage. The assessment will help provide critical feedback on the validity and viability of the card. A demonstration may be carried out as an extension of the study with developers to help evaluate the validity and reliability of the study at a determined time.

1.2.7 Communication

Finally, the design science research processes and findings are communicated in this master thesis, the final stage in the DSRM.

2 LITERATURE REVIEW

This chapter discusses the concept of Artificial Intelligence, Definition, Characteristics, Artificial Intelligent Systems, Machine Learning, Explainable Artificial Intelligent Systems, and Ethical Artificial Intelligence Systems.

2.1 Background

Artificial Intelligence (AI)I can be traced back to the 1950s with the design of the Turing machine by Alan Turing and his research on making computers more intelligent and capable of replicating the human brain (Simmons & Chappel,1988). The concept at the time met with limited success due to inadequacy in knowledge and technology, making the research dormant until the 1960s and 1970s, when further research paved the way for the current technology. According to Simmons and Chappel (1988), the term artificial intelligence was coined for expert systems (ES), which use knowledge-based application and inference procedures to solve problems. ES mimicked human intelligence as they had the capability of processing symbols as numbers which were deficient in computing systems at the time. (Tan et al., 2016.)

2.2 Definition of Artificial Intelligence

Attempts at fully defining Artificial Intelligence (AI) have resulted in many definitions. McCarthy et al. (2006), based on the Dartmouth AI project of 1955, defines AI as making machines behave in a manner that would be intelligent if humans behaved in a like manner. Rich (1983) describes it as the study of making computers do things that humans do better now. These definitions, centered on computers acting like humans, were too restrictive. It excluded the arm of AI that dealt with neural network research, leading to a contrasting definition. Haugeland (1989) defines AI as not just mimicking human intelligence but be-

ing full-blown machines with minds or machines that give a perception of having a mind of their own. This definition attributed the human thinking capabilities to computing systems bringing about confusion because AI was incapable of confounding human ability to overtly solve a specific class of problem and latently discover solutions for new problems (Simmons & Chappel, 1988). Thus, the concept of human intelligence is hard to define, with the definition of AI evolving depending on conceptualization (Arietta et al., 2020).

A common concept found in literature alludes to AI demonstration of behavior associated with human intelligence. Wierenga (2010) describes AI as representing human-like intelligence in computers and how it is harnessed in daily lives to improve businesses. Ertel (2018) describes AI as the simulation of computers or machines to perform tasks or processes which currently humans are better skilled at. Ma and Sun (2020) describe AI as the affordance of human intelligence to machines. But this concept does not accurately capture AI capabilities of computing and processing vast amounts of data and proffering solutions within a short period beyond human capability (Følstad, Nordheim & Bjørkli, 2018).

Artificial Intelligence's ability to process vast amounts of data while thinking based on knowledge enabled AI to be considered transparent and easily interpretable as knowledge-based Expert Systems (ES). However, over the years, with improved technology and the emergence of Machine Learning (ML) techniques, AI systems have become increasingly complex and opaque. Therefore, the definition of AI has evolved depending on its conceptualization as it is pretty challenging for one description to capture its entire essence. (Arietta et al., 2020.)

Guresen and Kayakutlub (2011) define AI in terms of artificial neural networks as a parallel fusion of simple processing units that acquire knowledge from the environment by learning a process and storing it in its connections. This definition extends the intelligence of AI from intelligent processing to include learning. Leijnen et al. (2020) describe AI as a series of different technologies that together produce intelligence. They explain that AI, in most cases, refers to applications in machine learning where computing machines deduce rules from data (Leijnen et al., 2020).

Kwon, Bae, and Shin (2020) describe AI as computing systems that can sense, comprehend, act, and learn from data to enable them to deliver value. AI emulates human thought processes, adaptivity, and reasoning by learning from human data. They operate using sensors and smart programs, enabling them to react to environments, communicate, plan, reason, problem-solve and represent data in information. AI is increasingly becoming capable of responding emotionally. Sentiment analysis combined with anthropomorphic features has enabled emotive behavior in AI (Følstad et al., 2018.)

Arrieta et al. (2020) define AI in terms of transparency *as machine learning systems that enable users to understand, trust appropriately and effectively manage artificially intelligent partners.* This definition by Arrieta et al. (2020) is considered the most suitable for this study. It explains AI in terms of transparency and alludes to ethics. It also describes how transparent AI can instill trust and confi-

dence in users enabling more informed users. For AI to make a positive impact, it needs to be thoroughly tested, explainable to its users, and have all its ethical considerations in place (Rudin & Radin, 2019).

2.2.1 AI Classification

Haenlein and Kaplan (2019) classify Artificial Intelligence (AI) into two categories: firstly, based on its evolution, and secondly, based on the intelligence it exhibits. According to Oosthuizen, Botha, Robertson, and Montecchi (2020), AI has evolved over the years along with the concept of intelligence. They delineate three stages of intelligence as narrow, general, and superintelligence.

- The narrow or artificial narrow intelligence phase represents the current application of human-level intelligence. In this stage, AI agents apply human-level intelligence (text, speech, and sound = data) to produce outputs such as voice and text recognition capabilities as exemplified by conversation and robotic agents such as Pepper, the customer service humanoid (HSBC, 2019). (Oosthuizen et al., 2020.)
- The artificial generalized intelligence phase signifies a strong human-level intelligence where AI systems develop the capability to perform tasks autonomously.
- The superintelligence level of AI signifies an above conscious human-level of intelligence where AI systems develop capabilities and instantaneously solve complex problems. (Oosthuizen et al., 2020.)

Most of the current AI technology and applications in most sectors fall under the artificial narrow intelligence category as AI is yet to attain the other two stages (Oosthuizen et al., 2020). The second stage of AI classification by Haenlein and Kaplan (2019) is based on the intelligence it exhibits, which are cognitive, emotional, and social. This classification is further explained in terms of analytical, human-inspired, and humanized capabilities, as illustrated in (table 1)

TABLE 1 Classification of AI (Haenlein & Kaplan 2019) Continued on next page

	Analytical AI	Human-inspired AI	Humanized AI	Human beings
Cognitive Intelligence	↓	↓	↓	↓
Emotional Intelligence	X	↓	↓	↓
Social Intelligence	X	X	↓	↓
	Supervised Learning, Unsupervised Learning, Reinforcement Learning	Supervised Learning, Unsupervised Learning, Reinforcement Learning	Supervised Learning, Unsupervised Learning, Reinforcement Learning	

- Analytical AI has characteristics consistent only with cognitive intelligence; AI systems in this category learn by using experience (data) to inform future decisions. The majority of AI applications fall under this category (Haenlein & Kaplan, 2019.)
- Human-inspired AI learns from cognitive and emotional intelligence and tends to understand and exhibit human emotions in their decision-making and interactions (Haenlein & Kaplan 2019). Sentiment analysis, machine learning capabilities combined with anthropomorphic features enable AI systems to interact and respond emotionally to users. AI systems are increasingly being trained to simulate human emotions accordingly to enhance their interaction with humans, but AI systems are currently unable to feel human emotions. (Følstad et al., 2018.)
- Humanized AI tends to learn from all three categories to develop its competencies and be self-conscious and self-aware in their interactions; however, these systems are yet to become available (Haenlein & Kaplan, 2019).

Artificial Intelligence's ability to learn from data is enabled by a learning process that is either supervised, unsupervised, and Reinforcement. Supervised learning refers to the traditional form of learning, which includes a set of given inputs used to derive a set of outputs. Unsupervised learning refers to training using a set of given data input, and the AI system must determine or infer the output from the data. In Reinforcement Learning, AI systems are allocated a set of output data and must maximize the output by a series of decisions. The AI system infers output from learning; as such, there is no way of accessing the accuracy of the output but to trust the system. (Haenlein & Kaplan 2019.)

2.3 Artificial Intelligent Systems

Rossi et al. (2019) define Artificial Intelligent systems (AIS) as human-designed software and possible hardware systems that act in the physical or digital dimension by perceiving their environment, acquiring data, interpreting the collected data and process, or reason the information derived from the data to decide the best course of action to achieve complex goals.

AIS is described as "rational" due to its capability to perceive its environment using sensors, which enable the collection and interpretation of data. Reason or process collected data to produce information; decide the best course of action most suitable, and use actuators to modify their environment based on their decision. AIS rationality is categorized as rational and learning rational. (Rossi et al., 2019.)

Rational AI systems modify their environment but do not adapt their behavior over time to better achieve their goal. On the other hand, a rational learning AIS takes action by evaluating the new state of the environment and adapting its reasoning rules and decision-making methods. AIS can use either a symbolic rule or learn a numeric model and adapt their behavior by analyzing how it affects previous actions (Rossi et al., 2019.) An illustration of an AI system is given in (figure 2):

AIS can be software-based operating in the virtual world like virtual assistants, image and speech recognition systems, search engines, or embedded in hardware such as autonomous vehicles, robots, and drones (Leijnen et al., 2020).

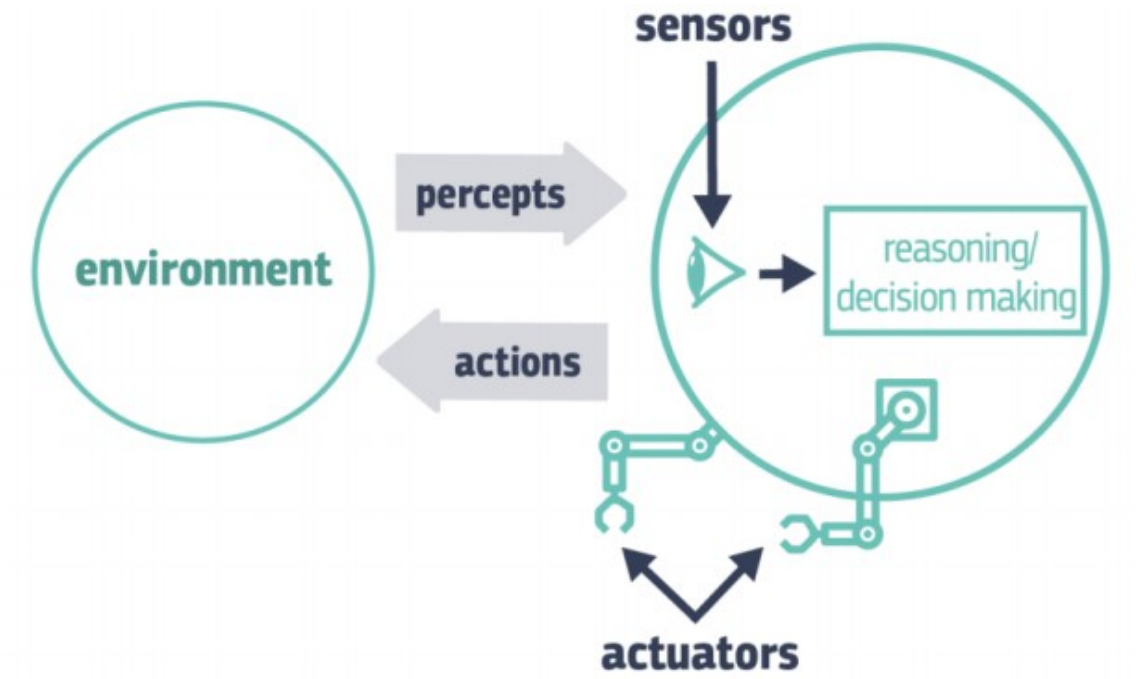

Figure 2 Schematic representation of an AI system (Rossi et al., 2019

The schematic representation is a basic depiction of how an AI system is structured. It shows the essential functions of an AI system in its operations. The functions are further explained below.

- Sensors refer to all input devices (microphones, keyboards, websites, and physical quantities (temperature, pressure, tactile sensors) that adequately perceive (percepts) the data present in the environment and relevant to the goal. (Rossi et al., 2019.)
- The reasoning and information-making module of AIS lie at their core and are responsible for inputting incoming data from sensors and propose an action to achieve the required goal. They interpret the data into numeric information models and then reason or process them to decide the best course of action. (Rossi et al., 2019.)

- Actuator: When AIS makes decisions on the best course of action, it outputs or performs the decision through its actuators which could be a physical entity or software which modifies the environment (Rossi et al., 2019.)

AIS techniques can be broadly attributed to their reasoning and learning capabilities (Rossi et al., 2019).

- Reasoning involves processing input data and representing the data as knowledge. It involves making inferences of the knowledge using symbolic rules, planning, and scheduling algorithms, searching through a large solution set to enable optimization among all possible solutions to make the best decision. Knowledge representation helps to transform or process data into information. According to Fikes and Garvey (2020), effective knowledge representation and reasoning methods are fundamental requirements for AIS.
- The Decision-making process or predictions of AIS is complex and requires a combination of several of these processes. AIS processing power in making decisions is far superior to human processing. AIS never gets tired, manages, and processes vast amounts of data by searching through a large solution set using symbolic rules, planning, and scheduling activities with fast and iterative intelligent algorithms in a short amount of time. As a result, most e-commerce and content platforms employ AIS in deploying their services. Websites such as Amazon or Netflix rely on AIS decision-making or prediction capability to provide recommendation services for their users. (Rossi et al., 2019.) In addition, AI systems learn from the outcomes and keep improving to produce better models. (Jarahi, 2018.) On the other hand, Jarrahi (2018) explains that AIS decision-making supports an analytical approach but is less capable of understanding the native intelligence situations like humans and less viable in uncertain and unpredictable environments outside their predefined knowledge domain. Humans use common-sense reasoning that is not based on fact but a judgment call and, as such, presents more intuitive decision-making (Jarahi, 2018).
- Learning involves using several techniques that teach AIS to solve problems that cannot be specified precisely or solution methods that cannot be described using symbolic reasoning rules. Problem sets that involve human cognitive capabilities such as perception, speech, language understanding, computer vision, and behavior prediction are challenging for AI systems paving the way for machine learning techniques. (Rossi et al., 2019.)

2.3.1 Machine Learning

Ma and Sun (2020) describe Machine Learning (ML) based on Mitchell (1997) definition as "A computer program is said to learn from experience E with re-

spect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience (p. 482)."

They explain that ML technology evolved separate from AI but has fast become the central paradigm of AIS and is considered a subfield of AI due to the vast amount of AI research being focused on ML since the 1990s. Increased interactions between computing systems and users continue to generate vast amounts of individualized digitized data prompting these heavy investments (Ma & Sun, 2020.)

ML techniques employ algorithms in generating numeric models to compute decisions from data and are effective where traditional quantitative methods are inefficient and capable of processing structured and unstructured data in real-time to provide accurate predictions. They are efficient in modeling predictive data analytics applications and computing tasks where the design of the algorithm is difficult or nearly impossible. (Ma & Sun, 2020.)

ML approach involves giving computing systems instructions that allow them to learn and improve performance from data without providing step-by-step instructions from the programmer, thus allowing them to be used for newer and complicated tasks that could not be programmed. (Ledesma et al., 2018). A BCC Research estimates the global market for ML solutions to grow annually at a rate of 43.6% to reach $8.8 billion by 2022 (BCC research, 2018).

Learning or training of data for ML algorithms are categorized as supervised, unsupervised learning, and reinforcement learning, as illustrated in figure 3.

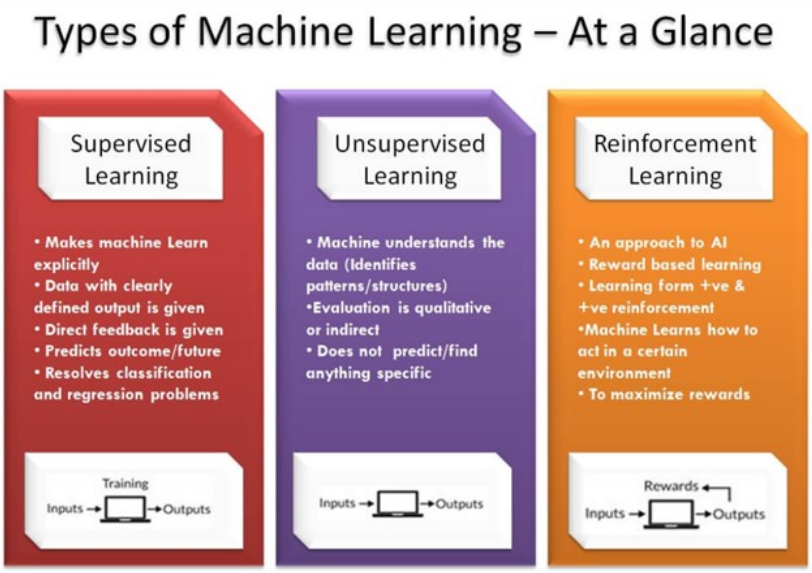

Figure 3 Representation of machine learning (NewTechdogo, 2018)

- Supervised learning enables ML algorithms to learn from a dataset provided for the instance, with prediction being the key focus. Different datasets are employed in evaluating the accuracy of predicted outcomes. Thus, the ML models are trained using the training data subsets and

tuned or selected using the validation subset. A key focus of supervised learning is predictions, as users are more interested in maximizing the prediction of the outcome than uncovering the linkage between the variables. (Ma & Sun, 2020.) Supervised ML is frequently used in building predictive models by extracting patterns from large datasets (Ledesma et al., 2018)

- Unsupervised learning involves training datasets containing only the input variables. The output variables are undefined or unknown, with the goal being to find or determine the hidden patterns or information from the dataset (Ma & Sun, 2020). Unsupervised learning is performed as part of an exploratory data analysis and involves ML algorithms analyzing datasets containing many different features and learning valuable features and properties from the datasets (Ledesma et al., 2018). The components extracted from the datasets possess vital information from the original datasets and can be interpreted or used for subsequent analysis. Usama et al. (2019) explain that unsupervised learning facilitates the analysis of raw datasets, generating analytic insights into unlabelled data (Usama et al., 2019).
- According to Ma and Sun (2020), reinforcement learning involves algorithms continually interacting with the surrounding environment by acting and observing feedback to optimize a specific objective function. Sutton and Barto (2015) explain reinforcement learning as learning what to do - how to map situations to actions - to maximize a numerical reward signal. In reinforcement learning, the learning agent is not told the course of action to take in a closed-loop problem and must discover which action yields the most reward by exploring all the possible courses of action as the action may affect not just the immediate compensation but all subsequent rewards. The three most significant factors in reinforcement learning being the closed-loop, no direct instructions on the course of action to take, and where the consequence and actions play out over extended periods. (Sutton & Barto, 2015.)

Machine learning effectively creates AIS due to its ability to process structured data, unstructured data, complex data structure. ML also accommodates data of hybrid formats, large data volumes, offers flexibility and prediction capabilities. However, ML techniques in AIS usually lack interpretability regarding a transparent model structure and clear linkage between variables. This lack of interpretability is mainly due to their reliance on engineering features and flexible model structure, resulting in a black box that delivers predictive accuracy rather than interpretive insights. (Ma and Sun, 2020.) Deep learning advances in ML have further fuelled a lack of interpretability owing to their capability of processing raw data without careful engineering and domain expertise (Usama et al., 2019).

2.3.2 Deep Learning

Deep Learning (DL) received little attention until 2006 as it was considered expensive due to the high computational costs of deep-learning procedures associated with it (Usama et al., 2019). Moreover, beliefs that deep learning training architectures in an unsupervised and supervised manner were considered intractable and poor performance riddled with errors. (Usama et al., 2019). However, this notion has been dispelled as DL has proven extremely useful and efficient in training datasets (Hinton, Osindero & Teh, 2006).

Deep learning can be described as a class of machine learning that utilizes hierarchical architectures for unsupervised learning. The models that are generated are used for classification and other related tasks. Hierarchical learning refers to learning simple and complex features from a hierarchy of multiple activations, whether linear or nonlinear, and reflects how DL is performed in modern multi-layer neural networks. (Usama et al., 2019.) According to Bengio (2009), DL methods aim to learn feature hierarchies with features from higher hierarchy levels formed by the composition of lower-level features. He explains that automatically learning features at multiple levels of abstraction enables a system to learn complex functions and become capable of mapping input to output directly from data instead of depending completely on crafted human features (Usama et al., 2019).

