Trustworthy AI: a fuzzy-multiple method for evaluating ethical principles in AI regulations

Trustworthy AI: A Fuzzy-Multiple Method for Evaluating Ethical Principles in AI Regulations

line 1: 1st Oksana Adamyk
line 2: *Loughborough Business School*
line 3: *Loughborough University*
line 4: Loughborough, United Kingdom
line 5: o.o.adamyk@lboro.ac.uk

line 1: 2nd Oksana Chereshnyuk
line 2: *Department of Financial Control and Audit*
line 3: *West Ukrainian National University*
line 4: Ternopil, Ukraine
line 5: oksana.duda@gmail.com

line 1: 3rd Bogdan Adamyk
line 2: *Aston Business School*
line 3: *Aston University*
line 4: Birmingham, United Kingdom
line 5: b.adamyk@aston.ac.uk

line 1: 4th Serhii Rylieiev
line 2: *Department of Finance, Accounting and Taxation*
line 3: *Chernivtsi Institute of Trade and Economics of the State University of Trade and Economics*
line 4: Chernivtsi, Ukraine
line 5: ryleev76@gmail.com

***Abstract*—In this study, we investigated the ethical principles of trustworthy AI and differentiated five prime factors essential for developing trust in AI and most widely presented in regulatory guidelines worldwide. By utilizing Fuzzy Logic Toolbox in MATLAB 9.4, we evaluated the impact of primary ethical principles on trustworthy AI systems in a systematic and structured manner. We discovered that the principle of Fairness and Non-discrimination is the most influential for the development of trustworthy AI, as it is the most represented in the regulatory guidelines. The proposed model offers two main benefits for developers and deployers of AI systems, including predicting the potential public trust in AI systems and assessment compliance with the regulatory frameworks. To ensure the continued trustworthiness of AI systems, the model should be used at all stages of the software life circle, including during development, before placing the system on the market, and at the stage of use to monitor compliance with the safeguards declared to users.**

Keywords—artificial intelligence, regulation, trustworthy AI, ethical principles of AI, trust.

I. INTRODUCTION

A. *Trust and trustworthy AI: the social need and regulatory initiatives*

The increasingly pervasive role of Artificial Intelligence (AI) in our societies is radically changing the way that social interaction takes place within all fields of knowledge (Gianni, R.et al. [1]). Along with the benefits of artificial intelligence, such as improving the ability to analyze and process vast amounts of data, the use of artificial intelligence systems also carries certain risks related to the self-learning capabilities of cognitive technologies. In addition, the implementation of AI technology also raises questions about its societal impacts, "which may be difficult to anticipate, identify or measure (e.g. on democracy, the rule of law and distributive justice, or on the human mind itself)" [2].

As a result, the specific risks and dangers associated with artificial intelligence, combined with the need to control the direction of AI development, have led to the creation of many national strategic documents. To effectively address the challenges and opportunities presented by AI systems, global solutions are necessary. Nowadays, there are 39 governance frameworks, were issued by national governments and intergovernmental bodies such as the UN High-Level Committee on Programmes, the OECD's Expert Group on AI or the High-Level Expert Group on Artificial Intelligence (AI HLEG). Policymakers thus have strong incentives to try and engineer trust.

It is important to distinguish between trust and trustworthiness. Both categories are complex and incorporate a variety of new ethical, legal, and social challenges. Trust remains the bedrock of societies, communities, economies and sustainable development. Trust is a complex phenomenon that has sparked many scholarly debates from researchers of diverse disciplines, including psychology, sociology, economics, management, computer science, and IS [3]. In its basic notion, trust is commonly defined as "the willingness of one party to expose themselves to a position of vulnerability towards a second party under conditions of risk and uncertainty as regards the intentions of that second party" [4]. Trust is improbable to be produced on demand [5], and impossible to achieve on command.

There is an inherent relationship between trust, trustworthiness, and the perceived acceptability of risks [6]. Trust in the development, deployment and use of AI systems concerns not only the technology's inherent properties but also the qualities of the socio-technical systems involving AI applications. The concept of trustworthiness performs an important normative function, as it helps in evaluating if people's actual levels of trust are normatively "justified" or "well-placed." This justification depends on whether their degree of trust in something matches its degree of trustworthiness. A person's trust can be "blind" or misplaced; so too can their mistrust. AI may then be merely reliable, but not trustable [7]. The degrees of trustworthiness and actual trust can thus be misaligned in society. This prompts the normative question of whether people's degree of trust is well-placed or justified [7].

