

Representation of Trustworthiness Components in AI Systems: A Formalized Approach Considering Interconnections and Overlap

Eduard Manziuk¹, Oleksandr Barmak¹, Pavlo Radiuk¹, Vladislav Kuznetsov² and Iurii Krak^{2,3}

¹ Khmelnytskyi National University, 11, Institutaska str., Khmelnytskyi, 29016, Ukraine

² Glushkov Institute of Cybernetics, 40, Glushkov ave., Kyiv, 03187, Ukraine

³ Taras Shevchenko National University of Kyiv, 63/13, Volodymyrska str., Kyiv, 01601, Ukraine

Abstract

The study addresses the problem of integrating trustworthiness components into artificial intelligence (AI) systems. A new method is proposed to determine the interdependence and intersection of concepts in the field of trustworthy AI. The approach provides a structured way to assess the interconnections and overlaps between different trustworthiness concepts, offering a more complete understanding of their complex interaction. A method for assessing the degree of coincidence between different trustworthiness components is proposed, which allows for a more accurate analysis of their interconnections. The results of experimental studies have shown the level of interconnection between the concepts, with an average level of overlap of about 67%. A formal model for integrating trustworthiness components into AI systems has also been developed. This model provides a framework for evaluating the actions of an AI agent against several trustworthiness criteria simultaneously. The approach takes into account different contexts and scenarios, providing a more robust and flexible assessment of trustworthy AI. By bridging the gap between theoretical concepts and practical implementation, this work contributes to the development of trustworthy AI systems. The proposed work also provides a more structured and formalized approach to understanding and implementing trustworthy AI. Furthermore, this research aims to contribute to the development of AI systems that are built on ethical principles and societal values.

Keywords

Trustworthy AI, concept interdependence, formal modeling trustworthiness, ethical AI, reliability assessment, trustworthiness integration, artificial intelligence ethics, AI safety ¹

1. Introduction

The rapid development and adoption of artificial intelligence (AI) systems has brought the concepts of trust and trustworthiness to the forefront of discussions on AI ethics and governance. As AI increasingly impacts important aspects of society, from healthcare decisions to financial services, the need for trustworthy AI has become paramount. This article explores the multifaceted nature of trustworthiness in AI systems, proposing a formalized approach to representing and integrating trustworthiness components, taking into account their interconnections and overlaps.


The concept of trustworthiness in AI covers various dimensions, including safety, security, reliability, and availability [1, 2]. Trust in AI is fundamentally based on the user's vulnerability and ability to predict AI decisions [3]. As the industry evolves, there is a growing recognition among policymakers, companies, and researchers of the importance of developing standardized models and ontologies to define key concepts and potential threats in AI systems [1, 4].

Although numerous studies have attempted to lay a conceptual foundation for understanding trust and trustworthiness in AI [5–8], there is still a need for more comprehensive and formalized approaches. This paper aims to address this gap by proposing a method for analyzing the

ProfIT AI 2024: 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024), September 25–27, 2024, Cambridge, MA, USA

✉ eduard.em.km@gmail.com (E. Manziuk); alexander.barmak@gmail.com (A. Barmak);

radiukpavlo@gmail.com (P. Radiuk); kuznetsow.wlad@gmail.com (V. Kuznetsov); yuri.krak@gmail.com (Iu. Krak)

ORCID  0000-0002-7310-2126 (E. Manziuk); 0000-0003-0739-9678 (A. Barmak); 0000-0003-3609-112X (P. Radiuk); 0000-0002-1068-769X (V. Kuznetsov); 0000-0002-8043-0785 (Iu. Krak)



© 2024 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



CEUR Workshop Proceedings (CEUR-WS.org)

interconnections and overlaps between different components of trustworthiness. Although decisions can be made using various approaches [9–11]. It builds on existing work on formalizing trustworthiness as computational concepts [12, 13], which is further extended to develop a more holistic understanding of trustworthiness in AI systems.

This paper makes several key contributions to the field of trustworthy AI:

- New formalized approach for representing and analyzing trustworthiness components in AI systems is recommended. The method provides a structured way to evaluate the interconnections and overlaps between different trustworthiness concepts, offering a more complete understanding of their complex interconnections.
- Method for assessing the degree of overlap between different trustworthiness components is recommended, which allows for a more accurate analysis of their interconnections.
- Formal model for integrating trustworthiness components into AI systems is developed. The model provides a framework for evaluating the actions of AI agents according to numerous trustworthiness criteria simultaneously. The approach takes into account different contexts and scenarios, providing a more reliable and flexible assessment of trustworthy AI.

By bridging the gap between theoretical concepts and practical implementation, the paper contributes to the development of trustworthy AI systems. By providing a more structured and formalized approach to understanding and implementing trustworthy AI, this research aims to contribute to the development of AI systems that are not only technically advanced but also aligned with ethical principles and societal values.

The article is organized as follows. The Related Works section reviews existing research in the area of trustworthy AI, identifies gaps, and formulates the purpose of the article. Sections 3 and 4 present a new approach to analyzing the interdependence of concepts in trustworthy AI and a formal model for integrating trustworthiness components in AI systems. The Experiments section presents experimental studies and their results of the developed methods. The "Discussion" section analyzes the results of the methods, including the degree of overlap between trustworthiness components, and discusses the implications for the development of AI systems. The "Conclusions" section summarizes the main results, emphasizes the importance of implementing trustworthy AI, and outlines future research directions.

2. Related Works

The concept of trustworthiness in AI systems is multifaceted, encompassing safety, security, reliability, and availability [2]. Trust in AI is based on user vulnerability and the ability to anticipate AI decisions, with trustworthiness detached from sociological notions [4]. Formalization of trustworthy AI involves developing standardized models and ontologies to identify key concepts and threats [6]. Trust and trustworthiness frame debates on AI governance, with policymakers and companies recognizing their importance [14]. Approaches to trust in AI include interpersonal, institutional, and epistemic perspectives. Improving trustworthiness requires formal specification and verification methods [2], as well as consideration of intrinsic reasoning and extrinsic behavior [4]. The effectiveness of self-defined commitments to ensure trustworthy AI development and governance remains questioned by academia and civil society [14].