DL employs the concept of artificial neural networks (ANN) or deep neural networks (DNN), a deep structure consisting of multiple hidden layers comprising numerous layers in each layer, nonlinear activation function, cost function, and backpropagation algorithm (Usama et al., 2019). ANN is modeled after the analogy of the human brain and comprises a network of neurons densely connected and programmed to identify similarities between datasets (Altman, 2017).

Complex functions map an input to output through deep layers of neurons. ANN models high-level abstraction in data transformation that attains depth deep enough where learning takes place to enable a machine to self-learn complex models or representations of data (Usama et al., 2019). A simple illustration of ANN is given in Figure 4.

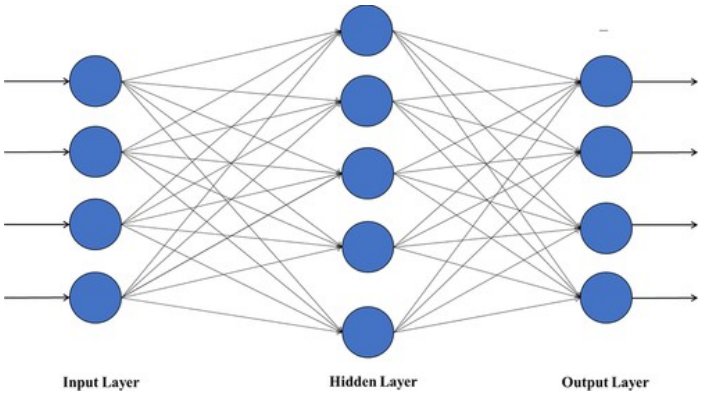

Figure 4 A simple illustration of ANN (Garg et al., 2020)

ANN comprises three layers: the input layer, the hidden layer, and the output layer, with each having different activation parameters. The input layer is made of the input neuron, the middle layer that is not exposed to the outside world is made up of multiple layers of neurons depending on the complexity of the task, and the output layer comprises the output neurons.

The learning process involves mapping optimal activation parameters that enable ANN map input to output with different problem sets requiring different layers. A problem may require multiple sets of hidden layers, which involves a long chain of computations, and others may not require such depth. ANN learning presents some ethical issues as the multiple sets of hidden layers are mainly responsible for unexplained decisions made by AIS (Usama et al., 2019.) Its versatility is harnessed in AIS, such as autonomous vehicles, speech processing, and image recognition. ANN used with problem sets requiring simple input and output data are interpretable and straightforward. However, larger datasets make the network topology time-consuming, complex, and challenging to interpret in the design of AIS and one of the leading ethical concerns in the design and development of AIS (Kumar, Reddy & Praveen, 2019).

2.4 Ethical Artificial Intelligence

The demand for more ethical and explainable AIS is on the rise because humans are becoming more reticent in adopting technologies that are not directly interpretable, tractable, and trustworthy (Arrieta et al., 2020). Jain et al. (2020) agree and explain that Ethical AIS will further explore the possibility of AIS being developed to influence fundamental societal values such as Sustainable Development Goals (SDGs), where the main objectives for developers will focus on maintaining sustainable order in the society.

Eitel-Porter (2021) describes the Ethics of AI as the practice of using AI with good intentions in empowering employees and businesses with a fair impact on customers and society. It also helps in engendering trust and in the scaling of AI technology with confidence. (Eitel-Porter, 2021). Theodorou and Dignum (2020) explain that ethical AI is not intended to give machines responsibilities for their actions and decisions but to give people and organizations more responsibility and make them more accountable.

Challenges posed by the black-box nature of AIS are some reasons for an increasing trend towards demands for ethical AIS that are more transparent given their proliferation in a critical context in application areas (Müller, 2020). Other reasons such as privacy and trust issues have been expressed as organizations employ unethical ways to collect customer data which could violate user's privacy (Culnan, 2019). In addition, AI reportedly has a history of unfairness regarding ethnicity, gender, and race (Sen, Dasgupta & Guptal, 2020). The EU analyses that sustained use of AIS could lead to breaches of fundamental hu-

man rights if ethical and governance policies are not implemented at the development stage (European Commission, 2020). A host of these concerns are addressed further.

2.4.1 Artificial Intelligence Concerns

A review of the literature reveals some AI concerns, as explained in (table 2). While some stem from risks associated with security and safety, most are rooted in ethics.

Table 2 Artificial Intelligence systems concern

Sector	AI concern	Author
Health	AI bias	Reddy, Allan, Coghlan & Cooper, 2019
	Governance	
	Ethical	Kiener, 2020
	Privacy	Culnan, 2019
	Trust	Sen et al., 2020
	Lack of transparency	Stembert & Harbers, 2019
	Regulatory	
	Unexplainable AI	
	Safety & Efficacy	
	Liability	
	Cyberattacks	
	Black box	
	Systematic bias	
	Mismatch	
Finance	Bias	Caron, 2019
	Variance in the accuracy of algorithms	Königstorfer & Thalmann, 2020
	Privacy	
	Governance	
	Transparent	
	Regulatory	
Social contract	Safety	Caron & Gupta, 2020
Politics	Deepfakes	Nguyen et al., 2019
Law	Explainability	Atkinson, Bench-Capon & Bollegala, 2020
Public Policy	Lack of transparency	Robinson, 2020
Human Rights	Vulnerability	Rodrigues, 2020
	Deepfakes	Kwon et al., 2020
	Ethics	
	Governance	
Education	Algorithm Bias	Yang, Ogata, Matsui & Chen, 2021
Cybersecurity	Data misclassification	Yamin, Ullah, Ullah & Katt, 2021
	Synthetic data generation	
	Data analysis	

Reddy et al. (2020) explain that with AI technologies such as Deep Learning (DL) gaining ground in sensitive sectors such as healthcare, a need exists for a more effective governing structure to tackle ethical and regulatory concerns embedded in issues like transparency, bias, privacy, and trust. They proffer a governing solution based on fairness, trustworthiness, transparency, and accountability. A proper governing of AIS can help tackle biased training datasets that do not represent the target population, inadequate or incomplete data, and inadvertent historical data, resulting in an overall bias in AIS and creating discrimination and disparities. (Yamin et al., 2021; Rodrigues, 2020.) Kwon et al. (2020) corroborate this by analyzing the need for guiding governing frameworks to guide ethical decision-making to address AI concerns. AI engenders ethical opportunities for abuse such as privacy breaches, the misuse of genetic data banks, information ethics, and social media. They explain that ethical issues of AIS do not stem from the technology itself but from human design and development; however, the results are felt by end-users. Cases of bias such as racist coding by biased humans or picture misrepresentation of animals to be humans represent some use cases (Kwon et al., 2020.)

AIS auto-learn from real-world use to continually improve their performance over time, presenting a challenge regarding regulation. This learning process is because some of their newly acquired features go beyond the initial approved policies and guidelines. In addition, regulatory standards that assess AIS safety and impact are yet to be formalized in many countries, which could yield unsafe practices in the use of AIS. The issue of liability and responsibility that occur from errors arising from the use of AIS poses an additional challenge as there are no delineated governing practices. (Reddy et al., 2020.) In addition, Lax data protection rules in some countries may permit collecting some form of data without users' consent leading to a breach of privacy. As a result, the need for a more structured governing structure that protects privacy breaches and can result in severe psychological and reputational harm is needed (Reddy et al., 2019; Yang et al., 2021.)

Nguyen et al. (2019) explain ethical concerns citing AI's versatile and dual nature, enabling their adoption for practical and destructive purposes. AI automation that enhances business processes can simultaneously be employed in maliciously carrying out cybercrimes and physical attacks such as the use of drones and deep fakes. Due to its non-transparent nature, ANN is effectively utilized in education, automation, and the arts to provide value for end-users. At the same time, they can also be misused in an unethical fashion to create deep fakes which are used in political or automated phishing attacks and yield potentially devastating effects. Thus ANN increases threats associated with privacy invasion and social manipulation (Zwetsloot & Dafoe, 2019.); Therefore, AIS should be developed using ethical methods to help check their misuse (Nguyen et al., 2019).

According to Ischen et al. (2020), the non-transparent nature of AIS such as chatbots creates ethical privacy concerns. They explain that chatbots in their

anthropomorphic interaction create an environment of comfort that enables easy interaction with users who provide personal information in exchange for a valuable recommendation. While this exchange is beneficial for both parties, users may be unaware that their information is being collected and stored; Raising concerns for users about data privacy, and they may feel exploited. (Ischen et al., 2020).

Königstorfer and Thalmann (2020) analyze the unethical exploitation by banks in using AIS to collate data from interactions between employees and customers without their full consent. They explain that such practices could be construed as an invasion of privacy, which could constitute misconduct under the European General Data Protection (GDPR). In addition, such practices could be counterproductive as employees and customers could become less transparent in their communication which negates the importance of capturing accurate quality interaction data. Moreover, customers trust their banks, and if their information is compromised at any time, it could lead to customer churn. (Königstorfer & Thalmann, 2020.)

Caron (2019) explains that the benefits of AIS in finance cannot be overemphasized but, it also raises concerns as using the same AI techniques to provide identical services such as risk profiling for a wide range of customers can trigger in herding behavior in financial systems. In herding behavior could potentially lead to financial crises. Thus, additional regulatory requirements are essential in developing AIS to help mitigate any technical, financial, and legal concerns such as compatibility and fragmentation of social classes resulting from ignorance and access restriction (Caron, 2019).

Atkinson et al. (2020) explain that the use of AIS in law poses a great advantage due to DL technology and brings with it several ethical challenges in terms of Explainability. The black-box nature of non-transparent AIS elucidates little trust in users and creates a barrier to its adoption. Legal AIS with feature selection decision-making has transitioned from manual to automated, making it hard to identify which feature was selected to make predictions due to DL's unexplainable nature. They explain that this makes it particularly challenging in the field of Law as legal decisions must be devoid of any discriminatory biases and based on existing law and principles of natural justice. But, DL-powered AIS learns automatically from training data; as such, there is no guarantee that feature selection data will be based on natural justice and existing law. (Atkinson et al., 2020.) Users of AIS have a right to an explanation as the challenge in explaining AI algorithm decisions emanates not from the complexity of the algorithms but in giving meaning to the data it draws from (Rodrigues, 2020).

Kiener (2020) discusses the need for increased security in the use of AIS due to their susceptibility to cyber-attack given the sensitivity of the applications they support in different fields. AIS increasingly faces security and trust issues, is vulnerable to web attacks, and requires particular attention in the security of its systems to enable trust in their adoption. Attacks such as input attacks, a particular vulnerability in AIS where data is engineered to give wrong results, lead to decreased confidence in AIS. These issues are rooted in the de-

sign and development of AIS, usually from the black-box nature of AI, and such vulnerabilities could not be satisfactorily exploited in other transparent information systems. (Kiener, 2020.) Robinson (2020) emphasizes the need for transparency in AIS through AI literacy education, clear algorithmic decision making, and openness by creating data lakes and data trust as the way forward in combating the issue of transparent AI systems. He analyses that the Nordic region has exemplified these values and striving towards explainable and transparent AI systems.

According to Caron and Gupta (2019), AIS should be viewed as a social contract. Their adoption and implementation should include a clear identification for their purpose at the time of design/development prior to deployment and scaling. They argue that identification needs to be done technically and through governing policies in an explicit, unambiguous, and clear human language with goals aligning with human rights, safety, and considerations. They further clarify that while most technologies possess an inherent level of safety risk for society, there are generally very high expectations for technologies to present very low probabilities of hazard or danger when used, and if there be any possibility of harm, then adequate warning and full disclosure of such risks should be brought to the knowledge of the users. (Caron & Gupta, 2019.)

2.4.2 Explainable Artificial Intelligence

The issues discussed in the section above have given rise to demands for better regulatory and governance practices in the development of AIS (Reddy et al., 2020). In addition, the need to better understand and explain AIS decisions in an ethical manner has led to explainable AI. According to Adadi and Berrada (2018), the problem of the explainability of AIS has existed since the 1970s in the study of expert systems; as such, there exists no standard and generally accepted definition of explainable AI (XAI). They describe XAI as the movement, initiatives, and efforts made in response to the concerns in AI with respect to trust and transparency more than a formal technical concept (Adadi & Berrada, 2018).

Gunning (2017) defines XAI as a suite of human learning techniques that will enable human users to understand, appropriately trust, and effectively manage the emerging generation of AIS. He explains that some of the current ML technologies are opaque, non-intuitive, and difficult for humans to understand. XAI aims to create ML ethical AIS that are transparent, trustworthy, and explainable for human users (Gunning, 2017).

Arrieta et al. (2020) define XAI as systems that produce details or reasons to make its functioning clear or easy to understand given an audience. They explain that XAI is sought by different audiences, as depicted in figure 5, and as such, the users should be the key focus in the design of XAI systems.

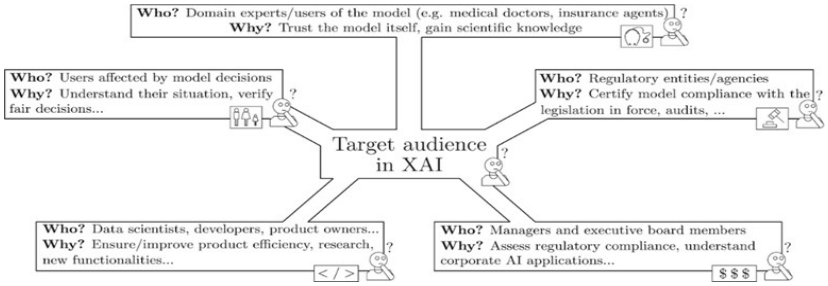

Figure 5 A representation of the different target audiences for XAI systems (Arrieta et al., 2020).

The importance of Explainability as an active attribute in XAI systems denotes any action or procedure is taken by the system with the intent of clarifying or detailing its internal functions (Arrieta et al., 2020). Ribeiro, Sing, and Guestrin (2016) list some desired characteristics for XAI: interpretability, understandability, comprehensibility, Explainability, and transparency. Interpretability is defined in terms of XAI as providing qualitative understanding between input variables and response and considers users' limitations, should be easy to understand with the level of understanding depending on the target audience. Understandability involves the capacity of the target audience to understand the XAI inner structure working or algorithm without the need to provide further explanation. Comprehensibility describes XAI's ability to present its learned knowledge in a human-understandable fashion. (Ribeiro et al., 2016.) Explainability denotes XAI's ability to act as an interface between humans that simultaneously serve as an accurate proxy for the decision-making and understanding of humans. Transparency Denotes understanding; XAI is transparent if it is understandable. (Arrieta et al., 2020.)

According to Rudin & Radin (2019), XAI or interpretable models provide a technical and possibly more ethical alternative to the black-box model. XAI is constrained to provide a better understanding of how algorithms make predictions, are simpler or decomposable, and provide a new level of insight than the black-box model. However, developers believe that the more transparent or explainable an AI model is, the less accurate it becomes, implying a trade-off between accuracy and Explainability. (Rudin & Radin 2019.)

But, Rudin and Radin (2019) argue that sacrificing accuracy for interpretability is inaccurate and could be a marketing ruse by developers to profit off complex black box models to the detriment of affected individuals where simple interpretable models are capable of producing the same results. Challenges raised by the black-box nature of AIS have inspired an increasing trend towards demands for ethical AIS that are more transparent given their proliferation in critical application context areas such as medicine, transportation, and security. Therefore, black box models need not be black box as in that state, they mask a series of potential mistakes, and even deep neural networks (DNN) that account for some of the most complex AI models can be made interpretable. (Rudin & Radin, 2019.)

Various ways for XAI systems implementation have been examined. Some suggest the participation of stakeholders involved to varying degrees if a level of success is to be achieved in creating ethical, trustworthy AIS. Jobin et al. (2019) identify 84 guidelines for ethical AIS, which include the EADe1 and the EU guidelines for trustworthy AIS. The EU High-Level Expert Group on AI [AI HLEG] was established to set up guidelines for developing ethical AIS in the European Union (European Commission, 2019). In turn, it has given rise to different frameworks and methods for developing ethical AIS.

Currently, Ethics in AIS is geared towards Ethically Aligned Designs (EAD) (IEADe1, 2019) in response to growing concerns to help raise awareness among industry professionals (Vakkuri & Abrahamsson, 2018). Research in AI ethics and how the concept can aid the practical implementation of ethics into AIS is a necessity that is still lacking (Vakkuri et al., 2019). But a shift from design to development has enabled awareness courses for AIS development to now be included in course curriculums in educational institutions to help with awareness creation from the development stage (Vakkuri et al., 2019).

3 TRUSTWORTHY ARTIFICIAL INTELLIGENCE

Wickramasinghe et al. (2020) explain that the trustworthiness concept was borne out of the need for human users of AIS to trust these systems due to the black box phenomenon. According to Theodorou and Dignum (2020), AIS is not what is ethical and trustworthy but the social component of the socio-technical system. Therefore, adequate responsibility and consideration within an ethical framework are needed to create trust in the overall system for society. (Wickramasinghe et al., 2020.)

Janssen, Brous, Estevez, Babosa & Janowski (2020) describe Trustworthiness as properties through which trusted entities serve the interest of the trustors directly or indirectly. Trust is explained as having different dimensions and interpretations, which starts with an initial trust where individuals have little or no prior experience with the other party and then develops to knowledge-based trust where the individual generates enough information on the other party to make predictions (Thiebes et al., 2020). In addition, trust in a technology influences or directly impacts users trusting intentions to engage in trust-related behavior such as sharing personal information, using a system for its functionalities, or the information it provides (Thiebes et al., 2020).

Hence for trust to be established between humans and AIS, a need exists for answers explaining questions such as why a specific decision was made over another? When does the AIS succeed? When does the AIS fail? When can humans trust the AIS? And when can humans correct an error with an AIS? The goal of trustworthy AI being to strengthen human trust in AIS. (Wickramasinghe et al., 2020.)

Therefore, Trustworthiness is based on the idea that trust is a fundamental foundation for the economy at large and sustainable development of the society in the development of AIS. An AIS can be considered *as being trustworthy by users when it is developed, deployed, and used in a manner that ensures its compliance with the relevant laws, is robust, especially in its adherence to general ethical principles, specifically the ethical principle of the (AI HEG).* (Thiebes et al., 2020.) This definition of Thiebes et al. (2020) is most suitable for this study.

TAIS addresses ethical impact in three main areas, which include: Ethics of data which focuses on the issues raised in the collection, analysis, profiling, advertising, and use of data sets; Ethics of algorithm, which focuses on autonomy and complexity of machine learning techniques and applications; and Ethics of practices which focus on responsibilities and liabilities involved in the lifecycle of AIS such as the organizations, developers, system users, adopters, and data scientists. (Wickramasinghe et al., 2020.) The scope of this study falls within ethics of practices as it covers development methods for developers of TAIS.

According to Mayer, Davies, and Schoorman (1995), three characteristics that describe the component of trustworthiness are ability, benevolence, and integrity. They explain that each of these characteristics is interrelated but separate and of equal importance. According to Wickramasinghe et al. (2020), five components of trustworthiness include inclusive growth, sustainable development, and well-being; Human-centered values and fairness; Transparency and Explainability; Robustness, Security, and safety; and Accountability. They analyze that all these components are rooted in the AI HLEG trustworthy guidelines, which is one of the effective guidelines for TAIS development. However, Mittelstadt (2019) argues that the principled approach to ethical TAI can lead to substantively different requirements in practice as conceptual ambiguity gives room for context speculation for ethical requirements for AI development.

The AI HLEG guidelines explain that developing TAIS can aid individual flourishing and collective wellbeing by helping to generate prosperity, value creation, and wealth maximization. Therefore, it is important to ensure that their impact on human lives is fair and in line with uncompromised values. In addition, TAIS will enable the greater realization of the vast potential of AI technology and the benefits that they bring. (European Commission, 2019.) Based on the guidelines of AI HLEG and the IEEE global initiative, a framework developed by Vakkuri et al. (2019) identifies Trustworthiness as having a higher-level value that is produced by constructs in these guidelines geared towards creating ethical AIS that users can trust. Both the EU AIHLEG and the EAD guidelines by the IEEE organization are adopted for this study as constructs for TAIS development methods at the higher level.

