Trustworthy AI Systems from Untrustworthy Components: Development von Neumann's Paradigm using Principle of Diversity

Vyacheslav Kharchenko[1] and Oleg Odarushchenko[2]

[1] *National Aerospace University KhAI, Vadym Manko str., 17, Kharkiv, 61070, Ukraine*
[2] *Poltava State Agrarian University, Skovorody str., 1/3, Poltava, 36003, Ukraine*

Abstract

The article discusses possibilities of creating trustworthy and explainable artificial intelligence (AI) and AI-based systems (AIS) using well-known von Neumann's paradigm (VNP). The models of AI and AIS quality are analyzed focusing on the most challengeable attributes related to trustworthiness of AI, safety and security of AISs. Framework of analysis, VPN formulations, methods of implementation, and stages of evolution VNP (in context of dependable and resilient systems and infrastructures) including stage of creating AISs and particularities of implementing the paradigm for various AI quality attributes are described. An approach and mathematical models describing application of diversity principles to built trustworthy AIS out of not enough trustworthy AI components (channels) are developed and investigated. A problem of AIS "immortality", the research results and future steps are discussed.

Keywords

Artificial intelligence, trustworthiness, safety, two-version AI system, common cause failure

1. Introduction

1.1. Motivation

The development and implementation of methods, tools and technologies of artificial intelligence (AI) take place in three main directions. The first direction concerns the improvement of various services to improve the quality of life, the performance of functions that provide greater comfort and convenience in everyday life, business and finance [1, 2]. The second direction is related to the use of AI for developing algorithms and control tools for industry, transport, power stations and grids, etc. [3, 4, 5, 6].

The third direction can be formulated as the one related to the reliability and security problems of artificial intelligence and by analogy with the well-known term safeware proposed by N. Levenson [7], it can be defined as AI safeware (AISaW) or AI secureware (AISeW). It is clear that this direction is related to the first two, since reliability and safety issues are very important there.

There are many cases when the unpredictable and erroneous behavior of AI means led to catastrophic consequences for services and systems of the first and second mentioned directions [8, 9]. Their analysis, as well as the forecast of an increase in AI vulnerabilities and threats of cyber attacks, as well as specific failures of intelligent systems, led to the reaction of well-known specialists with a call to slow down and even stop the development and distribution of AI products, the use of service ChatGPT, etc. [10, 11].

Therefore, it is urgent to find solutions that would harmonize the first two directions with the third one and ensure the predicted, reliable and safe functioning of AI systems. Such solutions can be based on the use of various types of testing AI behavior, application of redundancy, means of tolerance and protection from the consequences of anomalous behavior caused by hidden

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

v.kharchenko@csn.khai.edu (V. Kharchenko); odarushchenko@gmail.com (O. Odarushchenko)

0000-0001-5352-077X (V. Kharchenko); 0000-0003-3933-9637 (O. Odarushchenko)

CEUR Workshop Proceedings (CEUR-WS.org)

vulnerabilities and faults, non-compliance with requirements, as well as failures of software and hardware platforms of intelligent systems.

1.2. State of the art

In the context of safety and security, artificial intelligence is considered from three positions [12]: AI as a safety/security object (AI as an asset that must be protected, AIaSO); AI as a means of ensuring safety/security or the so-called AI powered protection (AI as an asset for protection, AIaSP); AI as a means of breaching safety/security or so-called AI powered attacks (AI as an asset for an attack, AIaSA). The same division is possible from the point of view of any other characteristics X (AIaXO, AIaXP, AIaXA) such as reliability, dependability, resilience, trustworthiness and others. According to [13], eighth main scenarios can be considered depending on cases Yes/No on three options, for example, for security and AI.

This research focuses on the first issue when it is necessary to ensure the reliability, safety, and specific characteristics of AI and AI systems using different kinds of redundancy. There are many publications related to direction AIaXO and dedicated to various aspects of assessment, development and implementation of methods and means for providing required characteristics X of intelligent systems [14, 15, 16]. However, we attended for further investigation based on classical work of John von Neumann "Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components" proposed in 1952 [17].

Concept of “wide” dependability as a federative attribute joining reliability, maintainability, availability, safety, integrity [18] became a logical development of the paradigm “building reliable computer systems with unreliable components” [19]. It is important to note that von Neumann had in mind the use of the principles of high availability and redundancy, primarily for hardware (HW) that was the main reason of computer system failures. In the process of development of computer technology and information systems, the share of HW failures gradually decreased compared to failures caused by software (SW) design faults [20].

This led to the development of the VNP by other researchers, in particular authors of [21] and [22] suggested methodology of N-version programming and the concept of building dependable systems from undependable components correspondingly. The next stage of development is the formulation of the dependability concept for a specific class of systems such as the concept of creating dependable service-oriented systems from not enough dependable web-components with uncertain characteristics [22]. The methodology of building safe systems and infrastructures from insufficiently safe systems is considered in [23]. The development of the VNP was extended to cloud IT-infrastructure and I&C systems with multipurpose maintenance [24, 25].

As a preliminary conclusion, it should be noted that, on the one hand, the specific features of AI models and tools are not taken into account in a certain way within the framework of the development of VNP principles; on the other hand, a large number of publications regarding the provision of AI trustworthiness, explainability, ethics, etc., almost do not take into account systemic problems of dependability, safety, etc.