A number of studies aim to lay the conceptual foundation for understanding trust and reliability in AI. The papers [2, 4, 6] analyze the relationship between AI ethics principles, reliability, and trust, emphasizing that reliability is a means to build trust. Next papers [5, 15–17] propose the perspective of modeling as a conceptual basis for understanding trust.

These works reflect researchers' attempts to clarify and formulate precise definitions and conceptual frameworks for AI trust and reliability as a starting point for further work. However, they largely remain at the theoretical and philosophical level.

Another group of studies aims to formalize the concepts of reliability and trust in the form of mathematical or computational models, [18, 19] propose a structural model that decomposes

reliability into trust in outcomes and trust in process. Papers [20–22] develop a formalization of trust as a computational concept. The paper [23] introduces a formal model of a source network and belief change operators, demonstrating how users can be misled into trusting information from an AI system.

Such formal approaches can be useful for clearly specifying requirements for reliable AI systems and subsequently developing methods to satisfy them. However, it is noted that these formalizations are often incomplete or too narrow to encompass all aspects of reliability in complex AI systems. A number of works [12, 24–26] focus on investigating individual dimensions of reliability, such as safety, transparency, fairness, or explainability. Papers [11, 18, 26] present a roadmap for achieving reliability by addressing issues of model verification, robustness to attacks, fairness, and transparency. The works [12, 25, 27] in their reviews also consider definitions, challenges, requirements, and prospects of reliable AI.

Transparency and explainability of AI systems are crucial for building trust. Papers [3, 7, 8, 28] provides a comprehensive review of the state of research in this area, discussing the components of trustworthy AI, the need for it in various industries, and the importance of transparent and post-hoc explanation models in the construction of explainable systems. The paper [29] proposes a 12-item scale for the subjective assessment of trust in human-centered AI systems by stakeholders, [30] presents a reliabilistic approach to justify the degree of confidence in trusting AI systems based on the calibration between perceived and actual trust.

These studies are important as they detail certain aspects of reliability and put forward requirements for AI systems to achieve them. However, there is a tendency towards fragmentation, with different groups focusing on only one or a few dimensions without integrating the various perspectives on reliability into a holistic system.

While some scholars question the feasibility or usefulness of the concept of "reliable AI," other authors [9, 31] defend its relevance and necessity and advocate for the idea of reliable AI as a critical benchmark for the ethical development and use of artificial intelligence in response to criticism and skepticism.

These works underscore the importance of adhering to reliability principles to ensure the positive impact of AI technologies on society and prevent potential risks. They can serve as motivation and a catalyst for further efforts in this area.

Unlike purely conceptual works, some studies [9, 25, 32] propose practical steps and mechanisms for achieving AI reliability. For instance, propose a model for implementing reliable AI through standardization, certification, and auditing mechanisms based on openness and transparency. Publication [25] provide key reliability metrics and principles for engineering reliable machine learning systems.

These approaches are important for transitioning from theory to the practical implementation of reliable AI in the form of standards, best practices, and widely accepted requirements that developers should follow [9]. However, the question remains as to the practical operationalization of such initiatives and their broader adoption by the industry [8, 33, 34].

Some works [16, 35] analyze AI reliability from economic and regulatory perspectives and consider the economic incentives and benefits of investing in reliable AI systems. [35] discuss potential mechanisms for regulation and auditing to ensure reliability.

These issues are important for creating incentives and regulatory requirements for companies, regulators, and the entire sector to improve the reliability of AI systems. However, views on optimal regulatory approaches may differ, and the economic dimension requires deeper analysis.

Collectively, the works reflect various perspectives and approaches to addressing the problem of building reliable artificial intelligence systems. To achieve substantial progress, an integration of conceptual frameworks, formal models, technical solutions, multidimensional standards, economic incentives, and regulatory mechanisms into a comprehensive strategy for ensuring AI reliability with the participation of all stakeholders is needed.

The goal of this paper is to propose a novel method for analyzing the interdependence of trustworthiness concepts in AI systems and to develop a formal model for integrating these components, thereby contributing to the practical implementation of trustworthy AI.

3. Method for determining the interconnections and overlap of concepts within a domain

For a formal presentation, we propose the following approach using a combination of methods.

The method of direct matching correspondence allows establishing a matching between concepts of two different domains or structured areas.

Formally, it can be expressed as follows:

$$C_1 \equiv C_2 \Leftrightarrow \forall x(x \in C_1 \Leftrightarrow x \in C_2), \quad (1)$$

where C_1 and C_2 – is a concept from different domains that we consider equivalent or relevant.

The full coverage method assumes that each concept from one domain completely covers (or is a subset of) a concept from another domain:

$$C_1 \subseteq C_2 \Leftrightarrow \forall x(x \in C_1 \rightarrow x \in C_2), \quad (2)$$

where C_1 is a subset of C_2 , i.e. every element from concept C_1 belongs to concept C_2 .

Proximity and interpretation methods allow quantifying how close concepts from different domains are to each other with a certain degree of correspondence.

Formally, this can be represented as a function of proximity $d(C_1, C_2)$:

$$d : C \times C \rightarrow [0,1], \quad (3)$$

where is the value of proximity between concepts C_1 and C_2 , which is in the interval between 0 and 1.