3.1 EU Guidelines for Trustworthy AI

The EU AI HLEG guideline describes three basic principles of lawful, ethical, and robust (technical) that comprise TAIS. From these basic principles, they formed seven guidelines with practices for fulfilling these requirements. The guidelines are all equally important, support each other, and apply to different stakeholders (developers, deployers, end-users, and the broader society) involved in the AI system life cycle.

They are *Human agency and oversight, Technical robustness and safety, privacy and data governance, Transparency, Diversity, non-discrimination and fairness, Socie-*

tal and environmental well-being, and Accountability, as represented in figure 6 (European Commission, 2019.)

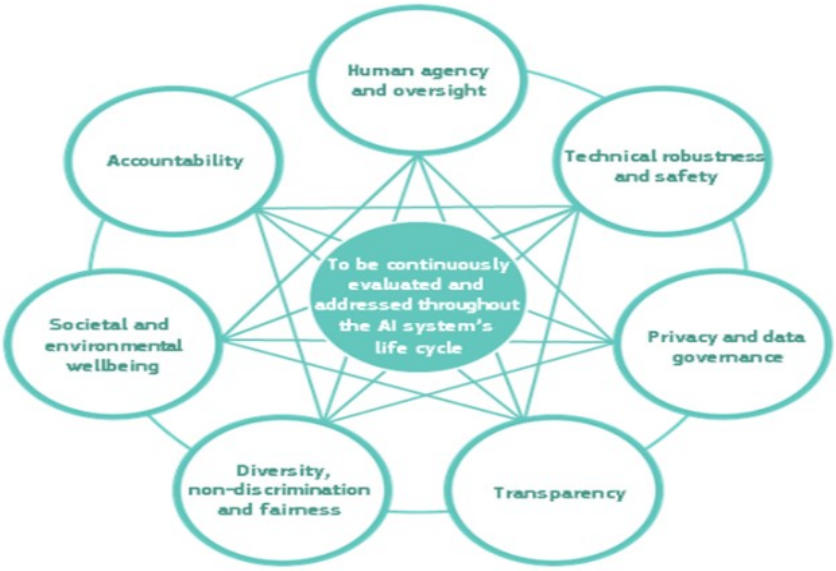

Figure 6 Representation of the seven requirements for Trustworthy AIS (European Commission, 2019)

- Human Agency and Oversight is based on the principle of respect for human autonomy and necessitates that AIS support human autonomy and decision making. It analyses that while AIS provides immense value in different application areas, it can also hamper or negatively affect fundamental human rights. Therefore, the fundamental risk assessment must be included in the design, development, and deployment (feedback mechanism) to help assess and mitigate usage. The guideline underlines the importance of human agents given the knowledge and tools to interact with AIS, make informed decisions, and, where applicable, challenge decisions with user autonomy being central to the system's functionality. Furthermore, it stresses the need for human oversight through governance mechanisms to ensure AIS does not undermine human autonomy or cause adverse effects. Governance mechanisms such as human-in-the-loop (HIL), human-in-command (HIC), and human-on-the-loop approaches give human agents the right to exercise oversight in line with their mandate. (European Commission, 2019.)
- Technical Robustness and Safety, grounded in the principle of prevention of harm, outlines the criticality of technical robustness. It emphasizes the importance of AIS being developed with a preventative approach to risk, reliably behaving as intended, and simultaneously minimizing unintentional and unexpected harm, and preventing unacceptable harm. This extends to AIS operating environments, any potential changes that

may occur, and the presence of other agents it interacts with. The guideline addresses the protection of AIS from adversaries that seek to exploit its vulnerabilities, options to make alternative decisions or shut down fully in the case of eventualities, securing AIS from abuse by malicious actors as well as unintended application. It denotes AIS having a fallback mechanism, whether human or artificial agents, should the need arise, providing the same output without negative outcomes. Processes that clarify and assess potential risks be established and relevant safety measures are developed and proactively tested. The accuracy of AIS is highly critical, and as such explicit and well-formed development processes that support, mitigate and correct unintended risks from inaccurate predictions be put in place. The guideline highlights the importance of AIS being reliable and reproducible, working properly with a range of inputs and situations. (European Commission, 2019.)

- Privacy and Data Governance emphasizes the importance of privacy. Privacy is a fundamental human right and underscores the prevention of harm to privacy, which encompasses data governance covering the integrity and quality of data used by AIS, its access protocols, and capabilities in processing data to ensure privacy is not compromised. Data and privacy must be guaranteed and protected throughout the AIS life cycle with the data not used unlawful or discriminatory. Data quality and integrity are highlighted here as of utmost importance. Data integrity must be ensured as well as documentation of all the necessary steps involved. Data protocols that govern data access should be in place, outlining access rights. (European Commission, 2019.)
- Transparency is closely linked to the principle of explicability. This guideline addresses data, systems, and business models of AI systems and is considered one of the most important principles. Transparency makes it possible to implement ethical principles in designing AIS and is listed as one of the key ethical principles in the EAD standards (Vakkuri et al., 2019). Creating an AI system to be transparent and traceable demands documentation to the best possible standard of datasets and processes that yield AIS decisions that could help with error tracking and detection. Transparency deals with Explainability and requires the technical processes and human decisions (application areas) supported by AIS to be understood and traceable by human beings with accuracy/explainability trade-offs being managed effectively and well documented to further enhance transparency. However, transparency is understood subjectively. Developers consider it as it pertains to algorithms and neural network (NN) architecture, and users view it from a less technical perspective (Vakkuri et al., 2019). Clear communication is emphasized, and human users of AIS should be made fully aware they are dealing with AIS without deception and ambiguity. Human users should be given the option to choose whom they choose to communicate with

and AIS capabilities and limitations made known. (European Commission, 2019.)

- Diversity, Non-discrimination, and Fairness must be a fundamental part of AIS and involves all affected stakeholders throughout the life cycle. As such, identifiable and discriminatory datasets that could lead to bias in AI systems should be eradicated in the collection phase. In addition, diversity in recruitment and oversight processes should be put in place to address systems purpose, constraints, requirements, and decisions clearly and transparently. AI systems should be user-centric and have a universal design allowing for use by everyone to gain access to their product or service irrespective of age, gender, race, abilities, or characteristics with acceptance and allowance for disabled persons engineered into their designs. Consultation of all relevant stakeholders (direct or indirect) is imperative, and solicitation of feedback is highly beneficial and strongly recommended throughout the AIs lifecycle. (European Commission, 2019.)
- Societal and Environmental Well-being guidelines emphasize the importance of considering the broader society, the environment, and other sentient beings as relevant stakeholders throughout the lifecycle of the AIS in line with the principle of fairness and prevention of harm. AIS that are environmentally friendly and in line with sustainable development goals be developed, deployed, used, and assessed in this regard, and measures that encourage such implementation. The ubiquitousness of AIS exposes them to virtually all social systems. While they enhance social skills, they are also associated with deterioration, especially in mental and physical wellbeing. In addition, they require effective monitoring of their effects. AIS usage should therefore be given careful consideration in matters of institutions, democracy, and society at large as it applies to electoral and political decision-making contexts. (European Commission, 2019.)
- Accountability compliments all the other guidelines and is linked to the principle of fairness. It requires that mechanisms be established for AIS that ensure responsibility, accountability, and for outcomes both before and after development, deployment, and usage. It highlights the need for AIS to be auditable, enabling the assessment of algorithms, datasets, and design processes by auditors with care taken to safeguard information regarding business models and intellectual property. Accountability encourages reporting of the negative minimization impact of AIS so that requisite action can be taken to ensure that reporters are duly protected should the need arise. The use of impact assessment that is proportionate to the risk the AIS poses should be in place. The guideline delineates tensions that arise in the implementation of AIS to be addressed rationally and methodologically. Accountability of relevant interests and values implicated by AIS should be identified and, where conflicts arise, trade-offs explicitly acknowledged and evaluated in terms of their risks to ethi-

> cal principles. Where no ethically acceptable trade-off is identified, the AIS is not continued with all decisions duly documented. It emphasizes that provision be made for redress, and in the event of an occurrence, adequate mechanisms are in place that prioritize vulnerable persons or groups. (European Commission, 2019.)

The guidelines recommend that trustworthiness be considered a prerequisite for people and society in developing, deploying, and using AIS (European Commission, 2019.) While the EU guideline does not explicitly deal with the lawful aspect of TAIS, it is expected that all systems are developed in accordance with the laws by which they are governed. The guideline explains that TAIS is expected to be aligned with ethical principles and technically robust in its design to instill confidence in users. (AI HEG, 2019.) However, each of these components is insufficient by itself, and an approach is required to balance tensions that may arise in their implementation (AI HEG, 2019).

3.2 Ethically Aligned Design

The Ethically Aligned Designs (EAD) (IEEE Standards Association [IEADe1], 2019) is considered one of the bedrocks for ethical design. This study explores its ethical constructs in its approach for developmental methods of TAIS at a higher level. The IEADe1 comprises eight guidelines of *Human Rights, Well-being, Data Agency, Effectiveness, Transparency, Accountability, Awareness of Misuse, and Competence* (IEEE Standard Association, 2019) and is represented in figure 7.

Figure 7 Representation of the general principle of ethically aligned design (Ead1, 2019)

- Human Rights denotes the creation, operation, and management of AIS to respect, promote, and protect internationally recognized human rights.
- Well-being denotes AIS creators and developers adopting human well-being as a key and primary success criterion.
- Data Agency denotes AIS developers ensuring the empowerment of individuals in accessing and securely sharing data to maintain their capacity to have control over their identity.
- Effectiveness denotes that the creators and developers of AIS provide evidence showing their efficacy and fitness for their purpose.
- Transparency denotes the need for the basis of a particular AIS decision to be discoverable.

- Accountability denotes AIS being created and operated to provide a clear rationale for all decisions made.
- Awareness of Misuse denotes AIS creators guard against all potential misuse and risks in operation.
- Competence denotes AIS creators specify, and operators adhere to knowledge and skill required for safe and effective operation. (IEADe1, 2019.)

The IEADe1 (2019) guidelines extensively reflect ethical standards, which align with the EU requirements for TAIS.

3.3 ECCOLA Method

ECCOLA method has been proposed as a possible solution to creating higher-level design and development of TAIS. ECCOLA is a tool for developers and product owners that seeks to implement AI ethics practically at a higher level and development decision level. Methods such as the Ethical framework for designing autonomous intelligent systems (Leikas, Koivisto & Gotcheva, 2019) focus on design at a higher level than development. They are not explicitly targeted at product owners and developers. (Vakkurri et al., 2020.)

Built on AI ethics research, ECCOLA utilizes existing theoretical research, conceptual research, and ethics guidelines from both the IEEE EAD and EU AI HLEG. ECCOLA is developed based on the Essence theory of software engineering and utilizes the philosophy of essentializing software engineering practices based on Jacobson et al. (2012) and utilizes cards to describe methods. Part of ECOLLA's development is iteratively based on the Cyclical Action Research (CAR) approach and aims to bridge the gap between research and practice in the field of AI ethics. ECCOLA is method agnostic and modular in practice and is presented as a deck of physical cards primarily due to the Essence of the theory software engineering approach. However, the use of cards as a method is

not uncommon as other methods such as Kanban employ a similar approach. (Vakkuri et al., 2020.)

ECCOLA method comprises an A5-sized game sheet that explains how the method and 21 cards. The cards are - Stakeholder Analysis, Types of Transparency, Explainability, Systems Reliability, Traceability, Documenting Trade-offs, Communication, Privacy and Data, Data Quality, Access to Data, Human Agency, Human Oversight, System Security, System Safety, Accessibility, Stakeholder Participation, Environmental Impact, Societal Effects, Auditability, Ability to Redress, and Minimizing Negative Impacts. The cards are split into eight themes, with each theme made up of one to six cards.

The method is made up of an analysis card (*analyze*) that evaluates the project and considers all the potential stakeholders. It also incorporates seven guidelines from the IEEE guidelines and the EU guidelines. The principles are *Transparency*, *Data*, *Agency 'and Oversight*, *Safety and Security*, *Fairness*, *Wellbeing*, and *accountability* (Vakkuri et al., 2020). Figure 8 shows the ECCOLA model.

Figure 8 Representation of ECCOLA (Vakkuri et al., 2020)

- *Analyse* Creates an awareness in the developers of the AIS regarding the stakeholder, including the end-users. It aids developers in identifying who the system will affect and how in terms of all the different stakeholders in different capacities.
- *Transparency*: This is arguably considered one of the most central and foundational ethical principles and is featured in both the IEEE and EU guidelines and involves understanding the AIS. ECCOLA explores the principle of transparency on various levels. On the basic level, it calls for developers of AIS to analyze who they are transparent with and what they are transparent about. On another level, ECCOLA analyses transparency based on **explainability.** It enables developers to question the explainability of AIS because if stakeholders are unable to understand

the actions of the system, then they will be unable to trust them. Transparency is analyzed from the **systems reliability** perspective and makes developers understand the need for stakeholders to understand how the system works and what influences its decisions. **Traceability** is explained in terms of transparency and entails developers consider stakeholders' need to understand why the AIS acts the way it does. Traceability is essential as it can help in discovering errors in the system as such documentation in terms of development, testing, and validation will enable transparency and ultimately trust in the system. **Documentation of trade-offs** is analyzed in terms of transparency. ECCOLA details the need for documenting decisions made or chosen over others and the importance of the reasoning behind why these decisions were chosen over others. **Communication** is analyzed under transparency and noted as instrumental in generating trust with stakeholders if it is clear and explains goals and relevance as it relates to them.

- *Data*: Data usage is extremely sensitive and is covered under the IEEE and EU guidelines by Data Governance and Data Agency. It includes protection of data throughout its lifecycle and awareness to users of their data being collected as protected under the GDPR regulation. ECCOLA explores the development of ethical AIS from the data perspective based on the quality of data and raises awareness for the developers on the integrity and quality of data being used. The practices enable developers to understand that the data used are in alignment with the system's goals. ECCOLA also explores data access and asks developers questions on who has access to collected data, the context, and the intended usage.
- *Agency and Oversight*: ECCOLA explores this principle through the concept of human agency and asks developers questions in human-machine interaction. It underlines the need for understanding by human users, the working and decisions made to support human decision-making, and allow humans to make their own decisions. ECCOLA examines human oversight by raising questions for developers concerning the measure of support AIS offers human users and if the system undermines human autonomy by overriding their decisions.
- *Safety and security*: ECCOLA method explores this principle from the perspective of system security and system safety. It brings to the awareness of developers such ethical issues as the potential forms of attacks the systems could be vulnerable to, their uniqueness and relevance to AIS, and the different risks and consequences cyber-physical AIS pose to the physical world.
- *Fairness*: ECCOLA explores fairness from the EU guideline of diversity, non-discrimination, and fairness. It enables developers to ask questions on if AIS allows for equal access in terms of a wide range of users regardless of disabilities and diverse groups, including the non-technically savvy groups of users. ECCOLA analyses stakeholder participation in the development and asks developers if the different stakeholders, especially

the target end-users, are included as part of the development of the system.

- *Wellbeing*: ECCOLA explores the ecological impact the development of AIS has on the environment and asks developers questions in this vein, making them examine the energy sources of data centers and their impact. Wellbeing is analyzed from the social effect perspective and asks developers questions to evaluate the broader impact of AIS usage beyond the immediate target users and the systemic effect it could have on society overall.
- *Accountability*: This is explored from auditability and examines the effect of audit on data and AIS stressing the importance of having an audit system in place for accountability should the need arise. Accountability is explored by creating awareness in developers of AIS in terms of the trust and peace of mind (redress) for stakeholders that can be enhanced if they understand that compensation exists for them from any harmful usage or impact arising from usage. The last aspect of accountability explored by ECCOLA is in the minimization of negative impacts from the usage of the system. ECCOLA asks developers questions arising from an overall perspective and explores them as it relates to the distinct perspective of the area of concern. (Vakkuri et al., 2020.)

ECCOLA method supports iterative development where each team can choose a theme or card that is relevant to the current iteration. Depending on the user, ECCOLA can assist product owners in creating non-functional user stories involving ethics. For developers, it facilitates communication on ethical issues that can translate to ethical decisions being made. Overall, it generates or creates an awareness of ethical issues that would otherwise not be realized. (Vakkuri et al., 2020.) The method has a threefold goal: the **creation of awareness for the importance of AI ethics (which forms the foundation for this study);** creation of an adaptable modular method that is suitable for a wide variety of software engineering (SE) context; and creation of an AI ethics method suitable for agile development. (Vakkuri et al., 2020.) Each card deals with an encapsulated theme in-depth, as represented in figure 9. Each card is split into three parts: ***Motivation-*** which deals with why the concept is important; ***What to do -*** which deals with tackling the problem; And ***Practical example*** of the topic that makes the issue more tangible. In addition, the cards are designed with a note-making space to enable users to process their thoughts and document accordingly. (Vakkuri et al., 2020.)

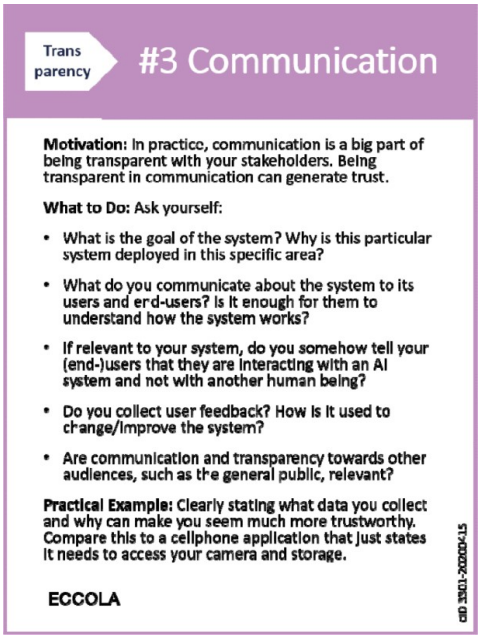

Figure 9 ECCOLA card illustrating Transparency construct (Vakkuri et al., 2020).

3.4 IDENTIFYING THE RESEARCH GAP IN ECCOLA

In line with the goal of this study to improve ECCOLA's robustness and widespread adoption by critically evaluating and analyzing its tenets, a conceptual framework by Tolvanen (2020) is employed to highlight areas within ECCOLA that display any perceived vulnerabilities for correction and improvement. This evaluation could lead to an extension of the method and serves as the problem-centered approach or problem-centered initiation in DSRM (Peffers et al., 2007).

3.5 Framework for Evaluation of Trustworthy AI

The framework is based on suggested requirements for both technical and non-technical methods towards the realization of TAIS (European Commission, 2019). The conceptual framework by Tolvanen's (2020) thesis is established based on 13 technical and non-technical suggested practices in the AI HLEG. The practices help implement Ethical guidelines in the development of TAIS in all stages of their lifecycle. They are based on the premise that the realization of TAIS is a continuous process that is constantly evolving in a dynamic environment. (European Commission, 2019.)

3.5.1 Technical Methods

The technical practices list five practices (table 3). They include Architecture for TAIS, Ethics and the rule of law by design (X-by-design), Explanation methods, Testing and Validating, and Quality of service indicators to help ensure that TAIS practices are integrated with the design, development, and use phases of an AIS. The method varies in terms of maturity and includes (European Commission, 2019).

Table 3 Technical Methods for the development of TAIS (European Commission, 2019)

Method	Description
Architectures for TAIS	Requirements for TAIS procedures should be implemented directly into the AIS architecture lifecycle.
Ethics and Rule of Law by design (X-by-design)	Development methods should implement ethical norms and legislation in the design of AIS.
Explanation Methods	Development methods should facilitate an explanation of the underlying mechanisms and outputs of AIS.
Testing and Validation	TAIS development methods should allow for the use of novel testing methods because of their non-deterministic nature. Early validation of data is recommended to help with this.
Quality of Service Indicators	Development methods should reflect the appropriate quality of service indicators to ensure there is a baseline understanding to determine if AIS has been tested and developed with safety and security in mind.