1.3. Objectives and structure

The aim of the research is to analyze possibilities of application von Neumann’s paradigm (VNP) and VNP-based solutions to improve trustworthiness and other characteristics of AI systems. The objectives are the following:

- to discuss AI quality models [24], their characteristics and sub-characteristics to determine which of them and how can be improved by use of VNP-based approach;
- to analyze stages of VNP evolution to justify possible options for implementation of the paradigm for providing trustworthiness and other AI characteristics;

- to develop and investigate models of trustworthy and safe AI systems which are based on application of diversity principle or version redundancy (VR) on creating redundant channels and implement VNP using such approach.

The structure of the paper is as follows: Section 2 analyzes models of AI and AI systems quality focusing on the most challengeable attributes related to trustworthiness of AI, safety and security of AI systems; Section 3 discussed structure and stages of development VNP (in context of dependable and resilient systems and infrastructures) including stage of creating AI systems and particularities of implementing the paradigm for various AI quality attributes; Section 4 presents approach, solution and mathematical models describing application of diversity principles to built trustworthy AI system out of not enough trustworthy AI components (channels); Section 5 discuss the results of investigations and concerns a problem of AI and AI systems immortality; the final Section 6 summarizes and describes future research directions.

2. Model of AI systems quality: trustworthiness

In order to answer the question of how VNP can be developed and used to improve the specific dependability related characteristics of intelligent systems, it is necessary to determine the features of the AI characteristics. For this, it is suggested to use the quality model suggested in [26]. The model of AI systems quality consists of two parts or sub-models. First one is the actual artificial intelligence quality model; the second part is the quality model of the software-hardware platform that implements the functional algorithms of artificial intelligence in accordance with the requirements. Table 1 describes a simplified, so-called [26] basic AI quality model, which generally has a three-level structure and includes 32 characteristics. The basic version provides a two-level model, with five characteristics of the first level and 16 sub-characteristics that form the second level and are detailed characteristics of AI.

Table 1

Model of AI quality simplifying according to [26]

Charac-teristics	Definition	Sub-characteristics	

Ethics, ETH	The ability of AI to meet current standards of morality on the results of functioning	Fairness, FRN;
		Graspability, GRS;
		Human agency, HMA;
		Redress, RDR
			]

Lawful-ness, LFL	Ability of AI to comply with laws and regulations	No	]

Explai-nability, EXP	The ability of AI to be understood and predictable in terms of purpose and behavior	Completeness, CMT;
		Comprehensibility, MH;
		Interpretability, INP;
		Interactivity, INR;
		Transparency, TRP;
		Verifiability, VFB
			]

Respon-sibility, RSP	Ability of AI to function considering the expectations of the client (user) in accordance with ethical norms, legal regulations, as well as to inform him in case of possible violation	No	]

Trustwor-thiness, TST	Ability of AI, characterized by the degree of confidence of the stakeholders, developers, auditors, etc.) that the AI meets and performs its functions in a predictable manner	Accuracy, ACR;
		Diversity, DVS;
		Resilience, RSL;
		Robustness, RBS;
		Safety, SFT;
		Security, SCR
			]

The table defines the characteristics of ethics (ETH), lawfulness (LFL), explainability (EXP), responsibility (RSP), trustworthiness (TST), as well as a list of relevant sub-characteristics and their codification in alphabet order.

The definition of characteristics and sub-characteristics was performed on the basis of the analysis of a large number of articles and regulatory documents in accordance with the methodology described in [26]. This technique was based on semantic analysis, selection, and harmonization of definitions. It should be noted that in two years many new normative documents of various levels regarding the characteristics of AI [27]. However, in our opinion, this does not fundamentally affect the conclusions of this study.

The second part of the quality model of AI systems, namely their platforms, consists of two subsets: a subset more traditional characteristics, namely [26]: auditability (ADT), availability (AVL), controllability (CNT), effectiveness (EFS), reliability (RLB), maintainability (MNT), sustainability (SST), usability (USB); a subset of sub-characteristics (so-called group, AIG) crossing with AI trustworthiness sub-characteristics such as accuracy (ACR), diversity (DVS), resiliency (RSL), robustness (RBS), safety (SFT), security (SCR), and sub-characteristics verifiability (VFB) of explainability.

The main differences between the AI quality model and the SW quality models: presence of specific characteristics, namely ethics, legality, etc.; definition as the main characteristics of trustworthiness, explainability and responsibility; subordination of such important, primary characteristics of traditional (critical) systems as safety, security, resilience and others to the key characteristic of AI trustworthiness; filling explainability with a set of known (VFB) and relatively new sub-characteristics such as comprehensibility, interpretability and others.

3. Evolution of Von Neumann' paradigm: a stage of developing trustworthy and explainable AI systems

The development and enhancing of intelligent systems contributes to the further advancement and expansion of the VNP. Initially, the development of VNP for AI systems can take place in an understandable way, when such systems are considered as software and hardware implementation of certain functions and the key is the question of their reliable (dependable) functioning.

Then VNP can be formulated as "a reliable AI system with insufficiently reliable (AI or any other) components" in the simplest option or detail this formulation in view of the evolution of computer systems as such.