The joint use of methods in the analysis of domain areas is shown in Fig. 1

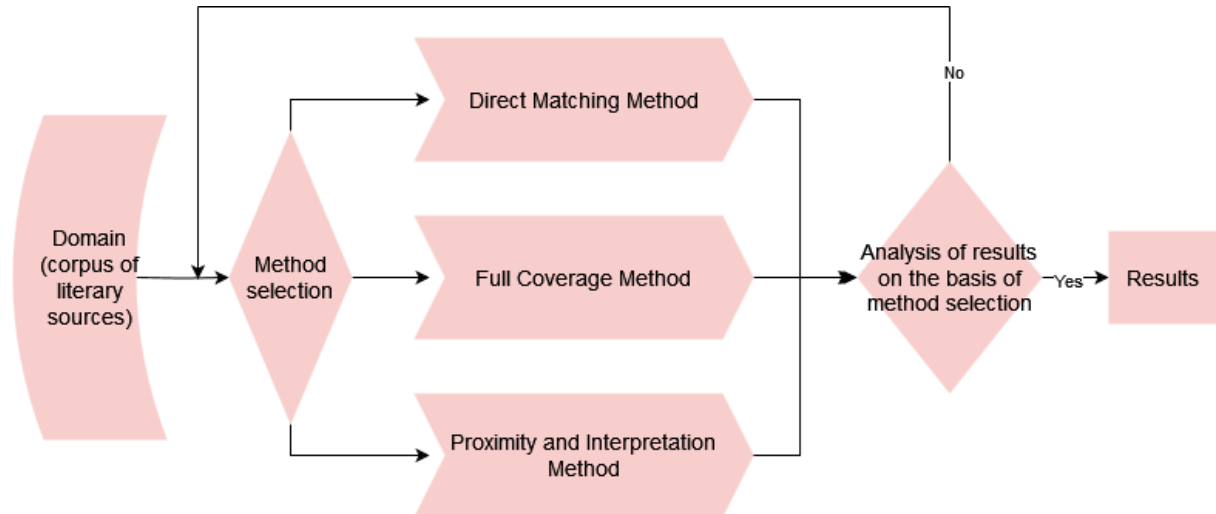


Figure 1: The process of selecting and applying domain analysis methods

The figure below shows the sequence of choosing and applying methods for analyzing a particular domain (corpus of literary sources). The process begins with defining the research area or corpus of sources to be analyzed. At the next stage, a decision is made to select one of three possible methods, which determines the further direction of the domain analysis. The choice of method opens up several possible paths.

If the direct match method is selected, the domain elements are analyzed for relevance based on the direct match. If the full coverage method is chosen, a comprehensive analysis is performed that covers all possible relationships and interactions. If the method of proximity and interpretation is selected, the analysis is based on the proximity of the elements and the interpretation of their meanings.

After applying the selected method, the results are subject to further analysis. At this stage, it is checked whether the results meet expectations or standards. If the results do not, it may be necessary to return to the previous steps to revise the method or reanalyze. If the results meet the requirements, the process moves to the final stage where the results are recorded and interpreted.

The diagram (Figure 1) shows the logical chain of choosing domain analysis methods and further processing the results. The process consists of several possible branches, depending on the choice of method and evaluation of the results. If the results meet the requirements, the process is completed, and if not, re-analysis may be necessary.

Thus, using a formal representation, it is possible to analyze trustworthiness concepts, establish their interconnections, and assess the level of proximity between different domains. The proposed methods are used to compare equal information domains in order to determine how concepts from one domain correspond to concepts in another domain. As the research results have shown, the use of these methods is advisable if the description of concepts in each of the comparison domains is sufficiently complete and detailed. However, there are a number of limitations and conditions under which the application of the methods is difficult. The methods work when there are sufficiently similar or analogous concepts. In some cases, it may be difficult to establish clear equivalence due to contextual differences. The methods do not take into account that some parts of a concept may be interdependent or overlap with other concepts, which is not always reflected in the description. Also, the definition of relations depends on the correct construction of the proximity function, which can be a difficult and subjective task. Proximity values can vary depending on the interpretation and context.

Accordingly, we propose a method for determining the interconnections and overlap of concepts within a domain. The method of assessing overlap for concept analysis begins with the selection of the concepts to be investigated. This may include, for example, trustworthiness components such as transparency, clarity, privacy, etc. It is important to determine which concepts will be analyzed, as this will affect the further stages of the research.

The next step is to find theoretical definitions of these concepts in the relevant literature. This can be either a specialized corpus of documents or an overview document covering the domain. The main task at this stage is to find and collect definitions of concepts that provide a clear idea of their meaning. If there are no clearly defined or generally accepted definitions for a particular concept, it is necessary to clarify them based on the interpretation in a particular scientific text. This involves a deeper analysis and interpretation of the definitions to ensure that the concept is accurate and understandable.

Next, the main components of each definition should be identified, i.e., those elements that are constant and essential to understanding the concept. This may include the key characteristics or aspects that make up the bulk of the definition. After identifying the main components, the next step is to compare these components between the two concepts to identify identical elements. This will help determine which components are in common between the different concepts and to what extent they overlap.

To quantify the share of each component of the total definition, the percentage of each component should be calculated. This will make it possible to estimate what percentage of the definition is made up of certain components. The next step is to estimate the share of identical components in the scope of the first concept. This will make it possible to determine what percentage of the first concept's component overlaps with the second concept's components. Converting this share to a percentage allows getting a preliminary level of overlap between concepts. This process must be repeated for each pair of concepts to obtain the full range of data for analysis.

The collected data on the percentage of overlap between different concepts should be processed by calculating the arithmetic mean of the overlap levels. This will provide a generalized result for all analyzed pairs of concepts. The last step is to analyze the results. At this stage, it is assessed what the percentages of overlap mean, how they reflect the interconnections between the concepts, and what can be drawn as general conclusions based on the analysis.