In recent times, numerous researchers, industry experts, and policymakers have created and released various frameworks and guidelines that advocate for ethical principles of Trustworthy Artificial Intelligence. In this paper, we will focus on normative documents only which were published by governmental, public and non-profit organizations.

The aim of this study is to critically examine the principles that constitute trustworthy AI and assess their impact on the development of AI systems that are considered trustworthy. To achieve this goal first we will analyze the principles of trustworthy AI (TAI), identify any patterns or

trends, and reveal any fundamental driving force behind the ongoing global conversation about the future of AI. Next, we will use a model developed by our team that uses the Fuzzy Logic Toolbox in the MATLAB 9.4 (R2018a) environment to explore any underlying patterns that may exist. For this, we will apply a scoring system for each principle based on the HLEG Trustworthy AI Checklist [2]. Finally, we aim to observe the impact of each variable (ethical principle) on the trustworthiness of AI, and we will be particularly interested in discerning any unpredictable outcomes.

B. Methodology

In our research, we focused specifically on normative documents issued by esteemed policymakers such as the UN High-Level Committee on Programs, the OECD's Expert Group on AI, the High-Level Expert Group on Artificial Intelligence, and similar authoritative bodies. We prioritized these documents as they reflect widely accepted perspectives and have the potential to shape a regulatory framework in the field of AI. Our research on the influence of principles of trustworthy AI is based on the paper of Fjeld, J. et al. [8] who analyzed the contents of 36 normative documents, defined by the authors as "widely influential, especially visible, and prominent" AI principles documents (further in this article - dataset). The substantial aspect of their findings is the eight key principles for trustworthy AI (called by authors "themes"), namely:

- Privacy principles which is present in 97% of documents in the dataset;
- Accountability principles is present in 97% of documents in the dataset;
- Safety and Security principles are present in 81% of documents in the dataset;
- Transparency and Explainability principles are present in 94% of documents in the dataset;
- Fairness and Non-discrimination principles are present in 100% of documents in the dataset;
- Human Control of Technology principles are present in 69% of documents in the dataset;
- Professional Responsibility principles are present in 78% of documents in the dataset;
- Promotion of Human Values principles are present in 69% of documents [8].

To classify the principles, we separated them into two distinct groups - primary and secondary. This classification is based on the percentage of how frequently a certain principle is mentioned in the selected regulatory documents. The principles that were mentioned in 90% or more of the sample are considered primary. In contrast, the principles that were found in less than 90% of the sample are regarded as secondary. Due to the length constraints, this study will focus solely on factors which we define as primary (Privacy, Accountability, Safety and Security, Transparency and Explainability, and Fairness and Non-discrimination). The scoring of the principles of trustworthy AI is based on the Trustworthy AI Assessment List recommended by the AI High-Level Expert Group (AI HLEG) [2], giving 1 point for each positively answered question and setting the maximum, medium or minimum number of points (see Table I).

To investigate the relationship between the prime factors and trustworthy AI, we proposed a model that employed Fuzzy Logic Toolbox in the MATLAB 9.4 (R2018a) environment. Initially, fuzzy systems were focused on theoretical aspects but the development of computer systems that imitate human reasoning has led to the development of a variety of fuzzy systems, including data mining, financial management, development tools, computational methods, development tools, calculation methods, and fuzzy control systems. The development of a fuzzy model for determining trustworthy artificial intelligence systems can serve as a source of obtaining information about the studied indicator, taking into account a large number of factors of its occurrence.

TABLE I. SCORING THE IMPACT OF PRIMARY PRINCIPLES ON TRUSTWORTHY AI

Table Head	**Number of questions included in the list**	**Score, based on the AI HLEG's Trustworthy AI assessment list [2]**		
		High	***Medium***	***Small***
Privacy	14	14	9	5
Accountability	13	13	8	5
Transparency and Explainability	30	30	19	10
Fairness and Non-discrimination	27	27	16	9
Safety and Security	33	33	19	10
Total	**84**	**84**	**52**	**29**

The impact of primary principles on trustworthy AI was modelled using established fuzzy input rules [9] by treating each question with equal weight. This makes the point allocation consistent and ensures that the evaluation process is fair and impartial (as shown in Fig. 1).