3.5.2 Non-technical Methods

The non-technical methods (table 4) described by the guidelines of the European Commission (2019) lists eight practices that can help provide a significant role in securing and maintaining TAIS on an ongoing basis. They include Regulation, Codes of conduct, Standardisation, Certification, Accountability via Governance frameworks, Education, and awareness to foster an Ethical mindset, Stakeholder participation ‘and social dialogue, and Diversity and Inclusive design teams. (European Commission, 2019.)

Table 4 Non-technical Methods for the development of TAIS (European Commission, 2019)

Method	**Description**
Regulation	Development methods should feature regulatory measures that support TAIS.
Code of Conduct	Stakeholders involved in TAIS development should adapt their corporate responsibility charter, Key performance indicators, codes of conduct, or internal policy documents to reflect striving towards TAIS.
Standardization	TAIS developers should strive to align standards employed to existing standards as a quality management check for AI stakeholders. Pending when a trustworthy AI label becomes available.
Certification	The development of TAIS should include certifications to the broader public on the transparency, accountability, and fairness of AIS. The certifications can employ standards from different application domains that appropriately align with industry and society standards of different contexts.
Accountability via Governance frameworks	The development of TAIS should include governance frameworks both internally and externally to ensure ethical dimensions of associated decisions by assigning governing personnel in charge of ethical issues or an ethical panel of the board to provide oversight.
Education and awareness to foster an ethical mind-set	The development of TAIS should encourage informed participation of all stakeholders. This practice can ensure that communication, education, and training of the potential impact of the systems is widespread and for fostering the basic literacy of AIS.
Stakeholder participation and Social dialogue	The development of TAIS should actively seek participation and dialogue from all stakeholders on the use and impact of AIS to support the evalua-

	tion of results and approaches.
Diversity and inclusive design teams	The development of TAIS should reflect the diversity of users and society in general as it contributes to objectivity, consideration of different perspectives, needs, and objectives. Diversity should be based on gender, culture, age, professional background, and skillsets.

3.5.3 Evaluation framework

The evaluation framework adopted is a conceptual framework by Tolvanen (2020) to identify areas with contributions in practice regarding TAIS development. The framework matches practices (both technical and non-technical) with tenets of ethical practices to identify areas where they are applied and where they are not. In ECCOLA, the framework will be used to evaluate and identify areas of vulnerability by matching the trustworthy tenets in ECCOLA with practice areas. This is to help determine how robustly the method tents incorporate ethical principles that translate to the development of TAIS and instill trust in users.

This can help identify areas that can be corrected to extend the method and make it more robust for wider adoption. For this study, the framework is modified to match ECCOLA's trustworthy tenets with the requirements of the method of the AI HLEG guidelines.

3.5.4 Evaluation

ECCOLA method is made up of an "analyze card that evaluates the project and considers all the potential stakeholders, and incorporates seven guidelines from the IEEE guidelines and the EU guidelines. The principles are *Analyse Transparency, Data, Agency, and Oversight, Safety and Security, Fairness, Wellbeing, and accountability* (Vakkuri et al., 2020). Analyse helps to creates an awareness in the developers of the AIS regarding the stakeholder, including the end-users. It aids developers in identifying who the system will affect and how in terms of all the different stakeholders in different capacities. Development practices that can contribute to this realization are presented in Table 5

Table 5 Analyse Practices with contributions for trustworthy AIS development (Tolvanen, 2020)

Development practice	**Relation**
Code of Conduct	ECCOLA promotes documentation of ethical practices towards developing TAIS
Education and awareness to foster an ethical mind-set	ECCOLA promotes education and creates awareness on ethical practices from all stakeholders
Stakeholder participation and Social dialogue	ECCOLA fosters active participation and social dialogue from all stakeholders on the use and impact of AIS in society.
Diversity and inclusive design teams	ECCOLA facilitates inclusion in the design for the different stakeholders.

Transparency is considered one of the most central and foundational ethical and trustworthy principles and is featured in both the IEEE and EU guidelines and involves understanding the AIS. ECCOLA looks at the principle of transparency on different levels. Transparency is explored in terms of the types of transparency, Explainability, Traceability, Documenting trade-offs, and Documentation. (European Commission, 2019). Development practices that can contribute to this realization are presented in Table 6

Table 6 Transparency Practices with contributions for TAIS (Tolvanen, 2020)

Development practice	**Relation**
Explanation Methods	ECCOLA facilitates the explanation of underlying mechanisms in the development of TAIS.
Code of Conduct	ECCOLA aids clear guidelines for the documentation of intentions to establish transparent operations.
Ethics and Rule of Law by design (X-by-design)	ECCOLA aids the implementation of ethical norms and legislation in the design of AIS.
Quality of Service Indicators	ECCOLA helps to reflect the appropriate quality of service indicators to ensure there is a baseline understanding that AIS has been tested and developed with safety and security in mind.
Certification	ECCOLA supports the transparency, accountability, and fairness of AIS that can lead to certifications from different application domains that appropriately align with industry and society standards of different contexts.

Data governance and Data agency is explored in ECCOLA in terms of Privacy and Data, Data Quality, and Access to Data (Vakkuri et al., 2020). ECCOLA explores the development of TAIS in terms of data based on the quality and integrity of data and to ensure alignment with the system's goals. Development practices that can contribute to this realization are presented in Table 7.

Table 7 Data governance and Data agency Practices with contributions to TAIS (Tolvanen, 2020)

Development practice	**Relation**
Testing and Validation	ECCOLA encourages practices for effective testing and validation of data.
Accountability via Governance frameworks	ECCOLA fosters practices for data management and governance.
Quality of Service Indicators	ECCOLA fosters practices for data management and governance.
Standardization	ECCOLA encourages practices that align with relevant standards or widely accepted protocols.
Regulation	ECCOLA fosters practices aimed at regulating and managing data such as GDPR.

Agency and Oversight are explored in ECCOLA through the concept of human agency, human oversight, and the EAD principle of human rights that aims to develop AIS that in their operation respect, promote, and protect human rights instill trust in users and trustworthiness of the system. Development practices that can contribute to this realization are presented in Table 8

Table 8 Agency and Oversight Practices with contributions to TAIS (Tolvanen, 2020)

Development practice	**Relation**
Architectures for TAIS	ECCOLA promotes the design and practices in the design process that lead to the development of TAIS.
Ethics and Rule of Law by design (X-by-design)	ECCOLA promotes ethics and the rule of law design for the development of TAIS
Explanation Methods	ECCOLA promotes practices that foster explanations made by the AIS.
Education and awareness to foster an ethical mind-set	ECCOLA fosters practices that educate stakeholders through ethical processes
Stakeholder participation and Social dialogue	ECCOLA fosters transparent and explainable practices, especially end-users, to help inform decision processes.
Quality of Service Indicators	ECCOLA fosters practices for the safety and security of human end-users.

Safety and Security are explored in ECCOLA from system security and system safety, which stems from the HLEG principle of technical robustness and safety and the EAD principle of competence. Development practices that can contribute to this realization are presented in Table 9

Table 9 Safety and Security Practices with contributions to TAIS (Tolvanen, 2020)

Method	**Relation**
Architectures for TAIS	ECCOLA promotes for trustworthy practices to be implemented in the architecture of AIS.
Ethics and Rule of Law by design (X-by-design)	ECCOLA promotes ethics and the rule of law design for the development of TAIS
Explanation Methods	ECCOLA promotes security and safety practices for and by the AIS.
Testing and Validation	ECCOLA encourages practices for adequate testing, validation of data to eliminate vulnerabilities.
Quality of Service Indicators	ECCOLA fosters practices for the safety and security of stakeholders, particularly human end-users.

ECCOLA explores fairness from the EU guideline of diversity, non-discrimination, and fairness. This tenet encourages the development of AIS that are all-inclusive and equal access in their operation and do exclude different parties or groups, which creates a sense of belonging to users and increases their trustworthiness. Development practices that can contribute to this realization are presented in Table 10.

Table 10 Fairness Practices with contributions to TAIS (Tolvanen, 2020)

Method	**Relation**
Ethics and Rule of Law by design (X-by-design)	ECCOLA promotes ethics and the rule of law design for the development of TAIS
Explanation Methods	ECCOLA fosters practices that explain the various stakeholders' involvement in the development of AIS.
Code of Conduct	ECCOLA promotes practices that reflect stakeholders' involvement in AIS development that include their

	responsibilities, Key performance indicators, and internal policy documents to reflect striving towards TAIS.
Standardization	ECCOLA promotes practices that strive towards meeting existing standards in terms of fairness.
Education and awareness to foster an ethical mind-set	ECCOLA promotes practices towards educating and creating an ethical mindset for stakeholders.
Stakeholder participation and Social dialogue	ECCOLA actively promotes practices for active stakeholder participation and social dialogue.
Diversity and inclusive design teams	ECCOLA actively promotes diversity to cover all spheres.

Wellbeing is explored through the principle of societal and environmental wellbeing of the EU guidelines and the Wellbeing principle of the EAD. Wellbeing is analyzed from the social effect perspective evaluates the broader impact of AIS usage beyond the immediate target users and the systemic effect it could have on society overall. Development practices that can contribute to this realization are presented in Table 11.

Table 11 Wellbeing Practices with contributions to TAIS (Tolvanen, 2020)

Method	**Relation**
Architectures for TAIS	ECCOLA promotes practices for the technology of AIS architecture geared towards the realization of TAIS.
Regulation	ECCOLA promotes practices for measures that promote environmental awareness of AIS geared towards the development of TAIS.
Education and awareness to foster an ethical mind-set	ECCOLA promotes practices that educated stakeholders on the environmental and societal effects of AIS geared towards TAIS development.
Ethics and Rule of Law by design (X-by-design)	ECCOLA promotes ethics and the rule of law design practices.
Stakeholder participation and Social dialogue	ECCOLA actively promotes environmental and societal practices geared towards the development of TAIS.

Accountability is explored in ECCOLA in terms of auditability, ability to redress, and minimizing negative effects. Development practices that can contribute to this realization are presented in Table 12.

Table 12 Accountability Practices with contributions to TAIS (Tolvanen, 2020)

Method	Relation
Regulations	ECCOLA fosters regulatory practices that can lead to an effective audit of the AIS geared towards TAIS development.
Architectures for TAIS	ECCOLA promotes practices for the architecture of AIS that can lead to an effective audit geared towards TAIS development.
Explanation Methods	ECCOLA promotes explanation practices that can lead to the effective audit of AIS geared towards TAIS development.
Quality of Service Indicators	ECCOLA promotes Quality of service practices that can lead to an effective audit of AIS geared towards the development of TAIS.
Code of Conduct	ECCOLA promotes practices for an effective code of conduct of AIS, which can be instrumental in audit situations and geared towards TAIS development.
Standardization	ECCOLA promotes awareness of Standardisation practices that can aid audit of AIS towards the development of TAIS.
Certification	ECCOLA promotes awareness for certifications practices that can aid in the audit of AIS towards TAIS development.

The evaluation is formed into a primary conceptual conclusion, as shown in figure 10, where colored squares represent contributions to practices and uncolored squares represent areas of less contribution.

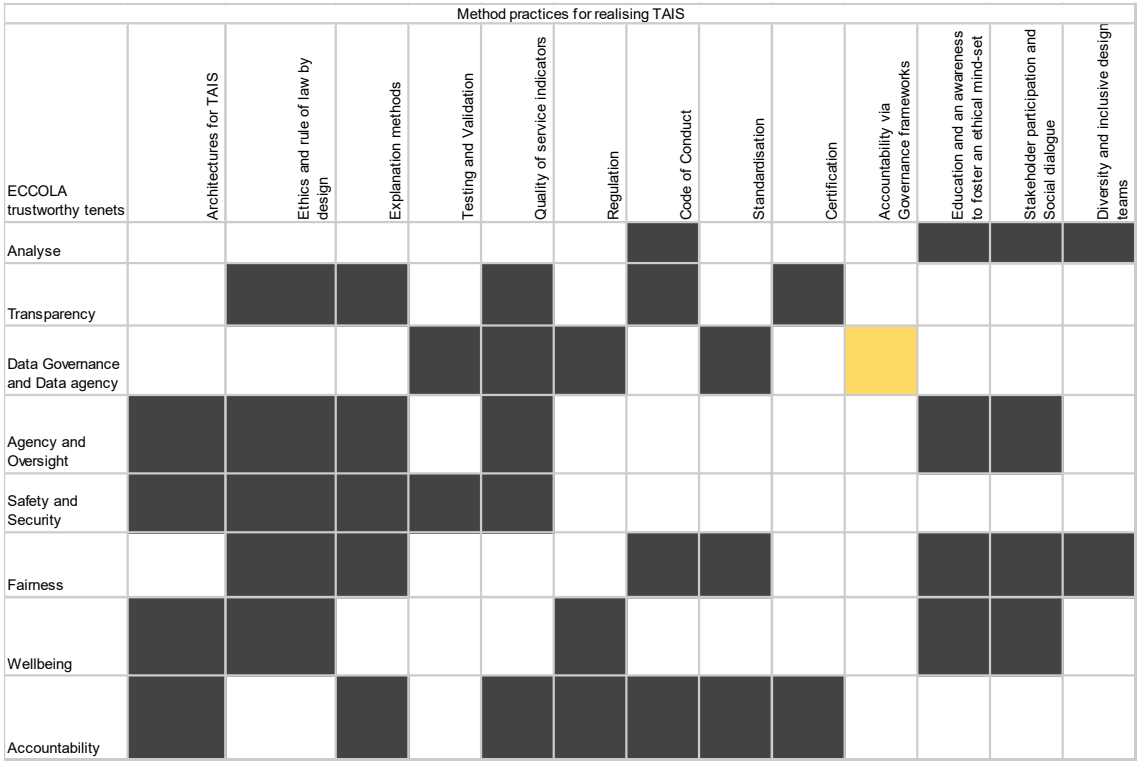

Figure 10 Primary conceptual conclusions showing ECCOLA evaluation against expected practices (Tolvanen, 2020)

In figure 9, ECCOLA infuses most ethical principles as part of the requirements for developing TAIS in method practices. It displays at least two colored squares in each category. However, it sparingly addresses the accountability via Governance frameworks as it only featured one colored square in that category.

Governance frameworks are governance structures that mirror interconnected relationships, factors, and other influences in an institution (Williamson, 1984). They usually comprise a conceptual structure and sets of rules that outline how an organization manages and controls its assets to perform at an efficient level—their influence cuts across all spheres of an organization. (Williamson, 1984). Governance frameworks are vital for directing interactions across organizations, stakeholders, regulatory bodies, and the general operations within organizations (Williamson, 1984).

Several governance frameworks exist, such as corporate governance frameworks, information security frameworks, information governance frameworks, and data governance frameworks (Veiga & Eloff, 2007), but due to the close association of information assets (IA) with development methods like ECCOLA, Information Governance framework will be explored in this study. Information Governance framework can help focus on sensitive practices of IA with development methods like ECCOLA (Veiga & Eloff, 2007), which can help improve the robustness of ECCOLA and extend it for developing TAIS instill confidence in users (Hamon et al., 2020). In addition, IG practices can help to instill trust in users by assuring users that their information is being managed in line with governing guidelines which provides confidence that their infor-

mation will not be accessed, mismanaged, or used in unethical ways. Therefore, it is critical further to explore the tenets of ECCOLA alongside IG principles to determine how incorporated these practices are in the method.

It is important to note that while the evaluation has highlighted other method practices such as certifications and testing and validation as having only two squares, these practices are outside the scope of this study.

4 INFORMATION GOVERNANCE

The EU guidelines recommend accountability via a governance framework to help with the robustness of TAIS (AI HLEG, 2019). It explains the importance of a governance structure both internally and externally to help with accountability of the ethical dimension associated with TAIS development (European Commission, 2019). According to Dafoe (2018), developing TAIS is closely related to its governance, with both helping humanity develop beneficial AIS with governance specifically focused on institutions and context within AIS are developed. Governance policies that constitute Information Governance (IG) for AIS seek to maximize the odds that people building and using them have the goals, incentives, worldview, time, training, resources, and the support to do so for the benefit of humanity (Dafoe, 2018). To this end, IG is explored to ensure that its practices are incorporated in ECCOLA.

4.1 Definition

Information Governance is a fundamental subset of corporate governance. It is described by Lomas (2010) as putting in place information governance programs that ensure information is controlled and appropriately made available without compromising its security. Bennet (2017) defines Information Governance as comprising the activities and technologies organizations employ to maximize the value of information and minimize associated risks and costs. Borgman, Heier, Bahli & Boekamp (2016) analyze IG as a framework that contains mechanisms for guiding the creation, collection, storage, analysis, use distribution, and deletion of information relevant to the business to achieve value creation. One of the most popular and all-encompassing definitions of IG is by Gartner as defined by Logan (2010) as

> "The specification of decision rights and an accountability framework to ensure appropriate behavior in the valuation, creation, storage, use, archiving, and deletion of information. It includes the processes, roles, and policies, standards, and metrics that

> ensure the effective and efficient use of information in enabling an organization to achieve its goals. (para.4)"

According to Kooper, Maes & Lindgreen (2011), IG is a logical alternative that focuses on seeking and finding creation and use and exchanging information, not solely based on its production. They further explain that IG answers the question "what information do we need, how do we make use of it, and who is responsible for it?" (Kooper et al., 2011, p. 196).

IG deals with information management throughout its lifecycle from creation to deletion and employs structural, procedural, relational mechanisms, decision rights, security, and privacy in its governance. Structural mechanisms include reporting structures, accountabilities, and governance body landscapes; procedural mechanisms consist of policies, processes, standards, and protocols that aim for information to be managed in line with proper guidelines; and relational mechanism deals with the collaboration between the various stakeholders. (Borgman et al., 2016.)

Information Governance is often confused with Data Governance (DG), a subset of IG that aims to control Information at the data level and ensures the maintenance of accurate, high-quality data by implementing appropriate systems and processes. Information is obtained from processing data which represents information in its basic form. (Bennet, 2017.) According to Borgman et al. (2016), information depends on contextualization and subjective interpretations, while data is composed of facts and raw numbers. Information is an unusual good that can function as both an end-to-end product or as an input into the creation of other information comprising several unique characteristics, which makes it difficult to evaluate and govern (Kooper et al., 2011).

Tallon Ramirez and Short (2013) explain that with data being transformed into information, a need exists for the realization that the value information provides is partly due to how it is governed over its lifecycle. However, independent of its content, the generic principles in understanding the value and governance of information can be recognized (Kooper et al., 2011). Kooper et al. (2011) analyze that the basis for the governance of information is the interaction concept. They explain that actors within certain environments have many interactions and do not possess the knowledge required to solve the complex, dynamic, and diversified challenges that come with these interrelations. As such, a governance approach is required for streamlining the patterns of interactions to create an understanding or make sense of the value of information created specifically in information exchange. In addition, within interactions, information becomes subjective as different actors give different interpretations to information based on their consumption or generation; as such, the governance of information should include human interactions of actors with people, data, and underlying systems (Kooper et al., 2011.)

Information Governance helps to provide a balance between the risk associated with information and the value that information provides. This balance emanates from IG covering the length, reason, and efficiency of use of information, usually requiring the participation of numerous stakeholders and aids

with legal compliance, operational transparency, and reduction of expenditure associated with legal discoveries. (Lomas, 2010.)

Lomas (2010) explains that organizations can organize logical and consistent frameworks to handle information using IG principles and procedures whereby they guide the proper use behavior on handling electronically stored information. Kooper et al. (2011) argue that IG may not be restricted to a particular framework since information does not restrict itself to the boundaries of an organization and may vary from policies, a way of working, the creation of space within a predefined settlement such as online communities or a strict framework such as privacy regulations. While the importance of IG cannot be overemphasized, it needs to be embedded as part of workplace culture to mitigate culture shock and organizational relationships that are not conducive to the division of labor required by IG (Hagmann, 2013).