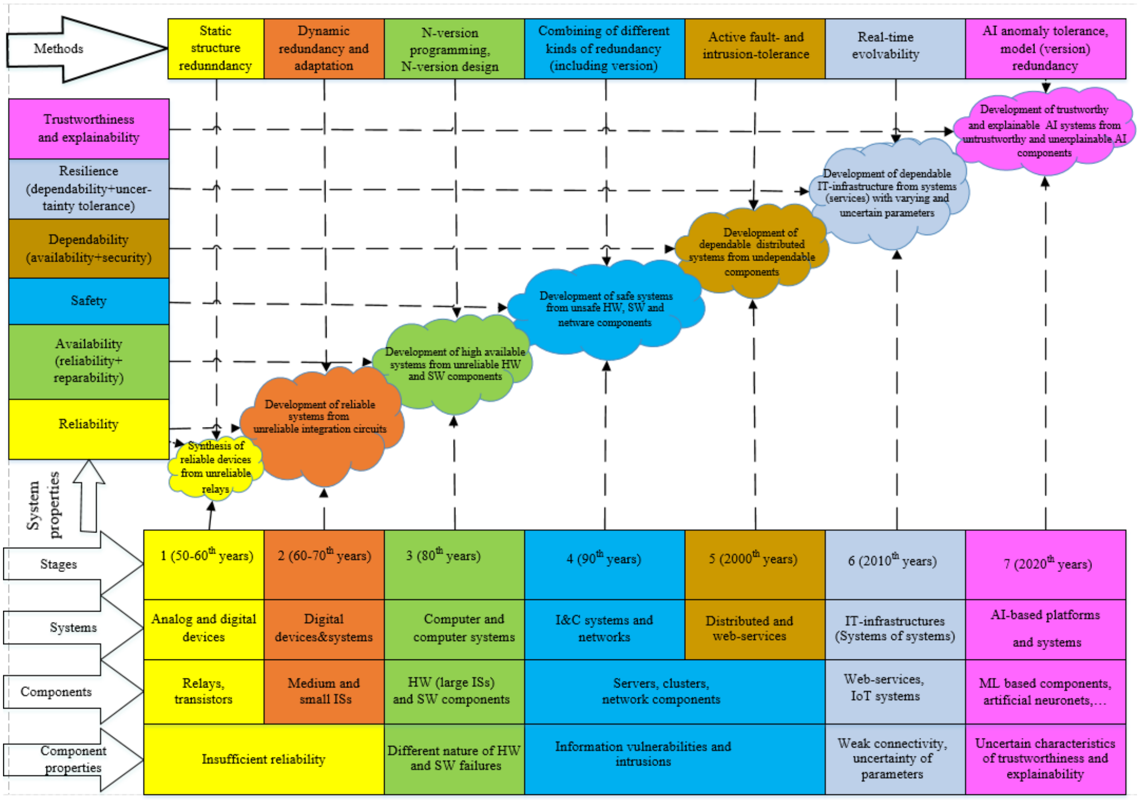
This approach is quite acceptable if we are talking about the software and hardware platform of the AI system, which is distinguished in the quality model of AI systems, the main component of which is the AI itself (the corresponding models and algorithms). However, if the specific attributes/characteristics of the qualities of AI itself, such as trustworthiness, explainability, ethics and so on, are taken into account, the paradigm should be formulated and developed more carefully.

This is due to the fact that, according to own ideology, AI can have so-called natural properties in some characteristics. In particular, it is about natural resilience, robustness, etc. [28].

Therefore, the formulation of VNP for AI systems should be based on the most important and quite specific AI characteristics, first of all, trustworthiness integrating several essential sub-characteristics, such as diversity, resilience, etc. Other specific AI characteristics (explainability, ethics, lawfulness) are interesting to analyze from the point of view of the possibility of applying and implementing VNP for their improvement.

3.1. VNP evolution analysis

Figure 1 provides describing the stages and entity of paradigm evolution in two-coordinate space “stages (systems/components/component properties) – system properties (reliability, availability, safety, dependability,...)” which is added by methods of VPN implementation.


Methods
Static structure redunndancy
Dynamic redundancy and adaptation
N-version programming, N-version design
Combining of different kinds of redundancy (including version)
Active fault- and intrusion-tolerance
Real-time evolvability
AI anomaly tolerance, model (version) redundancy
Trustworthiness and explainability
Resilience (dependability+uncer-tainty tolerance)
Dependability (availability+security)
Safety
Availability (reliability+ reparability)
Reliability
System properties
Synthesis of reliable devices from unreliable relays
Development of reliable systems from unreliable integration circuits
Development of high available systems from unreliable HW and SW components
Development of safe systems from unsafe HW, SW and netware components
Development of dependable distributed systems from undependable components
Development of dependable IT-infrastructure from systems (services) with varying and uncertain parameters
Development of trustworthy and explainable AI systems from untrustworthy and unexplainable AI components

Stages	1 (50-60th years)	2 (60-70th years)	3 (80th years)	4 (90th years)	5 (2000th years)	6 (2010th years)	7 (2020th years)
Systems	Analog and digital devices	Digital devices&systems	Computer and computer systems	I&C systems and networks	Distributed and web-services	IT-infrastructures (Systems of systems)	AI-based platforms and systems
Components	Relays, transistors	Medium and small ISs	HW (large ISs) and SW components	Servers, clusters, network components		Web-services, IoT systems	ML based components, artificial neuronets,…
Component properties	Insufficient reliability		Different nature of HW and SW failures	Information vulnerabilities and intrusions		Weak connectivity, uncertainty of parameters	Uncertain characteristics of trustworthiness and explainability

Figure 1: Diagram of VNP evolution including stage of AI systems development

The objects of analysis are evolutionary stages, systems, components, and their properties, which are the material embodiment of the respective stage. Elements of evolution-based methodology for analyzing transformation of VNP were suggested and investigated in work [29] describing evolution stages without AI systems. A formula of VNP can be presented by the following tuple:

$$VNP = < \{Proc\}, \{CharS\}, \{Syst\}, From, \{CharC\}, \{Comp\}>, \quad (1)$$

where {Proc} – a set of processes (synthesis, development, creation,…); {CharS} – a set of characteristics of system (reliable, dependable, safe, secure, resilient, trustworthy,…); {Syst} – a set of systems (device, system, infrastructure,… that are synthesized, developed, created,…); From – a preposition connecting system Syst and its components Comp (Syst X from Comp Y); {CharC} – a set of characteristics of component that part of system Syst (usually they are antipodes of CharS: unreliable or not enough reliable, unsafe, undependable, unresilient, untrustworthy); {Comp} – a set of components (relay, integral circuit/chip, hardware, software, system,… used to built system Syst.