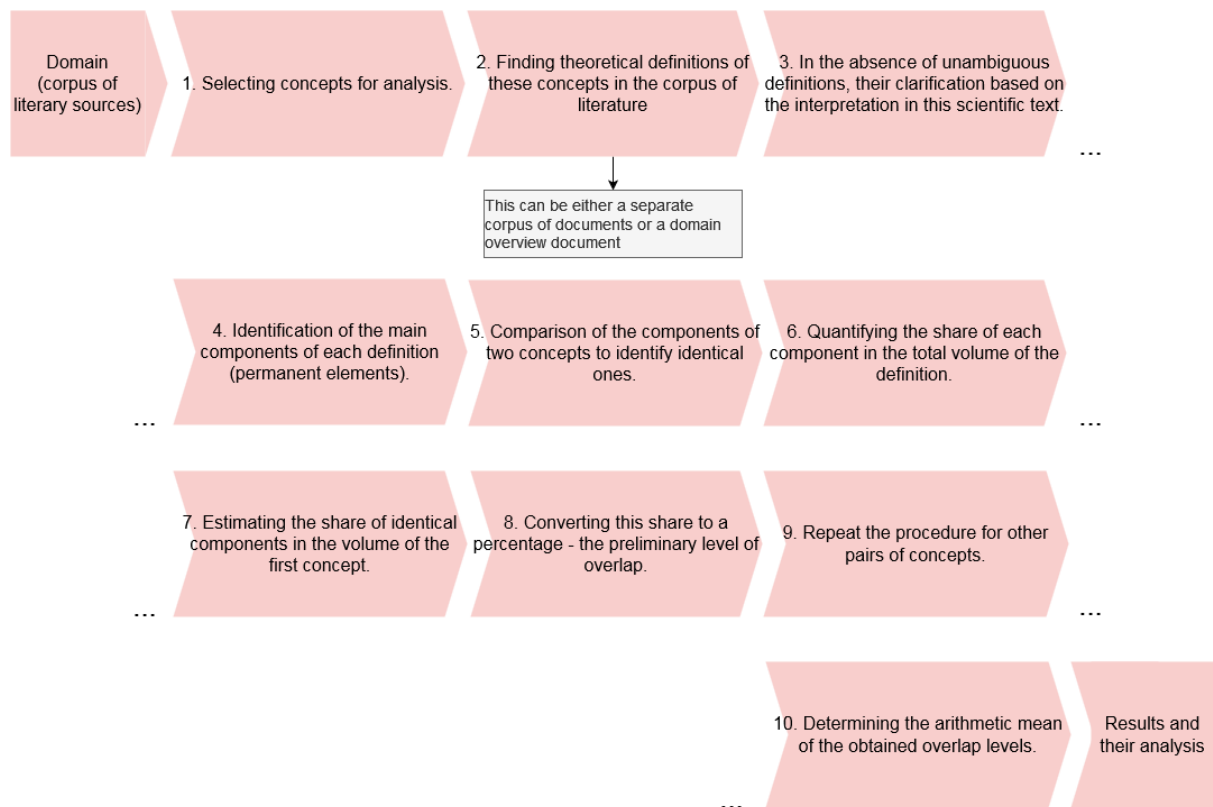


Figure 2: Method for determining the interconnections and overlap of concepts within a domain

This analysis of the interconnections and overlap of trustworthiness concepts in AI systems provides a deeper understanding of the complex nature of trustworthiness and its components. However, in order to apply this knowledge in the development and evaluation of AI systems, it is necessary to move from theoretical analysis to a formalized representation of these concepts and their integration into real systems.

With this in mind, the next step in the research is to develop a formal approach to integrating trustworthiness components into artificial intelligence systems and defining trustworthy agents. This approach will transform theoretical concepts into practical evaluation criteria that can be used to analyze the actions of AI agents.

Using the components of trustworthiness identified in the previous analysis, we will present them in a formalized form, allowing integrating these criteria into mathematical models and algorithms for assessing trustworthiness. This will not only provide a more accurate and objective assessment of the trustworthiness of AI systems, but also create a basis for the development of more ethical and trustworthy AI agents.

4. Formalization and integration of trustworthiness components into artificial intelligence systems: a model for evaluating the actions of trustworthy agents

After a detailed analysis of the interconnections between different trustworthiness concepts in AI systems, the logical next step is to consider the practical application of the knowledge gained. Understanding how these concepts intersect and interact provides the basis for developing a comprehensive approach to integrating trustworthiness principles into real-world AI systems. This section is dedicated to developing a formalized approach that would not only allow for the assessment but also for the active implementation of trustworthiness principles in AI agents. In this part of the study, we will focus on the development of such an approach. We will present a method for integrating trustworthiness components into artificial intelligence systems and propose

a formalized way to determine trustworthy agents. This approach will be based on the key components discussed in detail earlier, and it will allow moving from theoretical analysis to practical implementation of trustworthiness principles in AI systems.

This transition ensures a logical sequence between the analytical part of the study and its practical application, emphasizing the importance of preliminary analysis for the development of effective methods for implementing trustworthiness principles in AI systems.

We present an approach for integrating trustworthiness components into artificial intelligence systems and identifying trustworthy agents. We will use 11 components as criteria and present them in a formalized form. That is, we will represent the values of the components through identifiers (Id) and scores (Score).

$$\begin{aligned}
& \text{Transparency: } Trs \rightarrow Trs(trsId, trsScore); \\
& \text{Explainability: } Exp \rightarrow Exp(expId, expScore); \\
& \text{Privacy: } Prv \rightarrow Prv(prvId, prvScore); \\
& \text{Justice\&Faimess: } JstFms \rightarrow JstFms(jstfmsId, jstfmsScore); \\
& \text{Non – maleficence: } Nmlf \rightarrow Nmlf(nmlfId, nmlfScore); \\
& \text{Freedom \& Autonomy: } FrAtm \rightarrow FrAtm(fratmId, fratmScore); \\
& \text{Reliability: } Rlb \rightarrow Rlb(rlbId, rlbScore); \\
& \text{Trust: } Tr \rightarrow Tr(trId, trScore); \\
& \text{Resilience: } Rsl \rightarrow Rsl(rslId, rslScore); \\
& \text{Robustness: } Rbs \rightarrow Rbs(rbsId, rbsScore); \\
& \text{Responsibility: } Rsp \rightarrow Rsp(rspId, rspScore).
\end{aligned} \tag{4}$$

We also provide a list of active agent actions:

$$\text{AgentActions: } AgAc \rightarrow AgAc(agentId, action, result, \{concept_{i=1}^{11}\}Id) \tag{5}$$

A specific agent action is recorded, containing all the necessary connections to assess its compliance with the specified criteria. This allows to comprehensively analyze the actions of agents, checking them for compliance with the criteria presented in the form of concept identifiers $\{concept_{i=1}^{11}\}Id$.