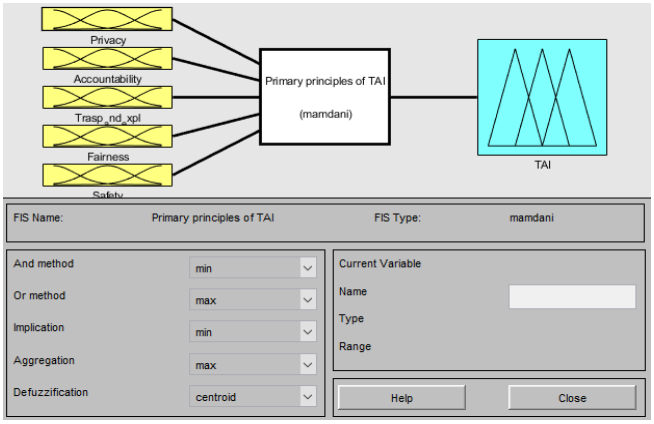

Figure1. General view of the Fuzzy System

We utilized three membership functions based on polynomial curves (pimf) for each variable, which represented three different values: low, medium, and high.

II. EVALUATION OF PRIMARY PRINCIPLES OF TRUSTWORTHY AI

A. Privacy

Privacy principles are present in 97% of documents in the dataset and are defined in many documents as a fundamental right closely related to the principle of prevention of harm. Artificial intelligence systems have a significant impact on privacy due to the fact it is not only implicated in prominent

implementations of AI but also behind the scenes, in the development and training of these systems. The EU General Data Protection Regulation (GDPR) has had a significant impact on establishing safeguards for personal data protection which include proper data governance, maintaining a high quality of data, that covers data quality and integrity, compliance with the deployment domain, secure access to protocols clearly identifying personal who can access data and under which circumstances.

The Trustworthy AI assessment list [2] which we use as a basis for scoring ethical principles, includes three sections for evaluating the principle of Privacy, with a total of 14 questions (Table II). Each positively answered question is given a single point, leading to a possible total score of 14 points. On the basis of the AI HLEG's Trustworthy AI assessment list, we developed recommendations for determining the high, medium and low values for variables in the proposed model. We consider that high values are obtained from 11 to 14 answers, medium from 6 to 10 answers, and low - from 0 to 5 answers.

TABLE II. SETTING BOUNDARIES OF FACTORS FOR FUZZY MODEL

Name of the primary principle	Boundaries [a]		
	High	***Medium***	***Small***
Privacy	11-14	6-10	0-5
Accountability	10-13	5-8	0-4
Transparency and Explainability	20-30	11-19	0-10
Fairness and Non-discrimination	18-27	10-17	0-9
Safety and Security	25-33	11-24	0-10

[a.] based on the AI HLEG's Trustworthy AI assessment list [2] *(Table footnote)*

To convert the number of positive answers from the questionnaire into numerical values, it is suggested to determine their indicator in the total number with a list of possible ones for each element. For Privacy, the maximum number of answers is 14, which compares to 1. Therefore, to obtain the result, the formula of attributing the available answers to the total number of questions is used, and the result is displayed as a decimal fraction. The following variables are offered for the Privacy membership function: s (small) within [0; 0.08; 0.3], m (average) – [0.15; 0.3; 0.7; 0.85], and h (high) – [0.7; 0.9; 1], which characterize a small, medium and large number of responses or characteristics (Fig. 2).