The Information governance initiative's annual report of 2014 clarifies that IG goes beyond retention and disposition (Information Governance Initiative, 2014) and has an extensive reach across every aspect of the organization. All areas of the organization that utilizes data incorporating Risk & Compliance, Cybersecurity, Privacy/Data Protection, Records & Information Management, Data Governance, eDiscovery, and Data analytics employ policies, procedures, technology, and people in its implementation illustrated in figure 11. (Bennet, 2017.)

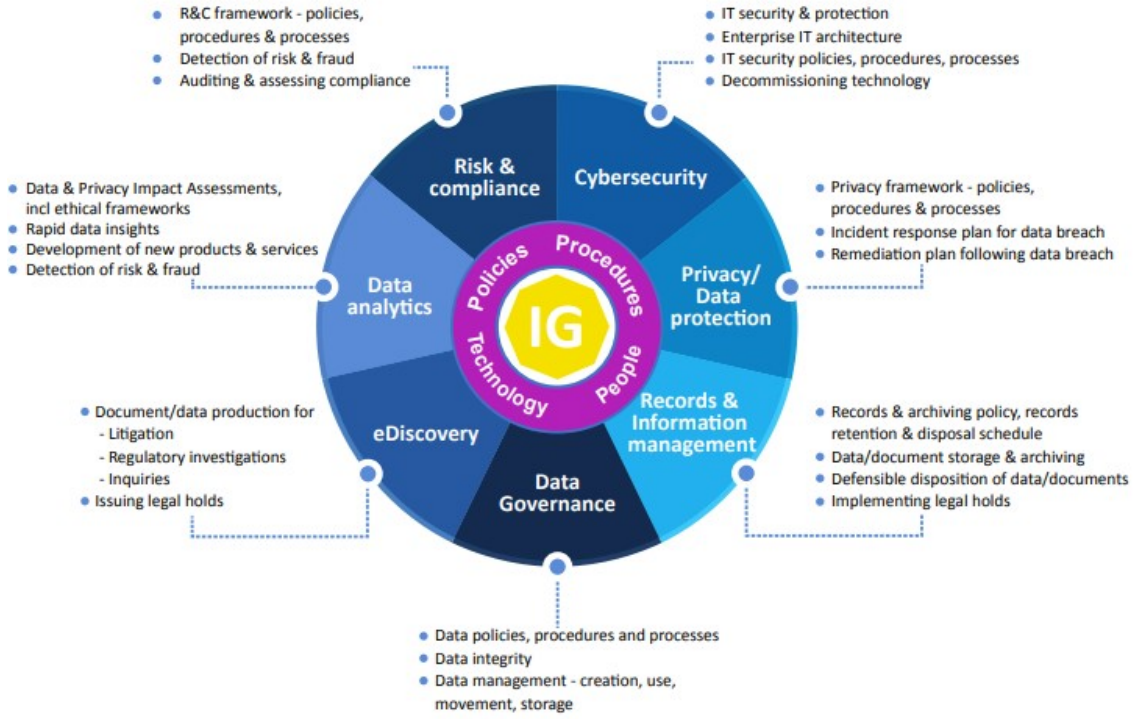

Figure 11 Basic overview of Information Governance framework (Bennet, 2017)

IG requires strategic thinking, leadership, and alignment of IG policies, procedures, technologies, and people across the organization. It also requires a collaborative and proactive culture that employs best practice IG frameworks across the different organizational silos (Risk & Compliance, Cybersecurity, Privacy/data protection, Records & Management Information, Data Governance, eDiscovery, and Data analytics) for effective implementation and outcomes to deliver results. Lastly, IG requires accountability where results are measured and audits carried out (Bennet, 2017.)

4.2 Information Governance frameworks

Notable IG frameworks exist for assessing and measuring information management and Government maturities, such as the **Information Governance model** and the **Information Lifecycle Management** (Hagmann, 2013). However, **Principles®** or **Generally Accepted Recordkeeping Principles® (GARP®)** by ARMA (2009) and its upgraded maturity model (2017) is identified as a globally accepted standard that identifies critical hallmarks and high-level framework of good practices for information governance.

4.3 Generally Accepted Recordkeeping Principles®(GARP®) by ARMA

GARP® is popular in the field of IG, and its principles provide organizations of all sizes, sectors, industry types, and geographic boundaries with standards of conduct for governing information as well as guidelines. GARP® provides guidance on information management and governance of record creation, organization, security maintenance, and other activities used to effectively support the recordkeeping of an organization (ARMA, 2009).

Record is defined as information that is recorded, received, or produced in the initiation, conduct, or completion of an organizational or individual activity comprising content, context, and structure sufficient to provide evidence of the activity (International Council on Archives [ICA], 2005). ICA (2005) further analyzes records as comprising content which could be data or information, contextual, that is how it relates to other records and to the organization that created it, and structural referring to an inherent logic to the information it contains and the metadata likely to define its context. Records can be physical or in digital forms such as databases, applications, and emails (ICA, 2005). The advent of the internet and the big data explosion has transformed and transitioned records management from the traditional practice to a larger and more effective information governance platform (ICA, 2005).

GARP® comprises eight principles - *accountability, transparency, integrity, protection, compliance, availability, retention, disposition, and a maturity model made up of five models (sub-standard, in Development, Essential, Proactive, and Transformational) that describe* characteristics typical for each level of maturity. ARMA principle describes and measures fundamental attributes of IG, is applicable to all classes of organizations and industries, and independent of local laws and customs. (ARMA, 2009.) However, the scope of this study requires the analysis of GARP® and not the maturity model.

- Accountability: Requires that a senior executive oversee information governance programs and delegate responsibilities accordingly for information management, policy, and procedure adoption that guide personnel and ensure auditability.
- Transparency: Requires that businesses and organizations document their activities and processes in an open and verifiable manner and documentation be made available to personnel and appropriate interested parties.
- Integrity: Requires that IG programs be constructed to reflect the authenticity and reliability of information assets generated or managed.
- Protection: Requires that IG program be constructed to ensure the appropriate level of protection for information assets in terms of their privacy, confidentiality, privilege, secrecy, classified, and how essential they are to business continuity or that otherwise require protection.
- Compliance: Requires that IG programs be constructed to comply with applicable laws, binding authorities, and organization policies.
- Availability: Requires that IG be exercised in information assets in a manner that ensures timely, efficient, and accurate retrieval.
- Retention: Requires that IG be exercised in a manner consistent with organizations maintaining its information assets for an appropriate period considering its legal, regulatory, fiscal, operational, and historical requirements.
- Disposition: Requires that IG be exercised in organizations to provide secure and appropriate disposal of information assets that no longer require to be maintained in compliance with organizational laws and policies.

(ARMA, 2009.)

A review of the literature reveals GARP® to be de facto IG principle with the widest Adoption. While mentioned in studies, the other IG principles pale in comparison to the adoption and usage exhibited by GARP®. For this study's purpose, the GARP® by ARMA is adopted as it appears to currently being the most robust IG practice in literature and practice.

5 EMPIRICAL ANALYSIS

This analysis aims to evaluate the trustworthy tenets of ECCOLA found in the 21 cards with GARP®) to identify how much each card contributes to IG practices. In accomplishing this, each ECCOLA card is analyzed against a corresponding GARP® principle.

5.1 Empirical method

Content analysis is used for analyzing the IG practices in GARP®, with each of the ECCOLA cards. Content analysis is a research tool that helps to determine themes, concepts, and specific words within a given qualitative data (Weber, 1990). The content analysis method enables the evaluation of language used within data such as text to search for a bias, enabling the researcher to make inferences within the text. It also helps to reduce data to concepts to describe the research phenomenon by creating categories. Arguments exist that content analysis can be subjective and reductive; however, they provide a great deal of flexibility, are transparent and replicable. (Elo et al., 2014; Colombia, n.d.) Sources of data that can be employed in Content Analysis include interviews, field research notes, and communicative languages such as books, guidelines, reports, and journals (Elo et al., 2014), making it the most suitable option for this study.

There are basically two types of content analysis, conceptual and relational. Conceptual Analysis deals with determining the existence and the frequency of the defined concept in a text, while relational analysis further develops the conceptual analysis by analyzing the relationships among concepts in the texts (Weber, 1990). The data for this study were analyzed using conceptual content analysis in identifying the concepts, and a relational approach is used in the analysis.

The IG principles in the GARP® were analyzed to determine a sample for analysis. Then each ECCOLA card is coded into a manageable content category.

The aim was to help determine the presence of certain IG words and concepts within the GARP® and analyze them against those of ECCOLA. Each card was coded primarily on IG practices, activities, and processes described in the card that indicate or lead to practices in line with the GARP®. A guide of IG practices was developed from each principle, as explained in table 13, to measure adherence to the GARP®. Each card was coded on a scale of "Exist," "Partially exist," and "Does not Exist." With exist implying identification of the references as outlined in the guide (clear and strong reference), partially exist - indicates an identification of some of the references as outlined in the guide (weak reference) and does not exist, implying non-identification of the references outlined in the guide (no reference). A sample of analysis is provided in the appendix. (Abroms, Padmanabhan, Thaweethai & Phillips, 2011.)

Table 13 Content Analysis Guide

Principle	Index
Accountability	• Reference to IG Accountability practices for Information asset (IA) such as an **accountability (decisions)** structure, approved **policies**, **documentation**, and **auditing** practices. • Reference to activities that indicate IG Accountability practices such as **accountability** structure, **documentation**, and **auditability** of IA. • Reference to processes that indicate IG Accountability practices such as documentation and auditability of IA.
Transparency	• Reference to IG transparency practices for IA relating to **documentation** (open and verifiable) and **accessibility** (by appropriate personnel). • References to activities that indicate IG transparency (**documentation** and **accessibility**) practices for IA • Reference to an indication of IG transparency practices such as documentation and accessibility of IA
Integrity	• Reference to IG integrity practices such as **reliability** and **authenticity** of IA. • References to activities that indicate IG integrity (reliability and authenticity) practices for IA • References to processes that indicate IG integrity (reliability and authenticity) practices for IA
Protection	• Reference to IG protection practices such as **protection** mechanisms for **designated** IA (**private**, **confidential**, **privileged**, **secret**, **classified**) and practices. • Reference to **activities** that indicate IG **protection** mechanisms for **designated** IA (**private**, **confidential**, **privileged**, **secret**, **classified**) and practices. • Reference to processes that indicate IG **protection**

	mechanisms for **designated** IA (**private**, **confidential**, **privileged**, **secret**, **classified**) and practices.
Compliance	• Reference to IG compliance practices such as **documentation** and **storage** that facilitate maintaining IA comply with applicable laws, organizational policies, and other binding authorities. • Reference to activities that indicate IG Compliance practices such as the preservation (documentation and storage) of information in line with relevant compliance regulations. • Reference to processes that indicate IG Compliance practices, such as the preservation (documentation and storage) of information in line with relevant compliance regulations.
Availability	• Reference to IG practices such as **accessibility**, **retrieval,** and **documentation** that facilitate maintenance of IA in a manner that ensures timely, efficient, and accurate retrieval. • Reference to activities that connote IG Availability practices such as documentation, accessibility, and retrieval of records and information. • Reference to processes that connote IG Availability practices such as documentation, accessibility, and retrieval of records and information.
Retention	• Refer to IG Retention practices such as documentation, retention/period of retention, and IA storage. • Reference to activities that connote IG Retention practices such as documentation, storage, or retention and or retention period of records and information. • Reference to processes that connote IG practices such as documentation, storage, retention, and or retention period of information and records.
Disposition	• Refer to IG Disposition practices such as transfer, disposition, and documentation of records or information compliance with applicable laws and policies. • Reference to activities that connote IG Disposition practices such as documentation, transfer, or disposition of records or information. • Reference to processes that connote IG Disposition practices such as documentation, transfer or disposition of records or information

5.2 Analysis

5.2.1 Accountability

The GARP® principle of accountability states that

> An organization shall assign a senior executive who will oversee the information governance program, delegate program responsibility to appropriate individuals, adopt policies and procedures to guide personnel, and ensure program auditability. (ARMA, 2009.)

The analysis for the principle of Accountability with ECCOLA reveals that 17 cards have a status of partially existing, and four cards have an "exit" status. Cards # (4,9,18 and 20) suggest IG practices in line with the GARP®. The cards indicate Transparent documentation of activities and processes, audit structure, and a reporting structure (who makes decisions) which indicates an accountability structure in alignment with regulatory bodies and policies, which is in line with GARP® IG practices. While some cards do not directly refer to these terms, their implications are inherent in the practices. As such, the guiding terms can be added to the cards specifically to improve their compliance or the cards left as they are. Analysis of card #4 reveals: *Practices from documenting trade-offs card indicates accountability IG practices. Transparent documentation of activities and processes and a reporting structure (who makes decisions on trade-offs) is identified, which indicates an accountability structure in alignment with regulatory bodies and policies that, if properly documented, can enable auditable practices in line with IG practices. Reference to auditability can further improve the GARP of accountability of the card.* Based on this, the first empirical conclusion (EC1) is made.

> *EC1: Accountability practices such as an accountability structure, approved policies, documentation, and auditing, which help to facilitate the GARP® principle of accounting, exist in cards 4,9,18, and 20.*

Cards # (0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 13,14, 15, 16, 17 and 19) indicate partial representation of GARP® principle of accountability. Each card, in varying degrees, indicates one of the activities or processes of accountability. Therefore, making reference to practices such as accountability structure, documentation of IA, and audits can improve the GARP® IG of these cards. For example, privacy and data (#7) are examined: Practices from privacy and data card indicate *accountability IG practices. Regulatory bodies exist for privacy and data issues signifying an accountability structure, albeit external and may also refer to an internal structure. Making reference to documentation of these practices, which can aid auditability, can help improve the GARP IG practices of this card.* Based on this, the second empirical conclusion is made.

EC2: Accountability practices such as an accountability structure in line with approved policies, documentation, and auditing which help to facilitate the GARP® principle of accounting partially exist in varying degrees in cards 0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 13,14, 15, 16, 17 and 19. To make these cards more compliant with IG practices in line with GARP®, explicit references to these practices can be indicated.

In total, there are two empirical conclusions for the GARP® Accountability principle. Based on these conclusions, the primary empirical conclusions PEC1 was formed.

PEC1: The GARP® of Accountability partially exists in 17 of the cards and exists in four of the cards. To improve this principle within the cards, practices, activities, and processes that reflect an accountability structure, documentation of IA and auditing can be referenced or alluded to in the cards that do not reflect them.

The empirical conclusion and the primary empirical conclusion for Accountability are presented in Table 14.

Table 14 Empirical conclusion for GARP® principle of accountability

Identifier	Empirical conclusion
EC1	Accountability practices such as an accountability structure, approved policies, documentation, and auditing, which help to facilitate the GARP® principle of accounting, exist in cards 4,9,18, and 20.
EC2	Accountability practices such as an accountability structure in line with approved policies, documentation, and auditing which help to facilitate the GARP® principle of accounting partially exist in varying degrees in cards 0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 13,14, 15, 16, 17 and 19. To make these cards more compliant with IG practices in line with GARP®, reference to these practices can be indicated.
PEC1	The GARP® of Accountability partially exists in 17 of the cards and exists in four of the cards. To improve this principle within the cards, practices, activities, and processes that reflect an accountability structure, documentation of IA and auditing can be referenced or alluded to in the cards that do not reflect them.

5.2.2 Transparency

The GARP® principle of transparency states that

> the processes and activities of an organization's Information Governance program shall be documented in an open, verifiable, and understandable manner available to all personnel and appropriate interested parties. (ARMA, para 2,2009.)

The analysis for the principle of Transparency reveals an exit status in five cards (4, 5 ,6, 9 and 10) and 16 cards (0, 1, 2, 3, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20) with a partially exit status. In cards (4, 5,6, 9, and 10) that the principle exists, there is a clear indication of practices, activities, and processes that facilitate documentation and accessibility of IA in an open and verifiable manner and accessible by the appropriate personnel. While specific mention of a practice may not be indicated in some of the cards, the activities and processes reflect these practices; however, specific reference can still be made if necessary. Analysis of card #6 reveals: *IA generated from system reliability practices can facilitate open and verifiable IG documentation and appropriate accessibility practices in line with the principle of transparency. The card references documentation, but reference to the accessibility of these IA can further improve GARP practices of the card.* Based on this, the empirical conclusion is made.

> *EC3: Transparency practices such as documentation and accessibility by appropriate personnel, which can help to facilitate the GARP® principle of transparency, exist in cards 4, 5, 6, 9, and 10).*

Cards # (0, 1, 2, 3, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20) indicate a partial representation of GARP® principle of transparency. Each card reflects either one of the IG practices, activities, and processes of documentation and accessibility in line with the GARP®. Therefore to improve the IG practice of transparency in these cards, both practices can be referenced in these cards. As an example, data quality card #8 is examined: *IA generated from data quality practices can facilitate open and verifiable IG documentation and appropriate accessibility practices in line with the principle of transparency. The card references accessibility, but reference to documentation is missing, and including it can improve GARP practices of the card.* The empirical conclusion is arrived at.

> *EC4: Cards (0, 1, 2, 3, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20) indicate partially exist status due to Transparency practices such as documentation (open and verifiable) and accessibility (by appropriate personnel) of IA not completely referenced in them.*

In total, there are two empirical conclusions for the Transparency GARP principle. Based on these conclusions, the primary empirical conclusion PEC2 was formed.

PEC2: The GARP® of Transparency exists in five cards and partially exists in 16 cards. To further improve this principle, practices such as transparent **documentation** in an open and verifiable manner and **accessibility** by approved personnel can be indicated or referenced in the cards that do not completely reflect them.

The empirical conclusion and the primary empirical conclusion for Transparency are presented in table 15.

Table 15 Empirical conclusion for GARP® principle of transparency

Identifier	Empirical conclusion
EC3	Transparency practices such as documentation and accessibility by appropriate personnel, which can help to facilitate the GARP® principle of transparency, exist in cards 4, 5, 6, 9, and 10).
EC4	Cards (0, 1, 2, 3, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20) indicate partially exist status due to Transparency practices such as documentation (open and verifiable) and accessibility (by appropriate personnel) of IA not completely referenced in them.
PEC2	The GARP® of Transparency exists in five of the cards and partially exists in 16 of the cards. To further improve this principle, practices such as transparent **documentation** in an open and verifiable manner and **accessibility** by approved personnel can be indicated or referenced in the cards that do not completely reflect them.

5.2.3 Integrity

The GARP® principle of integrity states that

> An Information Governance program shall be constructed so that the information assets generated or managed by or for the organization have a reasonable and suitable guarantee of authenticity and reliability. (ARMA, para3, 2009.)

The analysis for the principle of Integrity with ECCOLA reveals that all the cards fully incorporate GARP® that leads to integrity practices. Practices, activities, and processes that lead to reliable and authentic processing and management of IA are effectively represented in all the cards. For example, Card #14 Accessibility analysis is examined: Accessibility card helps ensure that IA generated from these practices is *authentic and reliable. It examines Information generated and seeks reliability in TAIS development, which is in line with GARP practice.* Most of the cards indicate these practices, and some cards make allusions to activities and processes in line with these practices.

> *EC5: All the ECCOLA cards reflect practices, activities, and processes that signify integrity IG, such as reliable and authentic practices of IA in line with GARP®.*

In total, there is one empirical conclusion for the Integrity GARP® principle. Based on this conclusion, the primary empirical conclusion PEC3 was formed.

PEC3: The GARP® principle of Integrity exists in all the cards in ECCOLA because all the cards reflect practices that indicate integrity in the management of IA. While some do not state them specifically, there is an allusion to these in the practices, activities, and processes. However, specific references to authenticity and reliability in the cards can make the principle more visible.

The empirical conclusion and the primary empirical conclusion for Integrity are presented in table 16.

Table 16 Empirical contribution for GARP® principle of integrity

Identifier	Empirical Contribution
EC5	All the ECCOLA cards reflect practices, activities, and processes that signify integrity IG, such as reliable and authentic practices of IA in line with GARP®.
PEC3	The GARP® principle of Integrity exists in all the cards in ECCOLA because all the cards reflect practices that indicate integrity in the management of IA. While some do not state them specifically, there is an allusion to these in the practices, activities, and processes. However, specific references to authenticity and reliability in the cards can make the principle more visible.