It is clear that a part of elements-tuples of the SVNP will be empty. Figure 1 describes a fragment of VNP evolution during seven stages (1950-2020 years), every of which is presented by one of the formulations (3) beginning from initial simplest expression “Synthesis of reliable devices form unreliable (not enough reliable) relays” (1950 years, applied method – structure static redundancy) and formulation of 2010th years “Development (deployment) of dependable IT-infrastructures from undependable systems (services with uncertain and varying characteristics)”. As for AI systems (seventh stage, 2020 years), several VNP formulations are also possible depending on which characteristics or sub-characteristics are considered. The most general is “Development of trustworthy (and/or explainable) AI systems from untrustworthy (unexplainable) AI components”.

3.2. VNP for AI systems: options

Table 2 describes possibilities of VNP implementation for AI systems taking into account various characteristics and sub-characteristics.

Table 2
Analysis of methods for VNP implementation for AI characteristics

AI characteristics	AI subcharacteristics	VNP application (Y/N)	Methods applied for VNP implementation	Notes
TST	In general	Yes	Methods applied for sub-characteristics	Compatability of these methods and means should be taken into account
	DVS	Yes	Version redundancy	Problem is developing and choosing versions with maximal diversity metrics
	RSL	Yes	Proactivity, version and structural redundancy, dynamical reconfiguration	The same. Besides, problems is in development and implementation of means providing proactivity and dynamical reconfiguration
	RBS	Yes	Version redundancy	Problem is developing and choosing versions with maximal diversity in point of view input data
	SFT	Yes	Version and struc-tural redundancy	Main criteria of implemented methods and means is minimizing risks of CCF
	SCR	Yes/ No	VR for integrity and accessibility	Use of VR must be defined considering impact on the different security attributes
	ACR	Yes	Version, time, structural redundancy	Main criteria is decreasing total errors
EXP	In general	No	-	Redundancy can increase level of unexplainability
RSP	In general	Yes	Version, structural redundancy, dynamical reconfiguration	RSP is dependent on credibility, explainability, etc. It should be taken into account on choice of the methods
ETH	In general	No	-	VR can be applied if the versions will provide different reactions on situations with ethically unacceptable alternatives
LFL	In general	No	-	VR can be applied if the versions will provide different reactions on so called situations with unacceptable alternatives in point of view lawfulness

VNP can be developed and applied to technologies in which AI-based solutions directly and effectively embedded. This applies, in particular, to IoT/IoE technologies. Known branches of these technologies are integrated with AI, namely the so-called Internet of Artificial Intelligence Things (IoAIT), Internet of Artificial Intelligence (IoAI), Artificial Intelligence of Things (AIoT) and so on.

AIoT is defined as the combination of Artificial intelligence technologies with the IoT infrastructure to achieve more efficient IoT operations, improve human-machine interactions and enhance data management and analytics.

Hence, VNP that was formulated for IoT system as Dependable Internet of Undependable (not enough Dependable) Things, DIoUDT, and implemented by application of redundant nodes and communications, can be reformulated as a Trustworthy Artificial Intelligence from Untrustworthy (not enough Trustworthy) Things, TAIoUTT. More problematic is developing such systems considering characteristics of explainability and ethics.

4. Application of diversity for creating trustworthy AI systems out of untrustworthy AI components

4.1. Principle of diversity for developing trustworthy and safe AI systems

In the early stages of development, VNP was based exclusively on the use of structural redundancy. Later, when the productivity of electronic components and computers increased, and, as a result, some reserves of time appeared, the redundancy based on the use of such reserves added the structural redundancy. Therefore, the temporal redundancy strengthened the structural one and improved the dynamics of the development of some branches of the VNP.

However, the next idea, the idea of N-version programming became truly revolutionary for VNP [21]. Later, it developed into the principles of multi-version design and multi-version systems [22, 25]. Version redundancy together with structural and temporal redundancy formed a specific redundancy base for VNP and fault- and intrusion tolerant systems of a very wide class.

The application of structural and temporal redundancy cannot protect the AI system, as well as digital systems in general, from failures caused by design faults, programming errors, giving rise to so-called common cause failures (CCF), since they are replicated on backup channels and repeated with additional phases of calculations.

Version redundancy, which equates to the principle of diversity, when the same task is implemented using different programming languages and teams of programmers, developers, verifiers, different environments and development tools, different software and hardware platforms, different life cycle models etc., significantly reduces the CCF risks [30].

The implementation of the principle of diversity for intelligent systems in terms of actual AI models and algorithms has its essential specificity. The task of classifying and researching types of diversity for AI is an independent task.

Diversification of the development of models and algorithms can be based on; different methods of construction (synthesis) of neural net model solutions; different methods of their training and retraining; diverse datasets, etc.

4.2. Models for assessing two-version AI systems

4.2.1. Theoretical-set model

Let’s consider two channel AI system that work according to principle “1 out of 2”. Both channels are equipped by embedded testing means that check up/down states of hardware and software components. If the channels have been implemented using the same version Va of AI (for example identical artificial neural networks), sets of input data both channels are described by the formula:

$$IDva = IDvao \ U \ IDvau \ U \ IDvas \ , \tag{2}$$

where IDvao is a subset of input data (ID) on which both AI channels using one version work correctly; IDvau is a subset of input data on which work of both AI channels is uncertainty; IDvas is a subset of input data on which both AI channels can work unsafely.