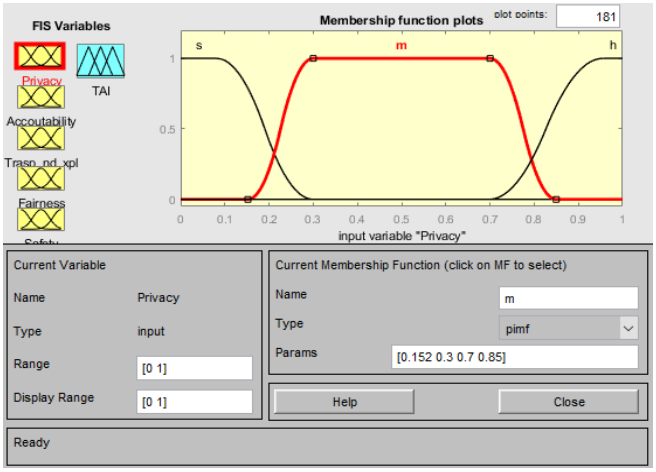

Figure 2. Membership function of Privacy principle

B. *Accountability*

The accountability principle is presented in 97% of documents in the dataset. In general, this principle demands the establishment of mechanisms that can ensure accountability to hold someone legally responsible in case of an AI failure, thus supporting the justice principle.

Almost all documents included in the dataset [8] mention the Accountability principle. The documents reflect diverse perspectives on the mechanisms through which accountability should be achieved. The White House AI Principles [10], on the other hand, refer to transparency and accountability within several of their ten principles but do not explicitly state both as a requirement for trustworthy AI [3].

The Trustworthy assessment list includes four sections for Accountability (Auditability; Minimisation and reporting of negative impacts; Trade-offs and Redress) with 13 questions in total (see Table II). Similar to the previous principle, each question received 1 mark for a total of 13 marks.

Taking into account the main Accountability principles and its main requirements, we consider it expedient to create a fuzzy model to define a high rate of receiving 10-13 positive answers, an average of 5-8 answers, and up to 4 will characterize a low result. Based on the ratio of the received answers to the maximum possible (13 answers), we will get the result of the Accountability input element for the fuzzy system in the form of a decimal fraction with the maximum possible value of 1. The membership function of Accountability is defined within s (small) – [0; 0.04; 0.36], m (medium) – [0.15; 0.35; 0.65; 0.85] and h (high) – [0.65; 0.85; 1], as shown in Fig. 3.

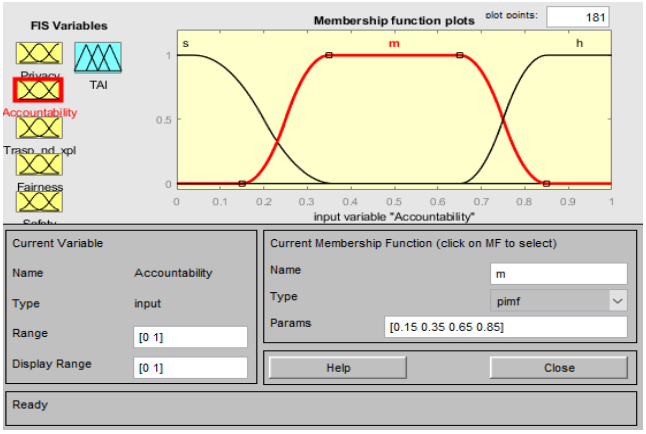

Figure 3. Membership function of the principle of Accountability

C. *Transparency.*

Transparency and Explainability principles are present in 94% of documents in the dataset. These principles are closely linked with the principle of explicability and encompass transparency of elements relevant to an AI system: the data, the system and the business models. While the AI HLEG Guidelines provide for three components of transparency – traceability, connectivity, and explainability – we believe the latter is the most important for developing trust in AI. The main reason for this is that current AI-based systems are complex systems that basically function as black boxes and therefore suffer from opacity and lack of accountability.

De Bruijn and others stress the complex relationship between explainability and transparency on the one side and

trust on the other side [11]. The authors state that explainable AI may not lead to trust and facilitate technology uptake if people "who do not know how AI works" do not have enough confidence in it [11].

This points to additional obstacles to a successful implementation of AI, namely Knowledge Asymmetries (or Knowledge Gap – lack of technical expertise of trustor). In our opinion Explainability can enhance trust if the public has developed trust in the mediating institution which interprets the technical processes and corresponding human decisions, monitors them and is also accountable to the public to protect their interests. However, in conditions where such an institution is missing, explainability would have an adverse rather than positive effect on public trust as requires special knowledge. We agree with Robinson [12] that "only individuals of high technical understanding can appreciate the complexity of systems," potentially eroding citizens' trust.