$$\begin{aligned}
TrsAc &= \pi_{agentID, action, result} (\sigma_{trsScore \geq threshold_{trs}} \bowtie_{AgAc.trsID=Trs.trsID} AgAc); \\
ExpAc &= \pi_{agentID, action, result} (\sigma_{expScore \geq threshold_{exp}} \bowtie_{AgAc.expID=Exp.expID} AgAc); \\
PrvAc &= \pi_{agentID, action, result} (\sigma_{prvScore \geq threshold_{prv}} \bowtie_{AgAc.prvID=Prv.prvID} AgAc); \\
JstFmsAc &= \\
\pi_{agentID, action, result} & (\sigma_{jstfmsScore \geq threshold_{jstfms}} \bowtie_{AgAc.jstfmsID=JstFms.jstfmsID} AgAc); \\
NmlfAc &= \\
\pi_{agentID, action, result} & (\sigma_{nmlfScore \geq threshold_{nmlf}} \bowtie_{AgAc.nmlfID=Nmlf.nmlfID} AgAc); \\
FrAtmAc &= \\
\pi_{agentID, action, result} & (\sigma_{fratmScore \geq threshold_{fratm}} \bowtie_{AgAc.fratmID=FrAtm.fratmID} AgAc); \\
RlbAc &= \pi_{agentID, action, result} (\sigma_{rlbScore \geq threshold_{rlb}} \bowtie_{AgAc.rlbID=Rlb.rlbID} AgAc); \\
TrAc &= \pi_{agentID, action, result} (\sigma_{trScore \geq threshold_{tr}} \bowtie_{AgAc.trID=Tr.trID} AgAc); \\
RslAc &= \pi_{agentID, action, result} (\sigma_{rslScore \geq threshold_{rsl}} \bowtie_{AgAc.rslID=Rsl.rslID} AgAc); \\
RbsAc &= \pi_{agentID, action, result} (\sigma_{rbsScore \geq threshold_{rbs}} \bowtie_{AgAc.rbsID=Rbs.rbsID} AgAc); \\
RspAc &= \pi_{agentID, action, result} (\sigma_{rspScore \geq threshold_{rsp}} \bowtie_{AgAc.rspID=Rsp.rspID} AgAc).
\end{aligned} \tag{6}$$

Each action is selected according to the threshold value of the concept $threshold_{concept}$. Accordingly, we select actions that have a value above the passing threshold for each concept, which allows setting the weighting of the importance of each concept.

The set of actions is represented as an intersection, in order to select only those actions of agents that are simultaneously present in all sets of actions corresponding to different concepts.

$$ComAcs = \bigcap \{Concept\}_{i=1}^{11}. \quad (7)$$

The result contains only those actions that satisfy all the listed concepts. This is how actions are selected when a threshold value is passed for each concept.

At the same time, actions must be true in all possible contexts or scenarios.

Accordingly, we impose the following limitations:

$$\Box ComAcs(a, u) = \bigwedge \{Concept(a, u)\}_{i=1}^{11} \wedge \Diamond ComAcs(a, u). \quad (8)$$

This allows to take into account different possible contexts, such as different behavioral models of agents, different access policies in the AI system, and other factors. In other words, for all active agents a and users u , if an agent performs a complex action, this action must meet all the criteria represented in the form of concepts, and there is a possibility that the agent can perform this complex action.

This model allows evaluating the actions of active agents in artificial intelligence systems by checking their compliance with key trustworthiness criteria.

The central element of this part is the proposed model for evaluating agent actions, which integrates trustworthiness components into AI systems and identifies trustworthy agents. This model includes a formal record of agents' active actions, covering all the necessary connections to assess compliance with the trustworthiness criteria.

An important aspect of the proposed approach is the projection of agent actions onto the plane of each concept. This is done by identifier, agent action, and outcome, using a threshold value to select actions that correspond to each trustworthiness concept. The study proposes a comprehensive evaluation method that allows selecting actions that simultaneously satisfy all of these trustworthiness concepts. This takes into account various possible contexts and application scenarios, which ensures the flexibility and versatility of the approach.

Particular attention is paid to constraints and contextualization. The introduced constraints ensure that agents' actions comply with all trustworthiness criteria in different contexts, taking into account different behavioral models of agents and access policies in the AI system.

5. Experiments

Applying the method of determining the interconnections and overlap of concepts within the domain to the data presented in [1, 7], we obtain the results presented in Figure 3. According to the presentation in this paper, the total number of trustworthiness components is 11, and they were grouped into holistic concepts according to the description provided.

Transparency and Explainability have 75% overlap. Both concepts relate to the ability to explain information to users. However, Transparency includes not only explanations, but also disclosure of details about the internal workings of the system and its limitations. This means that while Explainability is about breaking down complex information into understandable parts, Transparency also includes a caveat about fully disclosing possible biases or limitations of the system.

In the Privacy – Transparency pairing, Transparency involves disclosing important information about the system's processes and decisions, which may sometimes conflict with the need to protect personal privacy. Transparency requires disclosure of certain data, but this must be balanced against privacy so as not to compromise sensitive personal information.

Both the concepts of Justice & Fairness and Non-maleficence goal to prevent harm and promote fair treatment. Justice & Fairness emphasizes inclusivity and equal opportunities for all individuals, while Non-maleficence focuses more on avoiding harm. The overlap lies in their shared goal of protecting individuals from harm, although Justice & Fairness also covers broader social aspects.

Freedom & Autonomy has 55% overlap with Privacy. Here, Privacy is an important part of individual freedom and autonomy, but it is not the only factor. Freedom & Autonomy cover a wide range of considerations, including the ability to make independent choices and live without undue interference, which goes beyond privacy alone.

Reliability refers to the consistency and trustworthiness of a system, which is the basis for trustworthiness. However, Trust also includes a sense of security and confidence that goes beyond just trustworthiness, taking into account the morality and ethical behavior of the system or its creators.

Both the terms Resilience and Robustness are related to the ability of a system to withstand and recover from disruptions. Resilience includes the ability to adapt and continue functioning over time, while Robustness emphasizes the strength and stability of the system under stress. The overlap lies in their shared emphasis on durability, although Resilience also includes an adaptability aspect.