If the channels have been developed by use of two different versions (with different structure of neural networks or different techniques and datasets for learning and so on [31]), a set of input data can be divided into the following subsets:

- input data of correct behavior of the AI versions Va (a set of input data IDvao) and Vb (a set of input data IDvbo);
- input data (ID) of correct behavior both AI versions Va and Vb described by set

$$IDvabo = IDvao \ \Omega \ IDvbo; \tag{3}$$

- ID of correct behavior of the AI versions Va or Vb only described by two subsets

$$IDvao_{\backslash} = IDvao \backslash IDvabo, \tag{4}$$

$$IDvbo_{\backslash} = IDvbo \backslash IDvabo; \tag{5}$$

- ID of uncertain behavior of AI versions Va (a set of input data IDvau) and Vb (a set of input data IDvbu);
- ID of uncertain behavior both AI versions Va and Vb described by set

$$IDvabu = IDvau \ \Omega \ IDvbu; \tag{6}$$

- ID of uncertain behavior of the AI versions Va or Vb only described by two subsets

$$IDvau_{\backslash} = IDvau \backslash IDvabu, \tag{7}$$

$$IDvbu_{\backslash} = IDvbu \backslash IDvabu; \tag{8}$$

- ID of unsafe behavior of AI versions Va (set of ID, IDvas) and Vb (set of ID, IDvbs);
- ID of unsafe behavior both AI versions Va and Vb described by set

$$IDvabs = IDvas \ \Omega \ IDvbs; \tag{9}$$

- ID of unsafe behavior of the AI version Va or version Vb only described by two sets

$$IDvas_{\backslash} = IDvas \backslash IDvabs, \tag{10}$$

$$IDvbs_{\backslash} = IDvbs \backslash IDvabs. \tag{11}$$

Combinations of uncertain and unsafe states of versions are not anlysed, because such cases are identified as an unsafe (set IDvabs). Note, that sets IDvao and IDvbo are defined by datasets that were used for learning and are expected for versions Va and Vb.

4.2.2. Probabilistic models

Let's develop probabilistic models of one- and two-version two channel AI-systems. Assumptions for these models are the following: failure of the checking and reconfiguration means is failure of AI-systems; failures of the versions (SW and HW) are independent; switching on/off the channels in case of their failures is carried out instantly.

Dependency of trustworthy work probability for two-channel (duplicated) one-version AI system on the probabilities of states can be calculated using the following formula:

$$P_{AI1} = [P_0 + (1 – P_0)\ P_D]\ (1 – P_{U1} – P_{S1})\ P_{R1}, \tag{12}$$

where: P_0 is a probability of the channel up-state; P_{U1} is a probability that at the inputs there will be data from the set IDvau, which will lead to the transition of the channels and AI system to an uncertain state; P_{S1} is a probability that at the inputs there will be data from the set IDvas, which will lead to the transition of the channels and AI system to an unsafe state; P_{R1} is a probability of the checking and reconfiguration means for one-version two-channel AI system.

This indicator for two-channel and two-version AI system is determined as follows:

$$P_{AI2} = [P_0 + (1 – P_0)\ P_D]\ (1 – P_{U2} – P_{S2})\ P_{R2}, \tag{13}$$

where: P_{U2} is a probability that at the inputs there will be data from the set IDvabu, which will lead to the transition of the channels and AI system to an uncertain state; P_{S2} is a probability that at the inputs there will be data from the set IDvabs, which will lead to the transition of the channels and AI system to an unsafe state; P_{R1} is a probability of the checking and reconfiguration means for two-version two-channel AI system.

Sure that $P_{U2} < P_{U1}$, $P_{S2} < P_{S1}$, $P_{R1} > P_{R2}$. Let's calculate

$$\delta P_{AI2/AI1} = P_{AI2} / P_{AI1} = [(1 – P_{U2} – P_{S2}) / (1 – P_{U1} – P_{S1})]\ P_{R2} / P_{R1} \approx$$

$$[(1 – \beta_{U2} – \beta_{S2}) / (1 – \beta_{U1} – \beta_{S1})]\ P_{R2} / P_{R1}, \tag{14}$$

where: $\beta_{U1} = Card\ IDvau / Card\ IDva$, $\beta_{S1} = Card\ IDvas / Card\ IDva$ are coefficients (metrics) evaluating relative parts of input data which will lead to the transition of the channels and one-version AI system to uncertain and unsafe states correspondingly;

$\beta_{U2} = Card\ IDvabu / Card\ IDva$, $\beta_{S2} = Card\ IDvabs / Card\ IDva$ are coefficients (metrics) evaluating relative parts of input data which will lead to the transition of the channels and two-version AI system to uncertain and unsafe states correspondingly.

If assume that $P_{R1} \approx P_{R2}$, formula (13) will be the following:

$$\delta P_{AI2/AI1} \approx (1 - \beta_{U2} - \beta_{S2}) / (1 - \beta_{U1} - \beta_{S1}), \quad (15)$$

$$\delta Q_{AI1/AI2} = (1 - P_{AI})/ (1 - P_{AI2}) \approx (\beta_{U1} + \beta_{S1}) / (\beta_{U2} + \beta_{S2}). \quad (16)$$

If part of uncertain and unsafe input data IDvau and IDvas for version of one-version AI system equals O.1 and part of uncertain and unsafe input data IDvabu and IDvabs for versions of two-version AI system equals O.02, risk (probability) of unsafe or potentially unsafe states will be decreased in 5 times.