5.2.4 Protection

The GARP® principle of protection states that

> An Information Governance program shall be constructed to ensure a reasonable level of protection of information assets that are private, confidential, privileged, secret, or essential to business continuity. (ARMA, para4, 2009.)

The analysis for the principle of Protection with ECCOLA reveals eight of the cards have an existing status while 13 have a partially existing category. Cards # (6, 7, 8, 9, 12, 13, 18, and 20) are categorized as having the principle exist in them in terms of practices, activities, and processes, while cards # (0, 1, 2, 3, 4, 5, 10, 11, 14, 15, 16, 17 and 19) have either one of them partially existing in them. Card 12, which raises awareness for system security, is analyzed: *Forms of protection exist in system security practices. Information Assets are classified, and protection practices are suggested in line with GARP policies and compliance.* This gives rise to the next empirical conclusion.

> *EC6: Cards (6, 7, 8, 9, 12, 13, 18, and 20) with exist categorization indicates GARP® of protection such as protection mechanisms and categorization in the practices, activities, and processes of IA that result from them.*

On the other hand, Cards (0, 1, 2, 3, 4, 5, 10, 11, 14, 15, 16, 17, and 19) have either a protection mechanism or categorization in terms of protection of IA existing in them. An analysis of card 14, which raises awareness on practices on the accessibility of IA, is examined: *Classification of IA exists from Accessibility practices, but there is no reference to protection measures or mechanisms for IA that can result from these practices. Reference to protection practices for IA in line with policies can improve the GARP® of protection in this card.* This gives rise to the next empirical conclusion.

> *EC7: Cards (0, 1, 2, 3, 4, 5, 10, 11, 14, 15, 16, 17, and 19) with the partially exist categorization indicate either a practice, activity, or process relating to protection mechanisms or categorization of IA and not both practices in each card.*

In total, there are two empirical conclusions for the Protection GARP® principle. Based on this, the primary empirical conclusion PEC4 was formed.

PEC4: The GARP® principle of Protection exists in eight of the ECCOLA cards and partially exists in 13 of the cards. To further improve this principle in these cards with partially exist categorization, reference or indications can be made to protection mechanisms and categorization of IA in the practices, activities, and processes of the cards.

Empirical conclusions and the primary empirical conclusion for Protection are presented in table 17.

Table 17 Empirical conclusion for GARP® principle of protection

Identifier	**Empirical conclusion**
EC6	Cards (6, 7, 8, 9, 12, 13, 18, and 20) with exist categorization indicate GARP® of protection, such as protection mechanisms and categorization in the practices, activities, and processes of IA that result from them.
EC7	Cards (0, 1, 2, 3, 4, 5, 10, 11, 14, 15, 16, 17, and 19) with the partially exist categorization indicate either a practice, activity, or process relating to protection mechanisms or categorization of IA and not both practices in each card.
PEC4	The GARP® principle of Protection exists in eight of the ECCOLA cards and partially exists in 13 of the cards. To further improve this principle in the cards with partially exist categorization, reference or indications can be made to protection mechanisms and categorization of IA in the practices, activities, and processes of the cards.

5.2.5 Compliance

The GARP® principle of compliance states that

> An Information Governance program shall be constructed to comply with applicable laws and other binding authorities and the organization's policies.
>
> (ARMA, para5, 2009.)

The analysis for the principle of Compliance with ECCOLA reveals 13 cards with exist categorization and eight cards with a partially exist categorization. Cards (1, 6, 7, 8, 9, 13, 14, 15, 16, 17, 18, 19, and 20) indicate all relevant practices, activities, and processes of documentation and storage of IA in line with relevant compliance or policies. Card #9, which examines Access to data practices, is analyzed: *The card creates awareness for practices involved in accessing data in TAIS development. Reference is made to preserving information which includes documentation and storage of IA from Access to data practices in line with compliance policies denoting GARP of compliance.* This forms the empirical conclusion.

> *EC8: Cards (1, 6, 7, 8, 9, 13, 14, 15, 16, 17, 18, 19, and 20) with exist categorization indicates GARP® of compliance such as documentation and storage of IA in line with compliance policies.*

Cards (0, 2, 3, 4, 5, 10, 11, and 12) indicate either one of the practices of documentation and storage of IA in line with compliance and not both. For example, card # 10, which raises awareness for Human agency practices in the development of TAIS, is analyzed: The card references *the preservation of information from Human Agency practices such as documentation and storage, but no reference is made to regulatory or compliance policies. Reference to the preservation of information in line with compliance policies can improve the GARP of compliance.* This gives rise to the next empirical conclusion.

> *EC9: Cards (0, 2, 3, 4, 5, 10, 11, and 12) with the partially exists categorization indicate either one of the GARP® compliance index of documentation and storage of IA in line with compliance policies and not both.*

In total, there are two empirical conclusions for the Compliance GARP® principle. Based on this conclusion, the primary empirical conclusion PEC5 was formed.

PEC5: The GARP® principle of Compliance partially exists in eight of the ECCOLA cards and exists in 13 of the cards. In the cards with partially existing categorization, practices of preserving IA such as documentation and storage in line with compliance policies can be indicated or referenced in the practices, activities, and processes the cards represent.

Empirical conclusions and the primary empirical conclusion for Compliance are presented in table 18.

Table 18 Empirical conclusion for GARP® principle of compliance

Identifier	Empirical conclusion
EC8	Cards (1, 6, 7, 8, 9, 13, 14, 15, 16, 17, 18, 19, and 20) with exist categorization indicate GARP® of compliance such as documentation and storage of IA in line with compliance policies.
EC9	Cards (0, 2, 3, 4, 5, 10, 11, and 12) with the partially exists categorization indicate either one of the GARP® compliance indexes of documentation and storage of IA in line with compliance policies and not both.
PEC5	The GARP® principle of Compliance partially exists in eight of the ECCOLA cards and exists in 13 of the cards. In the cards with partially existing categorization, practices of preserving IA such as documentation and storage in line with compliance policies can be indicated or referenced in the practices, activities, and processes the cards represent.

5.2.6 Availability

The GARP® principle of availability states that

> An organization shall maintain information assets to ensure timely, efficient, and accurate retrieval of needed information.
>
> (ARMA, 2009.)

The analysis for the principle of Availability with ECCOLA reveals one card categorized exist and 20 cards as partially exists. Card #9 indicates that practices, activities, and processes such as accessibility, retrieval, and documentation that can facilitate maintenance of IA exists in a manner that ensures timely, efficient, and accurate retrieval in line with the GARP® principle. An analysis of the card, which creates awareness for practices associated with Access to data in the development of TAIS, reveals that: *Access to data card references governance frameworks, storage and use of IA, which alludes to documentation, accessibility, and retrieval practices that facilitate the availability of information in line with GARP.*

> *EC10: Card #9 with exist categorization indicates GARP® of Availability such as documentation, accessibility, and retrieval of IA that can facilitate the maintenance of IA in a manner that ensures timely, efficient, and accurate retrieval.*

Cards (0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20) reveal that practices, activities, and processes such as accessibility, retrieval and documentation that can facilitate maintenance of IA in a manner that ensures timely, efficient, and accurate retrieval in line with the GARP® principle of availa-

bility partially exists in these cards. This means these cards do not fully reflect these practices in their entirety. An analysis of card #11, which creates awareness for human agency practices in the development of TAIS, reveals: *Human oversight references governance practices that may be involved in transparent communication as regards human autonomy over system agent practices. However, there are no references to accessibility, retrieval, and documentation of IA from these practices, which could improve the GARP of availability of the card.* This gives rise to the next empirical conclusion.

> *EC11: Cards (0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20) with the partially exist categorization indicate representation of the GARP® of Availability such as documentation, accessibility and retrieval of IA that can facilitate the maintenance of IA in a manner that ensures timely, efficient, and accurate retrieval is partially represented as each of these cards do not fully reference these practices in their entirety.*

In total, there are two empirical conclusions for the Availability GARP® principle. Based on this conclusion, the primary empirical conclusion PEC6 was formed.

PEC6: The GARP® principle of Availability partially exists in 20 of the ECCOLA cards and exists in one of the cards. In the cards with partially exist categorization, practices of documentation, accessibility, and retrieval of IA that can facilitate maintenance of IA in a manner that ensures timely, efficient, and accurate retrieval can be referenced in the cards.

Empirical contribution and the primary empirical conclusion for availability are presented in table 19.

Table 19 Empirical conclusion for GARP® principle of Availability

Identifier	**Empirical conclusion**
EC10	Card #9 with existing categorization indicates GARP® of Availability such as documentation, accessibility, and retrieval of IA that can facilitate the maintenance of IA to ensure timely, efficient, and accurate retrieval.
EC11	Cards (0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20) with the partially exist categorization indicate representation of the GARP® of Availability such as documentation, accessibility, and retrieval of IA that can facilitate the maintenance of IA in a manner that ensures timely, efficient, and accurate retrieval is partially represented as each of these cards do not fully reference these practices in their entirety.
PEC6	The GARP® principle of Availability partially exists in 20 of the ECCOLA cards and exists in one of the cards. In the cards with

> partially exist categorization, practices of documentation, accessibility, and retrieval of IA that can facilitate maintenance of IA in a manner that ensures timely, efficient, and accurate retrieval can be referenced in the cards.

5.2.7 Retention

The GARP® principle of Retention states that

> An organization shall maintain its information assets for an appropriate time, taking into account legal, regulatory, fiscal, operational, and historical requirements. (ARMA, para7, 2009.)

The analysis of the principle of Retention with ECCOLA reveals nine of the cards have a partial exist status while 12 have a does not exist status. Cards (0, 4, 5, 6, 7, 8, 9, 18, and 20) indicate that principles of Retention such as documentation, storage, and retention (period of retention) of IA are not represented or referenced in their entirety in these cards. Each of the cards references one of the practices, activities, or processes. An analysis of card #4, which creates awareness for the documentation of trade-offs in TAIS development, is further analyzed: *Documenting Trade-offs card indicates the documentation aspect of the retention practice, but there is no reference to the retention/retention period of IA*. This forms the basis for the next empirical contribution.

> *EC12: Cards (0, 4, 5, 6, 7, 8, 9, 18, and 20) with a partial categorization indicate that IG practices of retention such as documentation, storage, and retention (retention period) are not fully referenced. Each of the cards contains an aspect of retention but not all the practices identified.*

Cards (1, 2, 3, 10, 11, 12, 13, 14, 15, 16, 17, and 19) indicate that IG principles of retention such as documentation, storage, and retention (retention period) are not referenced either in practice, activities, or processes. An analysis of card #13, which creates awareness for systems safety in the development of TAIS, reveals that: *While the card makes reference to measures for risks and safety, no indication or reference to practices of retention such as storage, retention, or retention period nor documentation of IA that can emanate from these practices.* This gives rise to the next empirical contribution.

> *EC13: Cards (1, 2, 3, 10, 11, 12, 13, 14, 15, 16, 17, and 19) with a does not exist categorization indicates that IG practices of retention such as documentation, storage, and retention (retention period) are not referenced. Each of the cards does not reflect any aspect of retention, with none of its practices identified.*

In total, there are two empirical contributions for the Retention GARP® principle. Based on this conclusion, the primary empirical conclusion PEC7 was formed.

PEC7: The GARP® principle of Retention partially exists in nine of the cards and does not exist in 12 of the cards. To improve the practices of retention in these cards, such as documentation, storage, and retention (period of retention) of IA, references or allusion to them can be made in the practices, activities, and processes.

Empirical conclusions and the primary empirical conclusion for Retention are presented in table 20.

Table 20 Empirical conclusion for GARP® principle of Retention

Identifier	Empirical conclusions
EC12	Cards (0, 4, 5, 6, 7, 8, 9, 18, and 20) with a partial categorization indicate that IG practices of retention such as documentation, storage, and retention (retention period) are not fully referenced. Each of the cards contains an aspect of retention but not all the practices are identified.
EC13	Cards (1, 2, 3, 10, 11, 12, 13, 14, 15, 16, 17, and 19) with a does not exist categorization indicates that IG practices of retention such as documentation, storage, and retention (retention period) are not referenced. Each of the cards does not reflect any aspect of retention, with none of its practices is identified.
PEC7	The GARP® principle of Retention partially exists in nine of the cards and does not exist in 12 of the cards. To improve the practices of retention in these cards, such as documentation, storage, and retention (period of retention) of IA, references or allusion to them can be made in the practices, activities, and processes.

5.2.8 Disposition

The GARP® Principle of Disposition states

> An organization shall provide a secure and appropriate disposition for information assets that are no longer required to be maintained by applicable laws and the organization's policies. (ARMA, para8, 2009.)

An analysis for the principle of Disposition with ECCOLA reveals that nine of the cards have a categorization of partial exist and 12 have a does not exist categorization. Cards (0, 4, 5, 6, 7, 8, 9, 18, and 20) indicate that principles of Disposition such as transfer, disposition, and documentation of records or information in compliance with applicable laws and policies are not represented or referenced in their entirety in these cards. Each of the cards references one of

the practices, activities, or processes. An analysis of card #7 which creates awareness for privacy and data practices in the development of TAIS, indicate that: *Whilst the Privacy and data card makes reference to organizational policies which may indicate governance practices including disposition, it does not indicate any practices of disposition of IA such as transfer, disposition, or documentation.* This forms the basis for the next empirical contribution.

> *EC14: Cards (0, 4, 5, 6, 7, 8, 9, 18, and 20) have partial categorization, which may suggest some IG practices of disposition such as transfer, disposition, or documentation of IA which comply with applicable laws and policies. Each of the cards contains only an aspect of disposition, and not all the practices are identified in each card.*

Cards (1, 2, 3, 10, 11, 12, 13, 14, 15, 16, 17, and 19) indicate that the GARP® principle of disposition practices such as transfer, disposition, or documentation of IA which comply with applicable laws and policies can be identified in these cards. These cards indicate a lack of reference to these practices in activities, practices, and processes. An analysis of card #11, which creates awareness of human oversight practices, indicates that: *while the card activities, processes, and practices generate awareness for human autonomy issues in the development of TAIS, it does not indicate any practices of disposition of IA such as transfer, disposition, or documentation of records.* This forms the basis for the next empirical contribution.

> *EC15: Cards (1, 2, 3, 10, 11, 12, 13, 14, 15, 16, 17 and 19) have a does not exist categorization. This indicates that IG practices of disposition such as transfer, disposition, and documentation of records that comply with applicable laws and policies have not been identified. None of the cards reflect or identify any aspect of retention in practices, activities, and processes.*

In total, there are two empirical conclusions for the Disposition GARP® principle. Based on this conclusion, the primary empirical conclusion PEC8 was formed.

> PEC8: The GARP® principle of Disposition partially exists in nine of the cards and does not exist in 12 of the cards. In improving the practices of disposition in these cards, such as transfer, disposition, and documentation of IA in compliance with applicable laws and policies, reference or allusion can be made to them in practices, activities, and processes.

Empirical conclusions and the primary empirical conclusion for Disposition are presented in table 21.

Table 21 Empirical conclusion for GARP® principle of disposition

Identifier	Empirical conclusion
EC14	Cards (0, 4, 5, 6, 7, 8, 9, 18, and 20) have partial categorization, suggesting some IG practices of disposition such as transfer, disposition, or documentation of IA that comply with applicable laws and policies. Each of the cards contains only an aspect of disposition, and not all the practices are identified in each card.
EC15	Cards (1, 2, 3, 10, 11, 12, 13, 14, 15, 16, 17 and 19) have a does not exist categorization. This indicates that IG practices of disposition such as transfer, disposition, and documentation of records that comply with applicable laws and policies have not been identified. None of the cards reflect or identify any aspect of retention in practices, activities, and processes.
PEC8	The GARP® principle of Disposition partially exists in nine of the cards and does not exist in 12 of the cards. In improving the practices of disposition in these cards, such as transfer, disposition, and documentation of IA in compliance with applicable laws and policies, reference or allusion can be made to them in practices, activities, and processes.

5.3 Summary

The overall findings from the analysis are made up of contributions (EC and PEC), each key principle of the GARP® that was critically analyzed with each ECCOLA card—resulting in 15 empirical conclusions which form the eight primary empirical conclusions as illustrated in the heat map in figure 12. The colored squares indicate if the GARP® exists, partially exists, or does not exist in each of the ECCOLA cards. The codes exist (color green), implying the card embodies the practices, processes, and activities of the principle, partially exist (color yellow), implying an indication of the practices, activities, and processes were identified in the card and does not exist (color red) implies none or no indication of the practices, activities or processes were identified in the card.

ECCOLA CARD	GARP® by ARMA							
	Accountability	Transparency	Integrity	Protection	Compliance	Availability	Retention	Disposition
#0 Stakeholder analysis	Partially exist	partially exist	exist	partially exist	partially exist	partially exist	partially exist	partially exist
#1 Types of transparency	Partially exist	partially exist	exist	partially exist	exist	partially exist	Does not exist	Does not exist
#2 Explainability	Partially exist	partially exist	exist	partially exist	partially exist	partially exist	Does not exist	Does not exist
#3 Communication	Partially exist	partially exist	exist	partially exist	partially exist	partially exist	Does not exist	Does not exist
#4 Documenting trade-offs	exist	exist	exist	partially exist	partially exist	partially exist	partially exist	partially exist
#5 Traceability	Partially exist	exist	exist	partially exist	partially exist	partially exist	partially exist	partially exist
#6 System Reliability	Partially exist	exist	exist	exist	exist	partially exist	partially exist	partially exist
#7 Privacy and Data	Partially exist	partially exist	exist	exist	exist	partially exist	partially exist	partially exist
#8 Data quality	Partially exist	partially exist	exist	exist	exist	partially exist	partially exist	partially exist
#9 Access to data	exist	exist	exist	exist	exist	exist	partially exist	partially exist
#10 Human agency	Partially exist	exist	exist	partially exist	partially exist	partially exist	Does not exist	Does not exist
#11 Human oversight	Partially exist	partially exist	exist	partially exist	partially exist	partially exist	Does not exist	Does not exist
#12 System security	Partially exist	partially exist	exist	exist	partially exist	partially exist	Does not exist	Does not exist
#13 System safety	Partially exist	partially exist	exist	exist	exist	partially exist	Does not exist	Does not exist
#14 Accessibility	Partially exist	partially exist	exist	partially exist	exist	partially exist	Does not exist	Does not exist
#15 Stakeholder participation	Partially exist	partially exist	exist	partially exist	exist	partially exist	Does not exist	Does not exist
#16 Environmental Impact	Partially exist	partially exist	exist	partially exist	exist	partially exist	Does not exist	Does not exist
#17 Societal Effects	Partially exist	partially exist	exist	partially exist	exist	partially exist	Does not exist	Does not exist
#18 Auditability	exist	partially exist	exist	exist	exist	partially exist	partially exist	partially exist
#19 Ability to redress	Partially exist	partially exist	exist	partially exist	exist	partially exist	Does not exist	Does not exist
#20 Minimizing negative impacts	exist	partially exist	exist	exist	exist	partially exist	partially exist	partially exist

Figure 12 Contributions towards the analysis of ECCOLA with GARP®

The result is based on the 15 empirical conclusions that make up the eight primary conclusions in table 22.

Table 22 Summary of Empirical conclusion for GARP® principles with ECCOLA

Identifier	Empirical conclusion
PEC1	The GARP® of Accountability partially exists in 17 of the cards and exists in four of the cards. To improve this principle within the cards, practices, activities, and processes that reflect an accountability structure, documentation of IA and auditing can be referenced or alluded to in the cards that do not reflect them.
PEC2	The GARP® of Transparency exists in five of the cards and partially exists in 16 of the cards. To further improve this principle, practices such as transparent documentation in an open and verifiable manner and accessibility by approved personnel can be indicated or referenced in the cards that do not completely reflect them.
PEC3	The GARP® principle of Integrity exists in all the cards in ECCOLA because all the cards reflect practices that indicate integrity in the management of IA. While some do not state them specifically, there is an allusion to these in the practices, activities, and processes. However, specific references to authenticity and reliability in the cards can make the principle more visible.
PEC4	The GARP® principle of Protection exists in eight of the ECCOLA cards and partially exists in 13 of the cards. To further improve this principle in the cards with partially exist categorization, reference or indications can be made to protection mechanisms and categorization of IA in the cards' practices, activities, and processes.