Responsibility implies taking responsibility for one's actions and ensuring ethical behavior, which contributes to the formation of trustworthiness. The concept of Trust depends on the belief that the responsible party will act honestly and predictably, but also includes other factors such as system reliability and overall user experience.

To better understand the method, let's illustrate this process with the following example. Consider two components: Transparency and Explainability. According to the document, Transparency refers to efforts to improve Explainability, interpretability, and other forms of communication and disclosure. Explainability is defined as the ability to explain information. These two components have in common the ability to communicate and disclose information to users.

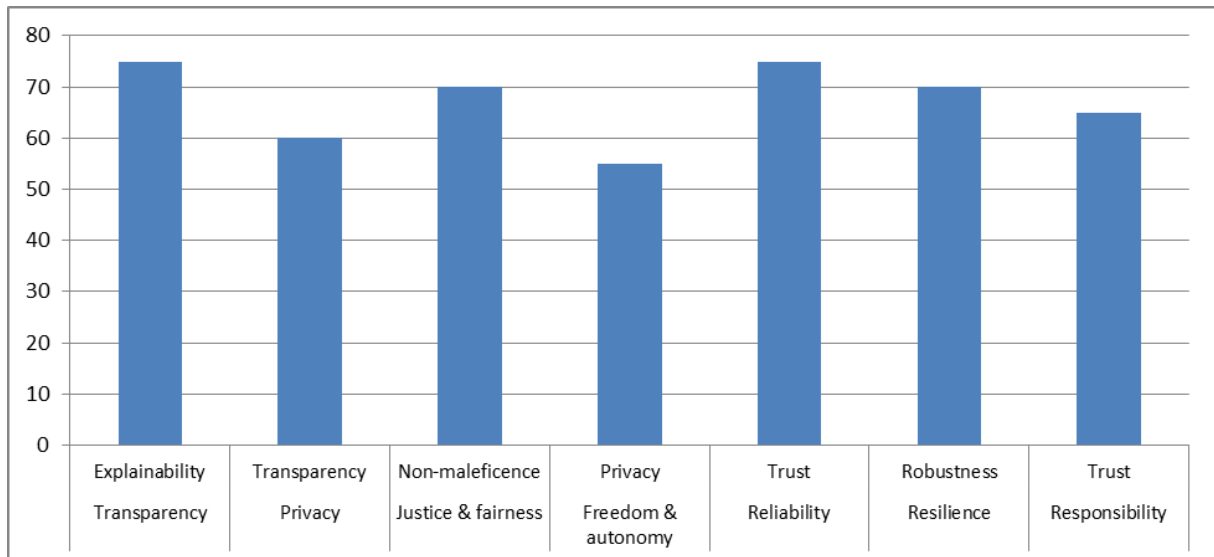


Figure 3: Level of overlap of domain concepts

The Transparency definition in the document covers efforts to improve the explainability, interpretability, and other forms of communication and disclosure of information about AI use, system code, data use, rationale for AI use, and limitations.

The definition of Explainability is the ability of an AI system or the organization that developed or implemented it to explain its decisions and actions to users or stakeholders.

The main common element of the definitions is the ability to communicate information. The author assesses what part of the definition of transparency is covered by this element. In general, the definition of transparency includes communication of different types of information (use of AI, code, data, etc.). Communication of information makes up about 3/4 of this definition.

The definition of Transparency includes:

- explainability;
- interpretability;
- communication;
- disclosure.

Of these elements, explainability, interpretability, and communication are directly related to the transfer of information. Disclosure also aims to convey information, although it is not directly related to communication.

Approximately 3/4 of the first definition is covered by its common element – the ability to communicate information. Thus, the level of overlap between transparency and explainability, calculated on the basis of the document's information, is 75%.

In general, the concepts related to trustworthiness and ethics demonstrate a high level of interconnection, which emphasizes their common nature and importance in systems. Transparency and Explainability are the most closely intertwined, reflecting their shared emphasis on the ability to explain information. Reliability and Trust also have a significant overlap, as reliability is an important basis for trust, although trust includes additional aspects.

Resilience and Robustness are closely related because of their shared focus on system resilience, although Resilience also encompasses the ability to adapt. Justice and Fairness have a similar goal to Non-maleficence, as they all aim to prevent harm, but Justice and Fairness also emphasize inclusiveness.

On the other hand, Privacy and Freedom have less overlap, as privacy is an importance, but not the only condition for freedom and autonomy. Accordingly, Responsibility and Trust have a moderate level of overlap, where responsibility contributes to trust, but trust itself encompasses other aspects such as reliability.

Thus, the concepts interact and intertwine, emphasizing their complexity and importance in creating trustworthy systems. Overall, the overlap between these 11 components can be estimated at about 67%, indicating that they are closely interrelated.

6. Discussions

The results of the concept overlap analysis demonstrate the complex interconnections and interdependencies between different components of trustworthiness in AI systems. The results of the case studies showed high levels of overlap for some pairs of concepts, such as Transparency – Explainability and Reliability – Trust, indicating their close interconnections and similar scope. However, the moderate to low levels of overlap for other pairs, such as Privacy – Transparency and Freedom & Autonomy – Privacy, indicate that these concepts, while related, differ in some key respects.

It is worth noting that the levels of overlap are relative and depend on the specific definitions and interpretations of each concept. Despite the attempt to identify the main components of the definitions as accurately as possible, some nuances and shades of meaning may be lost or misinterpreted. Therefore, the results should be considered as indicative indicators, not absolute values.

In addition, the process of determining the overlap of concepts can be somewhat subjective, as it depends on the researcher's judgment in identifying the key components of the definitions and assessing their share in the total. This can be partially compensated for by involving several experts to coordinate the estimates, but complete objectivity is hardly achievable.

The paper also proposes a formal model for the practical application of AI systems with a high level of trustworthiness. The approach provides a formal framework for assessing trustworthiness that can be adapted and extended for different types of AI systems and application scenarios.

However, there are also limitations of the proposed methods:

- The complexity and multifaceted nature of trustworthiness concepts themselves. These concepts often have fuzzy boundaries, overlap, and depend on context, which makes it difficult to formalize them clearly.
- The dynamic and rapidly changing nature of AI development. New technologies, approaches, and use cases of AI may change the understanding of trustworthiness concepts or require the introduction of new ones.

- Differences in interpretation and definitions of concepts by different stakeholders, such as developers, users, regulators, the public, etc. This can lead to differences in the formalization and analysis of interconnections.
- The interdisciplinary nature of the trustworthiness issue in AI, which encompasses technical, ethical, legal, social, and other aspects. Integration of different perspectives is a challenging task.
- Problems of validation and verification of the proposed relationship models in practice. It is difficult to conduct exhaustive empirical testing and potential influence of human factors and biases in determining the interconnections between concepts, as it often requires expert judgment and interpretation.

Thus, the problem of determining the interdependence of trustworthiness concepts in AI is complex and multifaceted, requiring an interdisciplinary approach, continuous model improvement, and close cooperation of various stakeholders.

Despite these limitations, the proposed method for assessing concept overlap is a useful tool for structuring and better understanding the interconnections between trustworthiness components in AI domains. The results of such an analysis can be the basis for further research, the development of more accurate concept definitions, and the creation of unified approaches to assessing the level of trustworthiness.

7. Conclusions

The study developed and applied a comprehensive approach to the analysis and integration of trustworthiness concepts in artificial intelligence systems. The proposed method for determining the interdependence and intersection of concepts within the domain allowed for an in-depth analysis of the interconnections between the key components of trustworthiness. This analysis revealed a complex network of interactions between different aspects of trustworthiness, emphasizing their interdependence and the importance of a comprehensive approach to building trustworthy AI systems. The analysis revealed a high level of interconnectedness between concepts, with an average overlap of about 67%.

Based on the results, a formalized approach to integrating trustworthiness components into AI systems was developed. This formal model provides a structured way to evaluate the actions of AI agents according to multiple trustworthiness criteria simultaneously. An important feature of the proposed approach is its flexibility and adaptability to different contexts and application scenarios, making it suitable for a wide range of AI systems.

The study also contributed to the understanding of the practical aspects of implementing trustworthiness principles in real AI systems. The proposed model for evaluating the actions of trustworthy agents creates a bridge between theoretical concepts and their practical implementation, providing developers with specific tools for creating trustworthy AI systems.

The main limitations of the proposed approaches to determining the interdependence of trustworthiness concepts in AI can be summarized as follows:

- The conceptual complexity and vagueness of the trustworthiness concepts themselves, their intersections, context dependence, and rapid changes in the AI industry.
- Different interpretations and definitions of the concepts by stakeholders, which makes it difficult to formalize and analyze the interconnections.
- The interdisciplinary nature of the problem, which requires the integration of technical, ethical, legal, and social aspects.
- Difficulties in empirical validation of interconnections models and the potential influence of human factors and biases in expert determination of the interconnections between concepts.

To overcome these limitations, interdisciplinary efforts are needed, involving expertise from different fields, developing flexible approaches to formalization, and constantly updating models in line with AI developments.

Overall, this research makes a significant contribution to the development of the theory and practice of building trustworthy artificial intelligence systems. It not only deepens the understanding of the complex interconnections between different aspects of trustworthiness, but also provides practical tools for integrating them into real systems. This creates a basis for further research and development in the field of ethical and trustworthy AI.

References

- [1] ISO/IEC TR 24028:2020 <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/76/77608.html> (accessed Jun 23, 2024).
- [2] Ma, Y.; Gao, X.; Zhou, W.; Chen, L. The trustworthiness measurement model of component based on defects. *Mathematical Problems in Engineering*, 2022, No. 1, 1–15. <https://doi.org/10.1155/2022/7290001>.
- [3] Chamola, V.; Hassija, V.; Sulthana, A. R.; Ghosh, D.; Dhingra, D.; Sikdar, B. A review of trustworthy and explainable artificial intelligence (xai). *IEEE Access*, 2023, **11**, 78994–79015. <https://doi.org/10.1109/ACCESS.2023.3294569>.
- [4] Jacovi, A.; Marasović, A.; Miller, T.; Goldberg, Y. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in ai. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*; FAccT '21; Association for Computing Machinery: New York, NY, USA, 2021; pp 624–635. <https://doi.org/10.1145/3442188.3445923>.
- [5] Danks, D. The value of trustworthy ai. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*; AIES '19; Association for Computing Machinery: New York, NY, USA, 2019; pp 521–522. <https://doi.org/10.1145/3306618.3314228>.
- [6] Manziuk, E.; Barmak, O.; Krak, I.; Mazurets, O.; Skrypnyk, T. Formal model of trustworthy artificial intelligence based on standardization. *CEUR Workshop Proceedings*; CEUR: Khmelnyskyi, Ukraine, 2021; Vol. 2853, pp 190–197.
- [7] Jobin, A.; Ienca, M.; Vayena, E. The global landscape of ai ethics guidelines. *Nat Mach Intell*, 2019, **1** (9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>.
- [8] Kaur, D.; Uslu, S.; Rittichier, K. J.; Durresi, A. Trustworthy artificial intelligence: a review. *ACM Comput. Surv.*, 2022, **55** (2), 39:1-39:38. <https://doi.org/10.1145/3491209>.
- [9] Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; Zhou, B. Trustworthy ai: from principles to practices. *ACM Comput. Surv.*, 2023, **55** (9), 177:1-177:46. <https://doi.org/10.1145/3555803>.
- [10] Manziuk, E. A.; Wójcik, W.; Barmak, O. V.; Krak, I. V.; Kulas, A. I.; Drabovska, V. A.; Puhach, V. M.; Sundetov, S.; Mussabekova, A. Approach to creating an ensemble on a hierarchy of clusters using model decisions correlation. *Przegląd elektrotechniczny*, 2020, **96** (9), 108–113.
- [11] Tidjon, L. N.; Khomh, F. Never trust, always verify: a roadmap for trustworthy ai? 2022. <https://doi.org/10.48550/arXiv.2206.11981>.
- [12] Carter, J.; Bitting, E.; Ghorbani, A. A. Reputation formalization for an information-sharing multi-agent system. *Computational Intelligence*, 2002, **18** (4), 515–534. <https://doi.org/10.1111/1467-8640.t01-1-00201>.
- [13] Seshia, S. A.; Sadigh, D.; Sastry, S. S. Toward verified artificial intelligence. *Commun. ACM*, 2022, **65** (7), 46–55. <https://doi.org/10.1145/3503914>.
- [14] Gillespie, N.; Curtis, C.; Bianchi, R.; Akbari, A.; Vlissingen, R. F. van. *Achieving trustworthy ai*; Report; KPMG, University of Queensland, 2020; p 40.
- [15] Tschopp, M.; Ruef, M. On trust in ai - a systemic approach; 2019.
- [16] Nalepa, G. J.; Otterlo, M. van; Bobek, S.; Atzmueller, M. From context mediation to declarative values and explainability. *IJCAI/ECAI Workshop on Explainable Artificial Intelligence (XAI-18)*; Stockholm, Sweden, 2018; pp 109–113.
- [17] Gillis, R.; Laux, J.; Mittelstadt, B. Trust and trustworthiness in artificial intelligence. *Handbook on Artificial Intelligence and Public Policy*, 2024, **33**, 153–169. <https://doi.org/10.2139/ssrn.4688574>.

- [18] Reinhardt, K. Trust and trustworthiness in ai ethics. *AI Ethics*, 2023, **3** (3), 735–744. <https://doi.org/10.1007/s43681-022-00200-5>.
- [19] Ries, S.; Habib, S. M.; Mühlhäuser, M.; Varadharajan, V. CertainLogic: a logic for modeling trust and uncertainty. *Trust and Trustworthy Computing*; McCune, J. M., Balacheff, B., Perrig, A., Sadeghi, A.-R., Sasse, A., Beres, Y., Eds.; Springer: Berlin, Heidelberg, 2011; pp 254–261. https://doi.org/10.1007/978-3-642-21599-5_19.
- [20] Wing, J. M. Trustworthy ai. *Commun. ACM*, 2021, **64** (10), 64–71. <https://doi.org/10.1145/3448248>.
- [21] Schlicker, N.; Langer, M. Towards warranted trust: a model on the relation between actual and perceived system trustworthiness. *Proceedings of Mensch und Computer 2021*; MuC '21; Association for Computing Machinery: New York, NY, USA, 2021; pp 325–329. <https://doi.org/10.1145/3473856.3474018>.
- [22] Krak, I.; Barmak, O.; Manziuk, E. Using visual analytics to develop human and machine-centric models: a review of approaches and proposed information technology. *Computational Intelligence*, 2022, **38** (3), 921–946. <https://doi.org/10.1111/coin.12289>.
- [23] Hunter, A. Interactions with ai systems: trust and transparency. *2023 IEEE Engineering Informatics*; 2023; pp 1–6. <https://doi.org/10.1109/IEEECONF58110.2023.10520626>.
- [24] Toreini, E.; Aitken, M.; Coopamootoo, K.; Elliott, K.; Zelaya, C. G.; van Moorsel, A. The relationship between trust in ai and trustworthy machine learning technologies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*; FAT* '20; Association for Computing Machinery: New York, NY, USA, 2020; pp 272–283. <https://doi.org/10.1145/3351095.3372834>.
- [25] Duenser, A.; Douglas, D. M. Who to trust, how and why: untangling ai ethics principles, trustworthiness and trust. 2023. <https://doi.org/10.48550/arXiv.2309.10318>.
- [26] Dolinek, L.; Wien, T.; Philipp, A.; Wintersberger, T. U.; Wien; Austria. Towards a generalized scale to measure situational trust in ai systems. *ACM CHI Conference on Human Factors in Computing Systems*; 2022.
- [27] Zanotti, G.; Petrolo, M.; Chiffi, D.; Schiaffonati, V. Keep trusting! a plea for the notion of trustworthy ai. *AI & Soc*, 2023. <https://doi.org/10.1007/s00146-023-01789-9>.
- [28] Krak, Y. V.; Barmak, O. V.; Mazurets, O. V. The practice implementation of the information technology for automated definition of semantic terms sets in the content of educational materials. *Problems in Programming*, 2018, No. 2–3, 245–254.
- [29] Shneiderman, B. Assessing trustworthiness. *Human-Centered AI*; Shneiderman, B., Ed.; Oxford University Press, 2022; p 0. <https://doi.org/10.1093/oso/9780192845290.003.0025>.
- [30] Ferrario, A. Justifying our credences in the trustworthiness of ai systems: a reliabilistic approach; Rochester, NY, 2023. <https://doi.org/10.2139/ssrn.4524678>.
- [31] O'Donovan, J.; Smyth, B. Trust in recommender systems. *Proceedings of the 10th international conference on Intelligent user interfaces*; IUI '05; Association for Computing Machinery: New York, NY, USA, 2005; pp 167–174. <https://doi.org/10.1145/1040830.1040870>.
- [32] Izonin, I.; Tkachenko, R.; Vitynskyi, P.; Zub, K.; Tkachenko, P.; Dronyuk, I. Stacking-based grnn-sgtm ensemble model for prediction tasks. *2020 International Conference on Decision Aid Sciences and Application (DASA)*; 2020; pp 326–330. <https://doi.org/10.1109/DASA51403.2020.9317124>.
- [33] Barmak, O. V.; Krak, Y. V.; Manziuk, E. A. Characteristics for choice of models in the ansables classification. *Problems in Programming*, 2018, No. 2–3, 171–179.
- [34] Barmak, O.; Manziuk, E.; Krak, I. Diversity as the basis for effective clustering-based classification. *CEUR Workshop Proceedings*; 2020; pp 53–67.
- [35] Holzinger, A. The next frontier: ai we can really trust. *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*; Kamp, M., Koprinska, I., Bibal, A., Bouadi, T., Frénay, B., Galárraga, L., Oramas, J., Adilova, L., Krishnamurthy, Y., Kang, B., et al., Eds.; Springer International Publishing: Cham, 2021; pp 427–440. https://doi.org/10.1007/978-3-030-93736-2_33.