It should be noted that the presented analytical models for calculating relevant indicators do not take into account other types of diversity and corresponding faults that can lead to system failures. AI systems are SW-HW solutions, and therefore, like any system with software or programmable hardware means, they are subject to design faults caused by developer errors, imperfection of the technical specifications and so on.

To tolerate their consequences, the principle of diversity is applied, but it refers purely to the use of different programming languages and technologies, hardware and software platforms, etc. [31, 32]. Note, that such diversity does not tolerate the specific problems of using AI models, which were discussed above. This is confirmed by the experience of using the driverless automotive systems [33], where diversity is actually used to protect against software (design) faults and certain HW (physical) faults, which in its absence can cause Common Cause Failures (CCFs) of redundant structures. However, such HW-SW diversity does not protect AI systems against vulnerabilities and complex kinds of CCFs caused by uncertainty of model behavior, and provide a trust guarantee of safe functioning.

Let's analyze models for one and two model-version AI-systems with one- and two-version SW (systems AI1-1, AI1-2, AI2-1, AI2-2, where first digital describes number of model versions, second one is number of SW versions) considering reliability of SW. The following formulas describe probabilities of up-states of these systems:

$$P_{AI1-1} = [P_{HW} + (1 - P_{HW})\, P_D\,]\, (1 - P_{U1} - P_{S1})\, P_{SW}P_{R1}, \quad (17)$$

$$P_{AI1-2} = [P_{HW}P_{SWr} + (1 - P_{HW}P_{SWr})\, P_D\,]\, (1 - P_{U1} - P_{S1})\, P_{SWa}P_{R1}, \quad (18)$$

$$P_{AI2-1} = [P_{HW} + (1 - P_{HW})\, P_D\,]\, (1 - P_{U2} - P_{S2})\, P_{SW}P_{R2}, \quad (19)$$

$$P_{AI2-2} = [P_{HW}P_{SWr} + (1 - P_{HW}P_{SWr})\, P_D\,]\, (1 - P_{U2} - P_{S2})\, P_{SWa}P_{R2}\,, \quad (20)$$

where (in case of independency of failures of HW and SW components of the channels): $P_0 = P_{HW}$ $P_{SW;}$ P_{HW} and P_{SW} are probabilities of the HW and SW up-states; $P_{SW} = P_{SWr}\, P_{SWa}$, P_{SWr} and P_{SWa} are probabilities of the SW up-state considering relative and absolute faults.

Hence, taking into account expressions (12-15, 17-20) and the insignificant difference in the probabilities of up-state the checking and reconfiguration means for the systems, the following formulas for calculating their relative differences can be given:

$$\delta P_{AI1-2/AI1-1} = P_{SWr}\, [P_{HW}P_{SWr} + (1 - P_{HW}P_{SWr})\, P_D\,] / [P_{HW} + (1 - P_{HW})\, P_D\,], \quad (21)$$

$$\delta P_{AI2-2/AI1-2} = (1 - P_{U2} - P_{S2}) / (1 - P_{U1} - P_{S1}) \approx$$
$$(1 - \beta_{U2} - \beta_{S2}) / (1 - \beta_{U1} - \beta_{S1}), \quad (22)$$

$$\delta P_{AI2-2/AI2-1} = P_{SWr}\, [P_{HW}P_{SWr} + (1 - P_{HW}P_{SWr})\, P_D\,]\, /\, [P_{HW} + (1 - P_{HW})\, P_D\,]. \quad (23)$$

These expressions allow specifying impact of diversity on the two levels and formulating requirements to AI versions. Formula (22) describes a simple linear dependency of AI2-2 system benefits in comparison with AI2-1. Formula (23) is traditional for evaluation of increasing safety due to using diversity in duplicated systems.

5. Discussion

5.1. VNP for AI systems: way to immortality

The VNP began its path of implementation on simple relay and then electronic devices.. Now, after 60 years, we have come (we are approaching) to the synthesis of reliable (trustworthy...) AI from insufficiently reliable (trustworthy) components. John von Neumann wrote about reliable organisms, not about conventional technical systems. He tried to expand the scope of research and consider bio-technical systems as certain heterogeneous formations. Perhaps the next his step would be related to purely biological systems and the ideas of redundancy and reconfiguration would be extended to them.

Considering the presence of a large number of AI quality characteristics, it is necessary to consider the possibilities of building “better” systems from the “worst” components separately for each of the characteristics. So, within the framework of this article, we got closer to artificial formations ("a little bit of organisms", since AI is a step in this direction) and then it remains to take the next step to reliable "organisms".

That is, we can conclude that, firstly, the VPN circle closes, so to speak, in the sense of "organisms", and the introduction of artificial intelligence, consideration of its dual nature as an object and a means of ensuring reliability is a way to create and research such reliable organisms. The transfer of AI to a new technological base, such as creating a bio-technical system, can be exemplified by the development of the Australian startup, Cortical Labs [34]. They are working on a new type of artificial intelligence that combines lab-grown human brain cells with computer chips. This approach further bridges the gap between AI and humans, potentially increasing the level of the various threats.

Secondly, a reliable organism made of insufficiently reliable components is a step to immortality! The path to it can be made both by reserving biological components and by replacing them with artificial means. As noted in [35], there are two threats to the problem of AI immortality. On the one hand, it is the possible loss of renewal that is a consequence of death, which can create the risk of weakening future generations and possible conflicts between them. On the other hand, AI immortality could create an "artificial intelligence-human" relationship similar to a "god-mortal" relationship. But in the context of this study, it is about immortality in view of the various types of failures of AI systems and the possible embedding of components to continue functioning.