PEC5 The GARP® principle of Compliance partially exists in eight of the ECCOLA cards and exists in 13 of the cards. In the cards with partially existing categorization, practices of preserving IA such as documentation and storage in line with compliance policies can be indicated or referenced in the practices, activities, and processes the cards represent.

PEC6 The GARP® principle of Availability partially exists in 20 of the ECCOLA cards and exists in one of the cards. In the cards with partially exist categorization, practices of documentation, accessibility, and retrieval of IA that can facilitate maintenance of IA in a manner that ensures timely, efficient, and accurate retrieval can be referenced in the cards.

PEC7 The GARP® principle of Retention partially exists in nine of the cards and does not exist in 12 of the cards. To improve the practices of retention in these cards, such as documentation, storage, and retention (period of retention) of IA, references or allusion to them can be made in practices, activities, and processes.

PEC8 The GARP® principle of Disposition partially exists in nine of the cards and does not exist in 12 of the cards. In improving the practices of disposition in these cards, such as transfer, disposition, and documentation of IA in compliance with applicable laws and policies, reference or allusion can be made to them in practices, activities, and processes.

6 DISCUSSION

This chapter reviews findings from the analysis and synthesizes them with the theoretical findings of this study.

6.1 Practical Contribution

The findings from the empirical analysis are discussed below and highlighted in table 23, which displays the PECs and their implication for practice.

The principle of accountability is based on the need for an accountability structure in line with approved policies that oversee IG practices such as documentation and management of IA are effectively carried out to facilitate audit practices. The findings from PEC1 reveal that while four of the cards exhibit these principles, 17 of the remaining cards partially demonstrate these practices. In practice, this can translate to crucial accountability practices being omitted in the use of these cards. Where they have not been indicated, developers may cultivate practices, activities, and processes that overlook accountability structure or do not comply with policies or non-documentation of crucial information. On the other hand, inculcating the need to include them in developmental practices of TAIS practices fully in the cards where they partially exist can help improve these practices and make them routine, which can improve IG practices. In addition, displaying a governance structure demonstrates that ECCOLA takes IG practices, policies, and responsibilities seriously.

The principle of Transparency is based on processes and activities of IG programs being conducted in an open and transparent manner and made available for appropriate personnel and interested parties. Being transparent is one of the strengths and debate for ethical TAIS. In PEC2 has transparency practices exist in five of the cards, and 16 of them exhibit partially exist. Whilst the importance of having transparent processes and activities in the development of TAIS cannot be over-emphasized, not applying transparent practices in the practices, activities, and processes that result from these cards can result in out-

comes that are not transparent or easily accessible by interested parties. In addition, it can also impede the IG process by being unable to provide transparent IA, thereby becoming a missing link in the process for developers. In addition, different parties will have different interests in information at different times. As such, it is important that IA generated by the practices, processes, and activities are transparent, understandable, and readily available.

The principle of Integrity deals with IA being generated or managed in a manner where they are suitably authentic or reliable. PEC3 reveals ECCOLA method ensures that IA generated or managed in the development of TAIS is based on integrity as all the practices, processes, and activities that emanate from the method are in agreement with the GARP® principle of Integrity. All the 21 cards indicate categorization exists, which implies that in practice, the activities and processes can lead to authentic and reliable outcomes in the IG process and ultimately in the systems being developed. For audit trails in IG, reliable information is crucial as they demonstrate that the practices, processes, and activities are generating reliable IA.

The principle of protection deals with conducting IG programs that ensure that Information Assets are properly categorized and protected. PEC4 reveals the GARP® principle of Protection exists in eight of the ECCOLA cards and partially exists in 13 of the cards. Non-categorization or improper categorization of IA can lead to the wrong level of protection being provided and could lead to disclosure or a breach if accessed by improperly authorized or unauthorized personnel or party for developers. Therefore, it is important to reference the proper categorization of IA in the development process of TAIS so that the correct level of categorization can be provided to avoid any serious consequences. In addition, it is important that practices, activities, and processes display reasonable safeguards for IA that can limit incidental access or disclosures.

The principle of Compliance deals with IG being constructed to comply with applicable laws, binding authorities, and policies. PEC4 reveals that the GARP® principle of Compliance partially exists in eight of the ECCOLA cards and exists in 13 of the cards. This suggests that practices of preservation of IA such as documentation and storage in line with compliance policies need to be further emphasized in the practices, activities, and processes the cards represent. These practices can help developers know what information they need to enter into their records to demonstrate that their activities are being conducted in compliance with laws, authorities, and policies. It will also help them maintain and monitor their IA in a manner consistent with regulations and laws.

The principle of Availability deals with IA being maintained to ensure timely, efficient, and accurate retrieval when needed. PEC6 reveals the GARP® principle of Availability partially exists in 20 of the ECCOLA cards and exists in one of the cards. Suggesting practices of documentation, accessibility, and retrieval of IA that can facilitate maintenance of IA in a manner that ensures timely, efficient, and accurate retrieval can be further emphasized. These practices can help developers minimize inconsistent and error-prone interpretation of IA

if they are maintained in an efficient and accurate manner that facilitates retrieval. In addition, it can help protect IA from being easily corrupted or lost and help facilitate audits and regulatory procedures.

The principle of Retention deals with IA being maintained for an appropriate time with legal, regulatory, fiscal, operational, and historical requirements being taken into account. PEC7 reveals that the GARP® principle of Retention partially exists in nine of the cards and does not exist in 12 of the cards. This implies that practices of retention such as documentation, storage, and retention (period of retention) of IA need to be emphasized in one category of the card, and in the other category, they need to be established. For developers, retention practices can help define which IA to maintain, for how long (retention period) and which is no longer valuable and needs to be disposed of and train the AIS to act accordingly. Such practices can help ascertain the risk associated with some IA and aid legal, regulatory, fiscal, regulatory, operational, and historical requirements.

The principle of disposition deals with IA being securely and appropriately disposed of when no longer required in line with applicable laws and the organization's policies. PEC8 reveals that the GARP® principle of Disposition partially exists in nine of the cards and does not exist in 12 of the cards. This implies the need for the emphasis on disposition practices such as transfer, disposition, and documentation of IA in compliance with applicable laws and policies. In practice, developers that establish practices, processes, and activities in line with disposition can help them determine which IA to maintain and which to dispose of to help mitigate the risks associated with them and train AIS accordingly. Efficient disposition of IA may also help instill trust in users as they know that their data/information will not be used indefinitely in the development of TAIS.

Table 23 PECs and implications for practice

Empirical conclusion	**Implication for practice**
PEC1	Accountability can help improve the accountability structure for IA in development practices in methods like ECCOLA. TAIS developers may become more mindful of IG practices for the IA they generate in activities and processes and incorporate them, knowing there is an accountability structure in place.
PEC2	Transparency can create transparent practices in the IG of IA and make them more accessible for developers using methods like ECCOLA. In addition, different parties with different interests in IA generated from practices, processes, and activities can easily access transparent and understandable information.
PEC3	Integrity practices can lead developers to authentic

	and reliable IA in IG and ultimately in the development of TAIS using methods like ECCOLA. For audit trails, the integrity of information is crucial and can help demonstrate that the practices, processes, and activities are generating reliable IA.
PEC4	Protection practices in development methods like ECCOLA can help developers with an important categorization of IA in the development process of TAIS and provide safeguards to mitigate risks from incidental access or disclosures.
PEC5	Compliance practices can help developers streamline information when using development methods like ECCOLA. It can help them maintain and monitor IA in a manner consistent with regulations and laws.
PEC6	Availability practices in development methods like ECCOLA can help developers minimize inconsistent and error-prone interpretations of IA and make them more accessible.
PEC7	Retention practices in development methods like ECCOLA can help developers define maintenance of IA to determine their retention period to aid legal, regulatory, fiscal, regulatory, operational, and historical requirements.
PEC8	Disposition practices in development methods like ECCOLA can help developers efficiently establish practices, processes, and activities in line with the disposition principle that can help determine which IA to maintain and which to dispose of to mitigate associated risks.

6.2 Theoretical Contribution

The empirical findings suggest IG practices are at varying stages in ECCOLA using the GARP® as a guide. The implication of having these principles distributed in the method is explored in theory.

The principle of accountability (PEC1) is based on the need for an accountability structure in line with approved policies that oversee IG practices such as documentation and management of IA are effectively carried out to facilitate audit practices. In theory, Reddy et al. (2020) analyses accountability as a challenge as regards implementation in terms of governance. They analyze that appropriate stages are needed for effective accountability practices. They stress the need for an approval structure by governing bodies or regulating authori-

ties that oversee and preview processes to ensure proper documentation of IA that can aid audits in governance (Yang et al., 2021). (Reddy et al., 2020.) Rodrigues (2020) explains that accountability in governance requires regulatory oversight and needs to be in place in development methods. They analyze that guiding actions by an accountability structure and explanations in the form of documentation of IA can facilitate audit either internally or externally. (Rodrigues et., 2020.) According to Yamin et al. (2021), having an accountability structure makes developers of AIS accountable in the documentation of their IA, and incorporating such practices can help reduce the opacity of governance frameworks such as IG (Hildebrandt, 2016).

The principle of transparency (PEC2) highlights the need for transparent practices in the IG of IA to make them more accessible for developers using methods like ECCOLA. Caron (2019) explains that transparent processes in AIS help to improve auditability. She explains that open and verifiable practices that generate IA need to be transparent in the development process. Transparency can aid understanding when these IA are accessed by appropriate personnel, which is important for auditability. Also, in the development process of TAIS, different cognitive biases and heuristics exist (Caron, 2019) and, as such, warrants the need for transparent obligations to be imposed in the form of auditable governance to help mitigate these practices so that when these IA are accessed there is a clear understanding that helps audit in governance. (Caron, 2019.) Kiener (2020) agrees with these practices and discusses the need for open and verifiable processes in the development of AIS in sensitive fields like medicine. He explains that transparent processes and activities of AIS can aid human oversight, risks, and audits in governance frameworks like IG.

PEC3 highlights the principle of integrity, which explains the need for reliability and authenticity in the processes, activities, and practices that generate IA in TAIS development to help facilitate proper audit in governance frameworks such as IG. Janson et al. (2020) explain that a governance framework like IG in the development of AIS can enable authentic and reliable IA. They analyze that a governance structure (IG) can enable IA of integrity by managing the quality, validity, security, and associated risks in practice to preserve the integrity.

PEC4 is based on the principle of protection. It emphasizes the need for protection practices for designated IA (private, confidential, privileged, secret, classified) such as protection mechanisms and practices that can provide safeguards to mitigate risks from incidental access or disclosures. Culnan (2019) supports this finding and explains that the protection of IA to ensure they are secure and ensuring they are properly categorized can help avert security and privacy breaches. Whilst the introduction of regulatory mechanisms such as the GDPR and the California Consumer Privacy Act (CCPA) exists to help provide protection for IA, but they are insufficient in protecting IA that are not properly categorized or labeled to ensure that the right form of protection is provided. Further adding that practices such as these can aid the audit processes involved in governance frameworks of AI developers. (Culnan, 2019.)

PEC5 findings relate to how compliance practices like documentation and storage of IA in a manner that complies with applicable laws, organizational policies, and other binding authorities can help aid governance frameworks such as IG. In (2018) explains that IG practices of maintaining IA in a manner that conforms to compliance (internal and external) can help mitigate risk and increase efficiency. He explains that when developers or organizations familiarise themselves with compliance practices and streamline the maintenance of their IA in line with governance frameworks like IG, then such practices become routine and make it easy to produce systems that are compliant with applicable laws and binding authorities. In addition, in legal matters and regulation, these practices can help audit processes in governance frameworks like IG. (In, 2018.)

PEC6 is based on the principle of availability. It explores how practices such as documentation, accessibility, and retrieval practices can facilitate the maintenance of IA in a manner that ensures timely efficiency. Accurate retrieval can be referenced in the cards. Hind et al. (2020) explains that developers are usually faced with the challenge of documentation of IA in the development of AIS as there are no clear guidelines on how much to document to provide enough clarity. Therefore, having a governance approach such as IG, which provides guidance from relevant stakeholders, can ensure the most effective and appropriate IA are documented and, upon retrieval, provides holistic information. Documentation of IA in line with a governance framework like IG also provides confidence that information made available is wholesome and suitable for all interested parties. In addition, these practices also aid the governance frameworks in audit processes to ensure unity. (Hind et al., 2020.)

PEC7 is based on the principle of retention and how practices like documentation, storage, and retention (period of retention) of IA can improve IG in development methods like ECCOLA. Kroll (2018) recommends the minimization of retention of collected records or the disposal of aggregate records where possible to enhance efficient governance of records. He explains that retention of IA should be properly documented and subject to a governance structure to reduce the risks of retaining them beyond their retention period. Retention of IA within a governance framework like IG can help reduce risk from legitimate requests from law enforcement. When Information Assets are retained beyond their lifecycle, they can pose a risk if authorities request them and utilize them beyond the scope for which they were acquired. (Kroll, 2018.)

PEC8 works on the principle of disposition and explains how disposition practices like transfer, disposition, and documentation of records or information (IA) in compliance with applicable laws and policies can improve IG or governance frameworks in development methods like ECCOLA. Kroll (2018) explains that regular disposal of aggregate and redundant records or reducing them to the lowest level of sensitivity can reduce privacy risks and increase the efficiency of IG. When Information Assets are maintained for a period, a need for them must be further established to enable them not to pose a risk of redundancy which may hamper efficiency. As such, a clear need exists for records or

information to be retained for a period and disposed of accordingly, as it can help to make development methods more trustworthy when audits of the IA are carried out in governance frameworks like IG. The findings are advised in table 24.

Table 24 PECs and contribution to Theory

Identifier	Empirical conclusion	Relation to existing theory
PEC1	The GARP® of Accountability partially exists in 17 of the cards and exists in four of the cards. To improve this principle within the cards, practices, activities, and processes that reflect an accountability structure, documentation of IA and auditing can be referenced or alluded to in the cards that do not reflect them.	Corresponding to previous research (Rodrigues, 2020; Yamin et al., 2021; Reddy et al., 2020; Hildebrandt, 2016; Yang et al., 2021)
PEC2	The GARP® of Transparency exists in five of the cards and partially exists in 16 of the cards. To further improve this principle, practices such as transparent documentation in an open and verifiable manner and accessibility by approved personnel can be indicated or referenced in the cards that do not completely reflect them.	Corresponding (Caron, 2019; Kiener, 2020)
PEC3	The GARP® principle of Integrity exists in all the cards in ECCOLA because all the cards reflect practices that indicate integrity in the management of IA. While some do not state them specifically, there is an allusion to these in the practices, activities, and processes. However, specific references to authenticity and reliability in the cards can make the principle more visible.	Corresponding (Janson et al., 2020)
PEC4	The GARP® principle of Protection exists in eight of the	Corresponding (Culnan, 2019)

	ECCOLA cards and partially exists in 13 of the cards. To further improve this principle in the cards with partially exist categorization, reference or indications can be made to protection mechanisms and categorization of IA in the cards' practices, activities, and processes.	
PEC5	The GARP® principle of Compliance partially exists in eight of the ECCOLA cards and exists in 13 of the cards. In the cards with partially existing categorization, practices of preserving IA such as documentation and storage in line with compliance policies can be indicated or referenced in the practices, activities, and processes the cards represent.	Corresponding (In, 2018)
PEC6	The GARP® principle of Availability partially exists in 20 of the ECCOLA cards and exists in one of the cards. In the cards with partially exist categorization, practices of documentation, accessibility, and retrieval of IA that can facilitate maintenance of IA in a manner that ensures timely, efficient, and accurate retrieval can be referenced in the cards.	Corresponding (Hind et al., 2020)
PEC7	The GARP® principle of Retention partially exists in nine of the cards and does not exist in 12 cards. To improve the practices of retention in these cards, such as documentation, storage, and retention (period of retention) of IA, references or allusion to them can be made in practices, activities, and processes.	Corresponding (Kroll, 2018)

PEC8	The GARP® principle of Disposition partially exists in nine of the cards and does not exist in 12 of the cards. In improving the practices of disposition in these cards, such as transfer, disposition, and documentation of IA in compliance with applicable laws and policies, reference or allusion can be made to them in practices, activities, and processes.	Corresponding (Kroll, 2018)

6.3 Main Contribution

The main contribution of this study in line with the DSRM is the development of an artifact as a solution to the problem-centered approach. A new card is proposed based on the main vulnerability discovered in ECCOLA of the virtually non-existent principles of retention and disposition. The principles of Retention and Disposition works on the premise that information assets be maintained for an appropriate period considering legal, regulatory, fiscal, operational, historical, and ethical requirements and disposed accordingly in line with IG laws and policies (ARMA, 2009). The incorporation of these principles can help promote ethical practices for developing TAIS. This is because openly communicating to users, the period of retention and disposition of their data and/or information will encourage their use of trustworthy AI systems as they know that their information will not be used indefinitely (Kroll, 2018). Also, frequently updating user's data and disposing of redundant data will enable trustworthy AI systems to generate current and better value (Kroll, 2018). Figure 13 is an illustration of how these principles can be represented in ECCOLA.

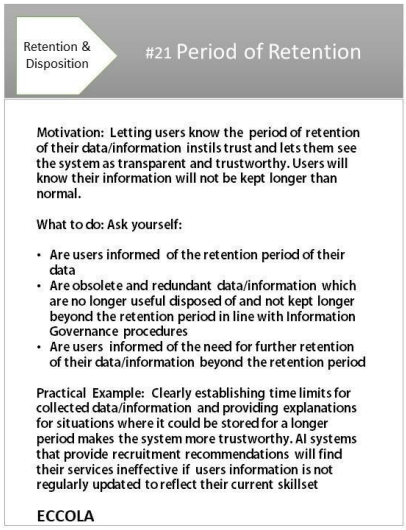

Figure 13 ECCOLA card #21 illustrating Retention and Disposition guideline

Motivation- Letting users know the length of time or life cycle their data/ information will be retained by the system provides transparency, leading to users trusting the system knowing that their information will not be kept for longer than necessary.

What to do: Ask yourself:

- Are users informed of the retention period of their data?
- Are Obsolete and redundant data/information no longer useful, destroyed, and not kept longer than the approved time frame in line with record keeping and legal requirements?
- Are users informed of the need for further retention of their data beyond the retention period?

Practical example: Establishing time limits for collected data and providing explanations for situations where it could be stored for an extended period makes the system more trustworthy. AI systems that provide recruitment recommendations will find their services ineffective if users' information is not regularly updated, as giving the right service for someone whose skillset has changed will be ineffective (European Union, 2021.)

Reviewing a development method like ECCOLA through the lens of GARP® IG practices is relatively new. As such, this study can serve as an addition to the body of knowledge specifically in the field of development methods for TAIS. According to In et al. (2018), Information Governance practices that are poorly cultivated, collected, or misgoverned diminishes the effect of its performance; therefore, in line with this reasoning, effectively incorporating all the GARP® principles as carried out in this study can help improve ECCOLA method in the efficient management and governance of information and can enhance its robustness.

7 CONCLUSION

The research aimed to evaluate AI ethics in developing trustworthy system in line with Information Governance (IG) (GARP®) practices by examining ECCOLA to answer the research question:

How to extend ECCOLA to cover Information Governance principles?