Communication (or informed consent). When people interact with AI systems, they have the right to know that they are not interacting with a human. This means that AI systems must be clearly identified as such. While one group of academics support the idea of implementing informed consent for fostering trust in AI [13], other criticize it as informed consent suffers from knowledge asymmetries in similar ways as transparency and explainability do [7]. Informed consent does not automatically promote trust in medical care [14]. Pickering argues that trust is "constant negotiation" - a dynamic and ongoing process between the trustor and the trustee [14].

The AI HLEG's Trustworthy assessment list which we used as a basis for scoring the impact of principles on trustworthy AI includes 30 questions about Transparency. Each question was assigned one point so that the sub-principles of traceability, explainability and communication were awarded 7, 10 and 13 points if the requirements were fully met and all questions were answered positively. Given this, we consider it appropriate to develop a fuzzy model and build a membership function to determine the high value when receiving from 20 to 30 positive answers, the average value for 11-19 positive answers and low - up to 10 answers.

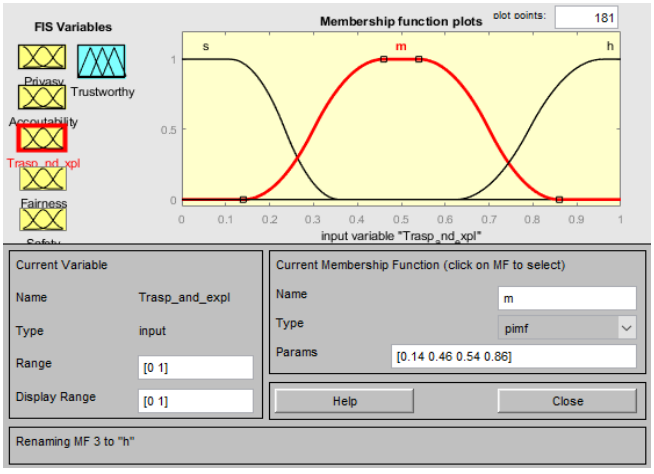

Figure 4. Membership function of Transparency and Explainability

To transform the received number of responses into input data for a fuzzy system, we applied the formula of the ratio of received positive responses to the maximum possible value (30 responses) and display the result as a decimal fraction. Membership function Transparency and explainability defined within s (small) – [0; 0.1; 0.35], m (medium) – [0.14; 0.46; 0.54; 0.86] and h (high) – [0.6; 0.95; 1] (Fig. 4).

D. Fairness and non-discrimination

Fairness and non-discrimination principles are one of the most included ones in our dataset [8] and are present in 100% of documents. That shows their importance and significance in ethical and social terms. The "non-discrimination and the prevention of bias" principle articulates that bias in AI – in the training data, technical design choices, or the technology's deployment – should be mitigated to prevent discriminatory impacts.

The Trustworthy AI assessment list which we use as a basis for scoring principles is comprised of 27 questions, each carrying a score of one point. In the event of an affirmative answer to all questions and full fulfilment of the criteria, the sub-principles receive 16, 9 and 2 points, respectively. In the model, it is proposed to define the Fairness and non-discrimination indicator as high when receiving from 18 to 27 positive responses, medium when receiving 10-17 responses, and low when receiving up to 9 responses. To transform the received number of responses in the block reflecting Fairness and non-discrimination into input data for a fuzzy system, we applied the formula of their ratio to the maximum possible value (27) and display the result as a decimal fraction. Membership function Fairness defined within s (small) – [0; 0.04; 0.35], m (medium) – [0.15; 0.45; 0.55; 0.85] and h (high) – [0.6; 0.9; 1] (Fig. 5).