Thirdly, since it is about how to build an organism with a specified value of reliability that is assessed by probability of up-state for a required time, a person can get a tool to check and control this level. Therefore, this person, which may also be a mean of artificial intelligence, will have multiple strategies for ensuring reliability through proactive repair, redundancy and reconfiguration to provide way to immortality.

5.2. Features and limitations of applying diversity for proving VNP to ensure AI trustworthiness

Despite the fact that such specific characteristics of AI as trustworthiness, explainability, ethics are top for intelligent systems, the characteristics of reliability, security and resilience are more understandable and familiar for developers and customers. This article did not have the task of delving deeply into the safety and security problems of AI, but they should always be close to the problems of evaluating and ensuring the necessary level of specific characteristics of AI, first of all, trustworthiness and explainability. The issue of determining qualitative, and especially quantitative, requirements for these characteristics is quite complex.

The principle of diversity can be quite effective from the point of view of as safety and trustworthiness. Regarding the well-known sub-characteristics of security, namely integrity, confidentiality and accessibility, the situation is somewhat more complicated, since application of diversity increases integrity and accessibility measures but can create risks for confidentiality considering the rule of “weak link”. Therefore, for a more thorough analysis and evaluation, it is necessary to consider one more the third level in addition to the characteristics and sub-characteristics.

Complex and contradictory is the question of the expediency of using the diversity principle for increasing ethical and lawfulness indicators. Table 2 provides a conclusion about the inappropriateness of using VNP to improve ethical indicators and notes that VR can be applied if the versions will provide different reactions on so called situations with ethically unacceptable alternatives (SEUA). It is theoretically possible to build versions in which such situations will be diversified to provide reducing the risk of SEUA for a common reason. The practical implementation of such a principle, for example, for driverless cars, requires the careful specification of the list of SEUAs, and the development of several AI versions using diversified techniques. Such an opportunity and ways of its implementation are quite complex and interesting for future research.

6. Conclusion and future work

The main contribution of this study is a framework for the formal presentation of VNP and its application in intelligent systems and methods of its implementation to ensure trustworthiness and other specific characteristics of AI. The importance of the proposed models and methods is that they can be detailed and developed to evaluate the feasibility and ways of using VNP methodology in creating trustworthy and safe AI systems.

However, in our opinion, the AI safe/secureware engineering has to be separated as an independent branch of intelligent engineering. It is fully justified in view of the uncertainty, threats and risks associated with the use of AI systems in critical domains and impact on consequences caused by failed/unpredictable behavior.

Diversity is a really important and promising principle that can be used to provide key trustworthiness, safety and security characteristics of AI systems. This applies to all elements of the triad AIaXO-AIaXP-AIaXA and scenarios of its implementation.

Future investigation could be connected with development of detailed models, methods and tools for assessing and providing specific characteristics and sub-characteristics of AI and AI systems. These steps should be added by enhancing and developing regulation requirements and justification of quantitative values for them.

References

[1] A. M. Rahmani, B. Rezazadeh, M. Haghparast, W. C. Chang and S. G. Ting, Applications of Artificial Intelligence in the Economy, Including Applications in Stock Trading, Market Analysis, and Risk Management, IEEE Access, 11 (2023): 80769-80793. doi:10.1109/ACCESS.2023.3300036.

[2] S. Sai, A. Garg, K. Jhawar, V. Chamola, B. Sikdar, A Comprehensive Survey on Artificial Intelligence for Unmanned Aerial Vehicles, IEEE Open J. Veh. Technol., 4 (2023): 1–26. doi:10.1109/ojvt.2023.3316181.

[3] A Kashtalian, S Lysenko, B Savenko, T Sochor, T Kysil, Principle and method of deception systems synthesizing for malware and computer attacks detection, Radioelectron. Comput. Syst., 11 (2023): 112–151. doi:10.32620/reks.2023.4.10.

[4] Chen, X.; Ma, D.; Liu, R.W. Application of AI in Maritime Transportation. J. Mar. Sci. Eng., 12, 439 (2024): 1-4. doi:10.3390/jmse12030439.

[5] Xiao, G.; Yang, D.; Xu, L.; Li, J.; Jiang, Z. The Application of Artificial Intelligence Technology in Shipping: A Bibliometric Review. J. Mar. Sci. Eng. 12, 624 (2024): 1-21. doi:10.3390/jmse12040624.

[6] Carpanzano, E. Editorial of the Special Issue “Advances in Artificial Intelligence Methods Applications in Industrial Control Systems”. Appl. Sci., 13, 16 (2023): 1-24 doi:10.3390/app13010016.

[7] Leveson N. Safeware: System safety and computers. Boston: Addison-Wesley; 1995.

[8] Williams, R.; Yampolskiy, R. Understanding and Avoiding AI Failures: A Practical Guide. Philosophies, 6, 53 (2021): 1-25. doi:10.3390/philosophies6030053

[9] Steimers, A.; Schneider, M. Sources of Risk of AI Systems. Int. J. Environ. Res. Public Health, 19, 3641 (2022): 1-26. doi:10.3390/ijerph19063641.

[10] W. Knight, P. Dave, In Sudden Alarm, Tech Doyens Call for a Pause on ChatGPT, Business, March 29, 2023. URL: https://www.wired.com/story/chatgpt-pause-ai-experiments-open-letter/

[11] The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. URL: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

[12] O. Illiashenko, V. Kharchenko, I. Babeshko, H. Fesenko, F. Di Giandomenico, Security-Informed Safety Analysis of Autonomous Transport Systems Considering AI-Powered Cyberattacks and Protection, Entropy 25.8, 1123 (2023): 1-35. doi:10.3390/e25081123.