In answering this question, the 21 ECCOLA cards were critically analyzed with the eight principles of GARP®) following the design science approach of initiating a problem. The results reveal that most of the GARP®) IG principles – Accountability, Transparency, Integrity, Protection, Compliance, and Availability are incorporated in ECCOLA to varying degrees. But the principles of retention and disposition are lacking. The resulting heatmap highlights the areas where IG practices exist and areas where these practices could be further incorporated to improve IG practices, and areas where the associated practices do not exist. To help extend ECCOLA and make it more robust and compliant in terms of accountability to governance frameworks, the lacking practices of retention and disposition, which were identified as the least incorporated IG practices, would need to be incorporated further in ECCOLA.

Incorporating or applying these principles is suggested in the form of a new card (card #21), which highlights the motivation for these GARP® IG practices. These practices can help improve ethical practices in TAIS development within ECCOLA by improving IG practices within the method. On the other hand, these practices can also be integrated within existing practices as both approaches can highlight their importance.

The findings helped to answer the research question. It helped extend and develop ECCOLA further using (GARP®) IG practices, thereby increasing its robustness in line with EU recommendation of ethical AIS development and improving its accountability to governance frameworks to achieve the study's goal.

7.1 Limitations of the Study

While the outcome from the analysis produced a solution in the form of a new card, #21, the card is yet to be tested within a real-life scenario or within a simulated environment to test its effectiveness. Testing of card #21 can help provide real-life insight into the efficacy of incorporating the principles of Retention and Disposition into ECCOLA and lend further credence to the study. In addition, testing card #21 can help generate feedback data that can be used to refine the approach further to improve ECCOLA.

Another limitation is the lack of literature on IG and TAIS development methods. Literature in this area is virtually non-existent. Most of the available literature focused on general AI governance frameworks, specifically on ML and algorithms, and negligible on methods development regarding IG.

The use of GARP® by ARMA as the only Information Governance (IG) approach for the study provides a limitation that constrains the study to IG practices covered under only GARP®, thereby narrowing the focus of the research. While this approach yielded some credible results, widening the scope of IG beyond GARP® could provide greater insight.

7.2 Future Research Opportunities

For Future research, IG could be further broken down to the data governance level and examined to see how it contributes to accountability via governance frameworks. While IG provides a holistic approach to managing records, a data governance angle can help bring more insight into the basic level of data management and how it contributes to this approach.

Testing of card #21 can help provide insight into the validity of this study to serve as an area for further research. Testing card #21 within a real-life scenario or a simulated environment can further refine the findings in this study to give a deeper insight into how the new card and practices improve IG practices in the method. Currently, in its untested state, there exist areas of improvement that are now undetected. In line with the Design science method, constantly refining an artifact through testing can help identify correction areas to improve the artifact and subsequently the development of TAIS.

Another area for possible further research can leverage this study's findings by determining the level of maturity of the GARP® IG the practices within the method can provide and how these can improve Information governance frameworks in the development of TAIS.

Additionally, a possible research area can be in the development of an IG framework for TAIS that can help with the IG of TAIS, which can constitute part of the coherent framework of the virtually non-existent IG literature where different practices can vary in terms of contextual particulars as recommended by (Wang & Siau, 2018).

The analysis carried out by the conceptual framework revealed other method practices such as certification and testing and validation as having only two squares. These practices can be further evaluated with ECCOLA to help determine how they can improve the method further.

8 REFERENCES

Abroms, L. C., Padmanabhan, N., Thaweethai, L., & Phillips, T. (2011). iPhone apps for smoking cessation: a content analysis. *American journal of preventive medicine*, *40*(3), 279-285.

Adadi, A., and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. Vol.6. 52138-52160

Altman, R. (2017). Artificial intelligence (AI) systems for interpreting complex medical datasets. *Clinical Pharmacology & Therapeutics.* Vol.100(5). 585-586.

American Health Information Management Association. (2014). Information governance principles for healthcare (IGPHC).

ARMA (2009). The Principles® (Generally Accepted Recordkeeping Principles®). Retrieved from https://www.arma.org/page/principles

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), 6-1.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information Fusion, 58, 82–115. doi.org/10.1016/j.inffus.2019.12.012

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems. Journal of the Association for Information Systems, 22(2), 8.

Atkinson, K., Bench-Capon, T., & Bollegala, D. (2020). Explanation in AI and law: Past, present, and future. *Artificial Intelligence*. Vol. (289) (December 2020).103387.

BBC. (2020). BBC.COM. Uber's self-driving operator charged over a fatal crash. [online]: https://www.bbc.com/news/technology-54175359 Accessed on 25/01/2021

BCC Research (2018). *Machine Learning: Global Markets to 2022.* Bccresearch.com. Available from https://www.bccresearch.com/market-research/information-technology/machine-learning-global-markets.html Accessed 03 February 2021

Bengio, Y. (2009) Learning Deep Architectures for AI", Foundations and Trends in Machine Learning: Vol. 2(1) 1-127. http://dx.doi.org/10.1561/2200000006

Bennett, S. (2017). What is information governance, and how does it differ from data governance? Governance Directions, 69(8), 462-467.

Borgman, H., Heier, H., Bahli, B., & Boekamp, T. (2016, January). "Dotting the I and crossing (out) the T in IT governance: New challenges for information governance. In 2016 49th Hawaii International Conference on System Sciences (HICSS) (pp. 4901-4909). IEEE.

Caron, M. S. (2019). The transformative effect of AI on the banking industry. *Banking & Finance Law Review*, *34*(2), 169-214. https://search.proquest.com/docview/2207836906?pqorigsite=gscholar&fromopenview=true

Caron, M. S., & Gupta, A. (2020). The Social Contract for AI. *arXiv preprint arXiv:2006.08140*.

Content Analysis. (n.d.). Retrieved May 24, 2021, from https://www.publichealth.columbia.edu/research/population-health-methods/content-analysis

Culnan, M. J. (2019). Policy to avoid a privacy disaster. *Journal of the Association for Information Systems*, *20*(6), 1.

Dafoe, A. (2018). AI governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*.

Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). Qualitative content analysis: A focus on trustworthiness. SAGE open, 4(1), 2158244014522633.

Eitel-Porter, R. (2021). Beyond the promise: implementing ethical AI. AI and Ethics, 1(1), 73-80.

Ertel, W. (2018). *Introduction to artificial intelligence*. Springer.

European Commission. (2019). Ethics Guidelines for Trustworthy AI. Brussels. Retrieved from https://ec.europa.eu/futurium/en/ai-allianceconsultation/guidelines

EU Commission. (2020). White Paper on Artificial Intelligence—A European Approach to Excellence and Trust. *COM (2020)*, *65*.

Explainable Artificial Intelligence (XAI). *IEEE Access.* Vol.6. 52138-52160

Fikes, R., & Garvey, T. (2020). Knowledge Representation and Reasoning--A History of DARPA Leadership. AI Magazine, 41(2).

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018, October). What makes users trust a chatbot for customer service? An exploratory interview study. In *International conference on internet science* (pp. 194-208). Springer, Cham.

Garg, A., Singh, S., Li, W., Gao, L., Cui, X., Wang, C., Peng, X., and Rajasekar, N. (2020). Illustration of experimental, machine learning, and characterization methods for the study of the performance of Li-ion batteries. International Journal of Energy Research. Vol. 44(2). 9513-9526.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), 2nd Web*, *2*(2).

Guresen, E., & Kayakutlu, G. (2011). Definition of artificial neural networks with comparison to other networks. *Procedia Computer Science*, *3*, 426-433.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, *61*(4), 5-14. https://doi.org/10.1177/0008125619864925

Hagmann, J. (2013). Information governance–beyond the buzz. *Records Management Journal*.

Hamon, R., Junklewitz, H., & Sanchez, I. (2020). Robustness and explainability of artificial intelligence. Publications Office of the European Union.

Hanid, M. B. (2014). *Design science research as an approach to develop conceptual solutions for improving cost management in construction* (Doctoral dissertation, University of Salford).

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT press.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 75-105.

Hildebrandt, M. (2016). The New Imbroglio–Living with Machine Algorithms.

Hind, M., Houde, S., Martino, J., Mojsilovic, A., Piorkowski, D., Richards, J., & Varshney, K. R. (2020, April). Experiences with improving the transparency of ai models and services. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-8).

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527-1554.

HSBC. (2019). *HSBC Bank and SoftBank Robotics America Enhance Beverly Hills Banking Experience with Pepper Robot.* HSBC.COM. Retrieved from https://www.about.us.hsbc.com/news-and-media

Hu, P., & Lu, Y. (2021). Dual humanness and trust in conversational AI: A person-centered approach. *Computers in Human Behavior*, *119*, 106727.

IEEE Standards Association. (2019).The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE, 2019. *https://ethicsinaction.ieee.org/?utm_campaign=EAD1e&utm_medium=PR&utm_source=Web&utm_content=geias*

In, J., Bradley, R., Bichescu, B. C., & Autry, C. W. (2018). Supply chain information governance: Toward a conceptual framework. The International Journal of Logistics Management.

International Council of Archives, (2016). ICA.org. *What are archives?* Retrieved from https://www.ica.org/ Accessed March 20, 2021.

Ischen, C., Araujo, T., Voorveld, H., van Noort, G., & Smit, E. (2019, November). Privacy concerns in chatbot interactions. In *International Workshop on Chatbot Research and Design*. Vol 11970, (pp. 34-48). Springer, Cham. https://doi.org/10.1007/978-3-030-39540-

Jain, S., Luthra, M., Sharma, S., & Fatima, M. (2020, March). Trustworthiness of Artificial Intelligence. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 907-912). IEEE.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. Government Information Quarterly, 37(3), 101493.

Jaremko, J. L., Azar, M., Bromwich, R., Lum, A., Alicia Cheong, L. H., & Giber, M. (2019). Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group. Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology. Can. Assoc. Radiol. J, 70, 107-118.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *61*(4), 577-586.

Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2

Kiener, M. (2020). Artificial intelligence in medicine and the disclosure of risks. *Ai & Society*, 1-9. DOI: 10.1007/s00146-020-01085-w

Königstorfer, F., & Thalmann, S. (2020). Applications of Artificial Intelligence in commercial banks–A research agenda for behavioral finance. *Journal of Behavioral and Experimental Finance*, *27*, 100352. DOI: 10.1016/j.jbef.2020.100352

Kooper, M. N., Maes, R., & Lindgreen, E. R. (2011). On the governance of information: Introducing a new concept of governance to support the management of information. *International journal of information management*, *31*(3), 195-200.

Kroll, J. A. (2018). Data science data governance [AI ethics]. IEEE Security & Privacy, 16(6), 61-70.

Kumar, A., Braud, T., Tarkoma, S., & Hui, P. (2020, March). Trustworthy AI in the Age of Pervasive Computing and Big Data. In 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops) (pp. 1-6). IEEE.

Kumar, R. R., Reddy, M. B., & Praveen, P. (2019). Text classification performance analysis on machine learning. *International Journal of Advanced Science and Technology*, *28*(20), 691-697.

Kwon, O., Bae, S., & Shin, B. (2020). Understanding the Adoption Intention of AI through the Ethics Lens. *Proceedings of the 53rd Hawaii International Conference on System Sciences.* (2020). 4972-4981. Retrieved from https://aisel.aisnet.org/hicss-53/ks/aspects_of_ai/3/

Ledesma, S., Ibarra-Manzano, M. A., Cabal-Yepez, E., Almanza-Ojeda, D. L., & Avina-Cervantes, J. G. (2018). Analysis of data sets with learning conflicts

for machine learning. *IEEE Access*, *6*, 45062-45070. DOI: 10.1109/ACCESS.2018.2865135.

Leijnen, S., Aldewereld, H., van Belkom, R., Bijvank, R., & Ossewaarde, R. (2020). An Agile Framework for Trustworthy AI.

Leikas, J., Koivisto, R., & Gotcheva, N. (2019). Ethical framework for designing autonomous intelligent systems. *Journal of Open Innovation: Technology, Market, and Complexity*, *5*(1), 18.

Logan, D. (January 2010). Gartner.com "What is Information Governance? And Why is it So Hard?" [online blog] Available from: https://blogs.gartner.com/debra_logan/2010/01/11/what-is-information-governance-and-why-is-it-so-hard/ Accessed 18 March 2021

Lomas, E. (2010). Information governance: information security and access within a UK context. *Records Management Journal*. Vol.20(2). 182-198

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing–Connecting computing power to human insights. *International Journal of Research in Marketing*, *37*(3), 481-504.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of management review, 20(3), 709-734.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, *27*(4), 12-12.

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS)*, *2*(2), 1-25.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501-507. http://dx.doi.org/10.2139/ssrn.3391293

Müller, V., C. (30 April 2020). Stanford Encyclopaedia of Philosophy. Accessed February 11, 2021

NewTechdogo. (March 2018). NewTechdogo.com. 13+ List of Machine Learning Algorithms with Details. [online blog] Available from: https://www.newtechdojo.com/list-machine-learning-algorithms/ Accessed on 05 February 2021.

Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). Deep Learning for Deepfakes Creation and Detection: A Survey. *arXiv preprint arXiv:1909.11573*.

Onwuegbuzie, A. J., & Frels, R. (2016). "Seven steps to a comprehensive literature review: A multimodal and cultural approach.

Oosthuizen, K., Botha, E., Robertson, J., & Montecchi, M. (2020). Artificial intelligence in retail: The AI-enabled value chain. *Australasian Marketing Journal*, j-ausmj. https://doi.org/10.1016/j.ausmj.2020.07.007

Peffers, K., Rothenberger, M., Tuunanen, T., & Vaezi, R. (2012, May). Design science research evaluation. In *International Conference on Design Science Research in Information Systems* (pp. 398-410). Springer, Berlin, Heidelberg.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77.

Rassloff, J. (2018, March). Recognizing the Value of AI. *KPMG,*

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in healthcare. Journal of the American Medical Informatics Association, 27(3), 491-497

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Rich, E Artificial Intelligence New York: McGraw-Hill 1983.

Robinson, S. C. (2020). Trust, transparency, and openness: How the inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). *Technology in Society*, Vol.63(101421). https://doi.org/10.1016/j.techsoc.2020.101421

Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges, and vulnerabilities. *Journal of Responsible Technology*. Vol. 4(2020).100005, 1-12 https://doi.org/10.1016/j.jrt.2020.100005

Rossi, F. (2018). Building trust in artificial intelligence. Journal of International Affairs, 72(1), 127-134.

Rossi, F., Ala-Pietilä, P., Bauer, W., Bergmann, U., Beliková, M., Bonefeld-Dahl, C., … Yeung, K. (2019). A DEFINITION OF AI: MAIN CAPABILITIES AND DISCIPLINES. Brussels: European Commission

Rudin, C., & Radin, J. (2019). Why are we using black-box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.5a8a3a3d

Sætra, H. S. (2021). AI in Context and the Sustainable Development Goals: Factoring in the Unsustainability of the Sociotechnical Systems. *Sustainability*, *13*(4), 1738.

Sen, S., Dasgupta, D., & Gupta, K. D. (2020, July). An empirical study on the algorithmic bias. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)* (pp. 1189-1194). IEEE.

Simmons, A-B., & Chappel, S-G. (1988). Artificial Intelligence-Definition and Practice. IEEE Journal of Oceanic Engineering. Vol. 13(2). April 1988. 14-42. DOI: 10.1109/48.551

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tallon, P. P., Ramirez, R. V., & Short, J. E. (2013). The information artifact in IT governance: toward a theory of information governance. *Journal of Management Information Systems*, *30*(3), 141-178.

Tan, C. F., Wahidin, L. S., Khalil, S. N., Tamaldin, N., Hu, J., & Rauterberg, G. W. M. (2016). The application of expert system: A review of research and applications. *ARPN Journal of Engineering and Applied Sciences*, *11*(4), 2448-2453.

Taschuk, M., & Wilson, G. (2017). Ten simple rules for making research software more robust. PLOS Computational Biology 13(4): e1005412. https://doi.org/10.1371/journal.pcbi.1005412

Theodorou, A., & Dignum, V. (2020). Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence*, *2*(1), 10-12.

Thiebes, S., Lins, S., & Sunyaev, A. (2020). Trustworthy artificial intelligence. Electronic Markets, 1-18.

Tolvanen, J. (2020). Development of Trustworthy Cyber-Physical Systems: Artificial Intelligence's Viewpoint.

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. L. A., Elkhatib, Y., ... & Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications, and research challenges. *IEEE Access*, *7*, 65579-65615. DOI: 10.1109/ACCESS.2019.2916648

Vakkuri, V., & Abrahamsson, P. (2018, June). "The key concepts of ethics of artificial intelligence. In *2018 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC)* (pp. 1-6). IEEE.

Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2020, August). ECCOLA-a method for implementing ethically aligned AI systems. In *2020 46th*

Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (pp. 195-204). IEEE.

Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2020, August). ECCOLA-a method for implementing ethically aligned AI systems. In *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)* (pp. 195-204). IEEE.

Vakkuri, V., Kemell, K. K., Kultanen, J., Siponen, M., & Abrahamsson, P. (2019). Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. *arXiv preprint arXiv:1906.07946*.

Veiga, A. D., & Eloff, J. H. (2007). An information security governance framework. *Information systems management*, *24*(4), 361-372.

Wang, W., & Siau, K. (2018). Artificial intelligence: a study on governance, policies, and regulations. MWAIS 2018 proceedings.

Weber, R. P. (1990). *Basic content analysis* (No. 49). Sage.

Wickramasinghe, C. S., Marino, D. L., Grandio, J., & Manic, M. (2020, June). Trustworthy AI Development Guidelines for Human System Interaction. In *2020 13th International Conference on Human System Interaction (HSI)* (pp. 130-136). IEEE.

Wierenga, B. (2010). Marketing and artificial intelligence: Great opportunities, reluctant partners. In *Marketing intelligent systems using soft computing* (pp. 1-8). Springer, Berlin, Heidelberg.

Williamson, O. E. (1984). The economics of governance: framework and implications. *Zeitschrift für die gesamte Staatswissenschaft/Journal of Institutional and Theoretical Economics*, (H. 1), 195-223.

Winby, S., & Mohrman, S. A. (2018). Digital sociotechnical system design. The Journal of Applied Behavioral Science, 54(4), 399-423.

Winby, S., & Mohrman, S. A. (2018). Digital sociotechnical system design. *The Journal of Applied Behavioral Science*, *54*(4), 399-423.

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The dark sides of artificial intelligence: An integrated AI governance framework for public administration. International Journal of Public Administration, 43(9), 818-829.

Yamin, M. M., Ullah, M., Ullah, H., & Katt, B. (2021). *Weaponized AI for cyber attacks. Journal of Information Security and Applications,* 57(2021).102722, 1-5 https://doi.org/10.1016/j.jisa.2020.102722

Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, *2* (2021) 100008. https://doi.org/10.1016/j.caeai.2021.100008

Zwetsloot, R., & Dafoe, A. (2019). Thinking about risks from AI: accidents, misuse, and structure. *Lawfare. February 11*, 2019.

9 APPENDIX

A sample of the analysis for the Accountability principle against card 5- Traceability is presented below in table 25

Table 25 Sample of analysis of card #5 and Accountability principle

Guide	GARP-Accountability	ECCOLA Cards	Analysis	Categorization
Reference to general IG Accountability practices for IA relating to an **accountability structure** with the collaboration of relevant stakeholders to **approved policies**, **documentation,** and **auditing** practices in a manner that complies with GARP practices.	An organization shall assign a senior executive who will oversee the information governance program, delegate program responsibility to appropriate individuals, adopt policies and procedures to guide personnel, and ensure program auditability. (ARMA, 2009)	Card #5 – Traceability What to do: **Document**. Different types of documentation (code, project, etc.) are typically crucial in producing transparency. Have you documented the development of the system, both in terms of code and decision-making? How was the model built or the AI train? How have you documented the testing and validation process? In terms of scenarios	Practices from the traceability card indicate accountability IG practices. Transparent documentation of processes and activities that align with regulatory bodies and policies that, if properly documented, can enable auditable practices of IA that meet relevant standards in line with IG practices. **Reference to an accountability structure and auditability practices** can further improve the GARP of accountability of the card.	Partially Exist

	and scenarios used etc. How do you document the actions of the systems? What about alternate activities (e.g., if the user was different but the situation otherwise the same)? (Vakkuri et al., 2020.)		