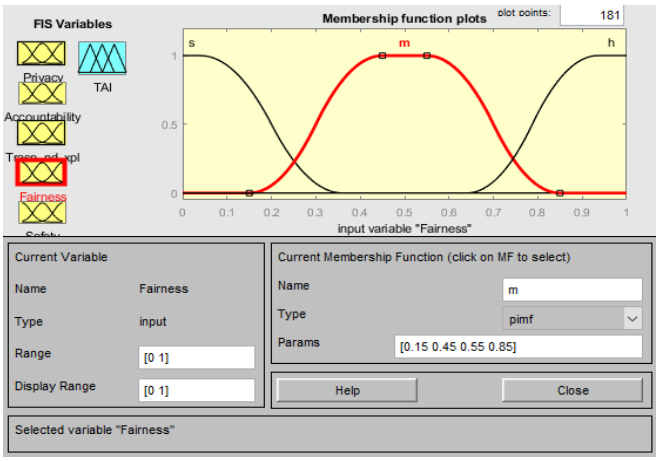

Figure 5. Membership function of the Fairness principle

E. Safety and Security Principles

Safety and security principles, which cover only 81% of documents in the database, were not classified as primary but included in the group of secondary principles. However, some recent regulatory documents which were not included in Fjeld, J. et al. [8] such as White House AI Principles [10], OpenAI [15], DeepMind Ethics & Society Principles [16] emphasize the criticality of these principles.

Fjeld, J. et al. [8] indicate that the principle of safety generally refers to the proper internal functioning of an AI system and the avoidance of unintended harm. Furthermore, they argue that when the term "reliability" is used in the papers in connection with artificial intelligence systems, it means that a reliable system is both secure because it cannot be hacked

by unauthorized third parties and secure because it operates under destination without errors.

The principles of Safety and Security are found in the least number of working documents, but they are reflected by a list of 33 questions. Therefore, we consider it expedient for building a fuzzy model to define the presence of 25 to 33 answers as a high value of this influence factor, 11-24 answers as medium, and less than 10 as low. Therefore, the formula of the ratio of the existing answers to the total number of questions regarding Safety and Security will be used to obtain the result, and the result will be displayed as a decimal fraction. The following variables are proposed for the membership of the function of Safety: s (small) within [0; 0.05; 0.36], m (medium) – [0.12; 0.3; 0.65; 0.9], and h (high) – [0.65; 0.9; 1]. These variables are used to identify a small, medium, or large number of score responses (Fig. 6).

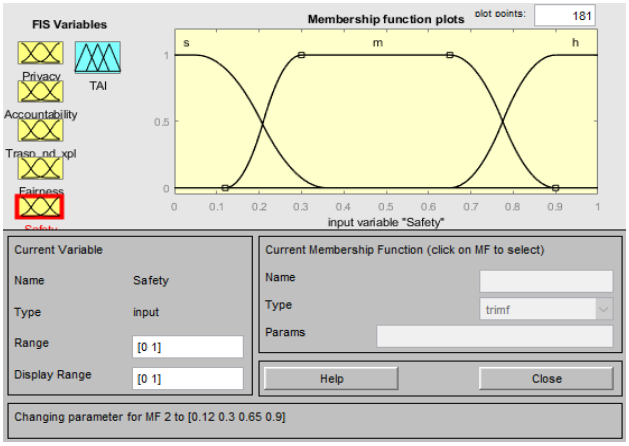

Figure 6. Membership function of Safety and Security principles

III. FINDINGS

The primary objective of evaluating the trustworthiness of artificial intelligence systems is to ensure user protection and foster public trust prior to deployment in the market. The greater the level of trustworthiness, the more positive feedback can be expected from users and the public. Conversely, if the AI system is deemed less trustworthy, people are less likely to adopt it. Software with a low level of trustworthiness can create a reputational risk which can come from a significant negative public response. This can jeopardize not only the current market position of the company but also make it difficult to overcome this negative perception in the future. Therefore, it is not advisable to release such programs on the market. It is difficult to predict the exact public response to AI software with a medium level of trustworthiness, as it can vary depending on several factors, such as the ability of the software to meet user needs, the level of competition in the market, the level of public awareness.

The proposed model provides insight into the process of assigning the original membership function of a trustworthy AI, showing that this is achieved through the use of a triangular function. This finding sheds light on the underlying mechanisms that drive the AI trustworthiness assessment and may have important implications for the design and deployment of such systems in a variety of contexts. The resulting values are classified as follows (Fig. 7):

- ‘s’ (small) category, if values fall into the interval [0, 0.4], indicating that the reliability of a trustworthy AI system is characterized by a low level of trust;
- ‘m’ (medium) category refers to values that are within the range of [0.1; 0.9], suggesting that a trustworthy AI system has an average level of trust.
- h (high) category, if values fall within the range of [0.6, 1], it indicates that the trustworthiness of AI is low.

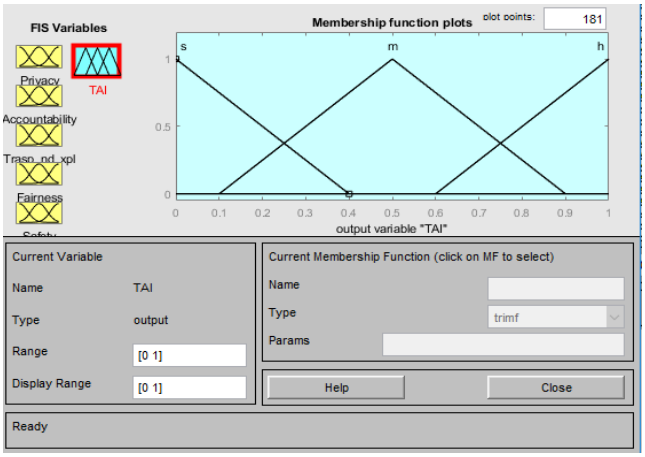

Figure 7. Membership function of TAI

The knowledge database for the construction of such a fuzzy model consists of 971 rules of the type “if – and… and – then” (Fig. 8). Since Fairness and non-discrimination principles are present in all documents, we consider its value to be one of the most significant when creating fuzzy model rules for determining TAI. Therefore, to form the rules of a fuzzy system, we consider it appropriate to consider the average and high values of these principles as a greater influence on the value of trustworthy AI. Safety and Security are represented in only 81% of documents, therefore, when developing the rules of the model, the value of this factor is smaller than others. All the input variables have three fuzzy states, and the result represents three variants of the estimation for the determination degree of TAI.

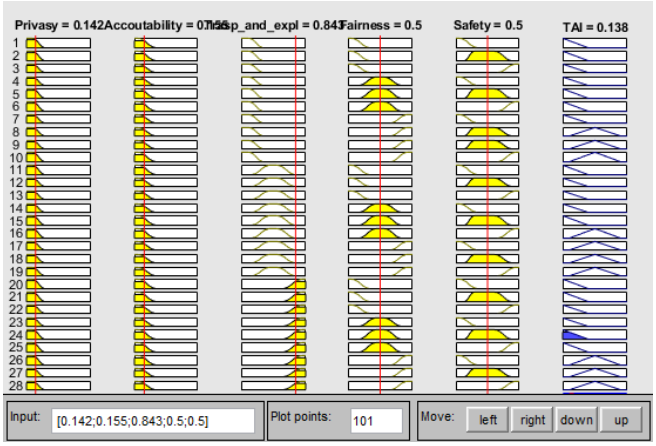

Figure 8. The example of implementation of the proposed model of Fuzzy Logic Toolbox for the assessment of the impact of principal factors on trustworthy AI

Examples of implementation. The majority of factor combinations tend to produce predictable and logical outcomes, for example:

- Rule 5, the levels of privacy, accountability, and transparency are low, while fairness and safety are at a medium level. As a result, we can expect AI to have a predictably low level of trustworthiness, that is predictable.

- According to Rule 26, the levels of privacy, accountability, and safety are low, while transparency and fairness are high. Therefore, we might reasonably expect the level of trustworthiness in AI to be at a medium level.

However, the proposed model of the Fuzzy Logic Toolbox for the assessment of the impact of principal factors on trustworthy AI also recognizes unpredictable results. A vivid example of it is rule 24 (see Fig. 8) according to which given the low values for privacy and accountability, as well as the medium values for the variables of fairness and safety, it is reasonable to expect that a high level of transparency would result in a medium level of trustworthiness. However, this assumption is incorrect. Instead, when these factors are considered, the level of trustworthiness in AI is actually low.

IV. The Area of Implementation of the Proposed Model

As there are no mandatory regulations established up to date, the proposed model refers to two benefits for developers and deployers of AI systems:

- *Assess the potential public trust in AI systems.* Using a triangular function, developers and deploying organizations can more accurately predict how AI systems will perceive the public and take steps to solve any problems that are concerned about the release of the system to the market.
- *Predicting the degree of compliance* with the regulatory framework, as it includes principles which are assigned in 90% or more of regulatory frameworks worldwide. This can be particularly valuable for developers and deployers, which potentially will fall under the EU AI Act and seek to ensure that their AI systems meet regulatory requirements and meet legal and ethical standards. The model can detect any fields of failure and identify necessary corrections before the system deploys.

Trustworthy AI is not about ticking boxes, but about continuously identifying requirements, evaluating solutions and ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders therein [2]. To ensure the continued trustworthiness of AI systems, we propose that our model be utilized at all stages of the software life circle: during development; before placing to market; and at the stage of use to monitor compliance with the safeguards declared to users.

Developers of AI systems are required to conduct an initial compliance assessment before placing high-risk AI systems on the market, while post-market monitoring systems must be established to document and analyze the performance of high-risk AI systems throughout their lifetime. It is also worth noting that compliance tends to deteriorate over time, which emphasizes the need for ongoing monitoring and maintenance. Implementing our proposed model at these key milestones will enable stakeholders to take proactive steps to maintain trust and mitigate risk.

V. Conclusions

While the proposed model refers to developers and deployers of AI to self-review, the effectiveness of ethics-based auditing is directly related the governmental and institutional support and regulation. The developers and deployers of AI systems are primarily motivated by commercial interests and profitability and may not prioritize ethical considerations. Therefore, it is imperative that government and regulatory agencies take responsibility for protecting their citizens in this regard.

This field cannot and should not be self-regulating due to the knowledge asymmetry that exists between developers and deployers of AI systems on the one hand, and users and the public on the other. The public may not possess the specialized knowledge and expertise to assess the potential harm that AI systems might cause. As such, government and regulatory agencies should use their powers to ensure that AI systems are designed and deployed in a manner that prioritizes ethical considerations and minimizes potential harm to users.

References

[1] R. Gianni, S. Lehtinen and M. Nieminen "Governance of Responsible AI: From Ethical Guidelines to Cooperative Policies". Frontiers in Computer Science. May 2022, Volume 4. Article 873437. doi:10.3389/fcomp.2022.873437

[2] Ethics Guidelines for Trustworthy AI. High-level expert group on artificial intelligence set up by the European Commission. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[3] S. Thiebes, S. Lins and Ali Sunyaev "Trustworthy artificial intelligence" Electron Markets: 31, 447–464 (2021). https://doi.org/10.1007/s12525-020-00441-4

[4] F. Bannister and R. Connolly "Trust and transformational government: A proposed framework for research". Government Information Quarterly, 28, 2011, pp. 137–147,

[5] K. S. Cook and J. J. Santana "The Routledge Handbook of Trust and Philosophy". Routledge Press, 2020, pp. 189–204.

[6] W. Poortinga and N. F. Pidgeon "Trust in risk regulation: Cause or consequence of the acceptability of GM food?" Risk Analysis, 25(1), 2005, pp. 199–209

[7] Johann Laux, Sandra Wachter and Brent Mittelstadt "Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk" Regulation & Governance, 2023. doi:10.1111/rego.12512

[8] J. Fjeld, A. Nele, H. Hilligoss, A. Nagy, and M. Srikumar. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center for Internet & Society, 2020. Available at the link: http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420

[9] S. Shtovba "Introduction to the Theory of Fuzzy Sets and Fuzzy Logic" [Online]. Available at: http://matlab.exponenta.ru/fuzzylogic/book1/

[10] Making Automated Systems Work for The American People. Blueprint for an AI Bill of Rights. Available at: https://www.whitehouse.gov/ostp/ai-bill-of-rights/#human

[11] H. de Bruijn, M. Warnier and M. Janssen. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Government Information Quarterly, 2022, 39, 101666.

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

[13] C. Wilson and M. van der Velden "Sustainable AI: An integrated model to guide public sector decision-making". Technology in Society, 68 (2022). https://doi.org/10.1016/j.techsoc.2022.101926

[14] B. Pickering "Trust, but verify: Informed consent, AI technologies, and public health emergencies". Future Internet, 13 (5). 2021, https://doi.org/10.3390/fi13050132

[15] Open AI: Product safety standards. Available at: https://openai.com/safety-standards

[16] DeepMind Ethics & Society Principles. Available at: Safety & Ethics (deepmind.com)