[13] O. Veprytska, V. Kharchenko, AI powered attacks against AI powered protection: classification, scenarios and risk analysis, 12th International Conference on Dependable Systems, Services and Technologies (DESSERT), IEEE, 2022, pp. 1-7. doi:10.1109/dessert58054.2022.10018770.

[14] Ding, W.; Abdel-Basset, M.; Hawash, H.; Ali, A.M. Explainability of AI methods, applications and challenges: A comprehensive survey. Inf. Sci., 615 (C) (2022): 238–292. doi: 10.1016/j.ins.2022.10.013.

[15] Hohma, E.; Lütge, C. From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems. AI, 4, (2023): 904-925. doi:10.3390/ai4040046.

[16] Wanner, J.; Herm, L.-V.; Heinrich, K.; Janiesch, C. The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. Electron. Mark., 32 (2022): 2079–2102.

[17] Neumann, J. von. Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components. Ed. by C. E. Shannon, J. McCarthy, Princeton University Press, 1956, pp. 43-98. doi:10.1515/9781400882618-003

[18] Avizienis, A.; Laprie, J.-C.; Randell, B.; Landwehr, C. Basic concepts and taxonomy of dependable and secure computing. IEEE Trans. Dependable Secur. Comput., 1 (2004): 11–33, doi:10.1109/TDSC.2004.2

[19] Z. Peng, Building reliable embedded systems with unreliable components, ICSES 2010 Intern. Conference on Signals and Electronic Circuits, Gliwice, Poland, 2010.

[20] M. Grottke, A. P. Nikora and K. S. Trivedi. An empirical investigation of fault types in space mission system software, Intern. Conf. on Dependable Systems & Networks, Chicago, IL, USA, 2010: 447-456. doi: 10.1109/DSN.2010.5544284.

[21] A. Avizienis, The N-Version Approach to Fault-Tolerant Software, IEEE Transactions on Software Engineering, vol. SE-11, 12 (1985): 1491-1501. doi: 10.1109/TSE.1985.231893.

[22] Gorbenko, A., Kharchenko, V. and Romanovsky, A. On composing Dependable Web Services using independent web components. International Journal of Simulation and Process Modeling, vol. 3(1/2) (2007): 45-54. doi: 10.1504/IJSPM.2007.014714.

[23] Y. Brezhniev. Multilevel Fuzzy Logic-Based Approach for Critical Energy Infrastructure's Cyber Resilience Assessment, 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT), Leeds, UK, 2019, pp. 213-217, doi: 10.1109/DESSERT.2019.8770034.

[24] E. Brezhniev, O. Ivanchenko. NPP-Smart Grid Mutual Safety and Cyber Security Assurance, Research Anthology on Smart Grid and Microgrid Development, IGI-Global, USA, 2022. doi: 10.4018/978-1-6684-3666-0.ch047.

[25] Ponochovnyi Y., Kharchenko V. Dependability Assurance Methodology of Information and Control Systems. Radioelectron. Comput. Syst., 3 (2020): 43–58. doi: 10.32620/reks.2020.3.05.

[26] V. Kharchenko, H. Fesenko, O. Illiashenko, Quality Models for Artificial Intelligence Systems: Characteristic-Based Approach, Development and Application, Sensors 22.13, 4865(2022): 1-32. doi: 10.3390/s22134865.

[27] ISO/IEC TR 24030:2024. Information technology. Artificial Intelligence. Use cases

[28] Moskalenko, V.; Kharchenko, V.; Moskalenko, A.; Kuzikov, B. Resilience and Resilient Systems of Artificial Intelligence: Taxonomy, Models and Methods. Algorithms, 16, 165 (2023): 1-34. doi: 10.3390/a16030165.

[29] V. Kharchenko, A. Gorbenko, Evolution of von Neumann's paradigm: Dependable and green computing, East-West Design & Test Symposium (EWDTS 2013), Rostov on Don, Russia, 2013, pp. 1-6, doi: 10.1109/EWDTS.2013.6673090.

[30] Diversity Strategies for Nuclear Power Plant Instrumentation and Control Systems, NUREG/CR-7007 Office of Nuclear Regulatory Research, 2010.

[31] Tao Zhou (ed.) Advanced Artificial Intelligence Models and Applications, Mathe-matics, October 2023, 182 p. doi: 10.3390/books978-3-0365-9133-9.

[32] Dini, P.; Diana, L.; Elhanashi, A.; Saponara, S. Overview of AI-Models and Tools in Embedded IIoT Applications. Electronics, 13, 2322 (2024): 1-27. doi: 10.3390/electronics13122322.

[33] Julitz, Tim Maurice; Tordeux, Antoine; Löwer, Manuel. Computer-Aided Design of fault-tolerant Hardware Architectures for Autonomous Driving Systems. International Conference on Engineering Design, ICED23, Bordeaux, France, 24-28 July 2023, pp. 1047-1056. doi: 10.1017/pds.2023.105.

[34] Daniel Van Boom. The Australian startup building computers out of human brains, Capital Brief, 18 January 2024 https://www.capitalbrief.com/article/the-australian-startup-building-computers-out-of-human-brains-a30d6821-cf2d-47db-b0cf-85a6fe294cb8/preview/

[35] Mike Thomas. Risks and Dangers of Artificial Intelligence, Built-In. July 25, 2024. URL: https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence/