DOI: https://doi.org/10.15276/hait.09.2026.45
UDC 004.8:316.4:005.52

A fuzzy multicriteria model for assessing social risk based on public attitudes toward artificially generated content

Volodymyr V. Polishchuk[1,2)]
ORCID: https://orcid.org/0000-0003-4586-1333; volodymyr.polishchuk@tuke.sk. Scopus Author ID: 57204564304
Iwona Włoch[3)]
ORCID: https://orcid.org/0000-0002-9969-0827; iwloch@prz.edu.pl. Scopus Author ID: 6506235489
Andrii A. Matei[2)]
ORCID: https://orcid.org/0009-0001-0280-1763; andrii.matei@uzhnu.edu.ua.
Yurii Yu. Mlavets[2)]
ORCID: https://orcid.org/0000-0002-1480-9017; yurii.mlavets@uzhnu.edu.ua. Scopus Author ID: 55123275500
[1)] Technical University of Košice, 1/9 Letná, 042 00, Slovakia
[2)] Uzhhorod National University, 3, Narodna Square, 88 000, Ukraine
[3)] Rzeszow University of Technology, 12, al. Powstańców Warszawy, 35 029, Poland

ABSTRACT

Relevance. Artificially generated content can undermine information trust, increase vulnerability to manipulation, and create group-level social risks that are difficult to represent by crisp indicators **Purpose.** To develop and empirically test a fuzzy multicriteria model that transforms heterogeneous survey responses into a group-level social-risk assessment **Tasks.** To formalize social indicators and criteria, risk-orient linguistic responses, compare socio-demographic groups, and define decision rules for interpreting the resulting index. **Methods.** Anonymous survey data from three hundred eighty-one respondents in Ukraine were processed using ordinal recoding, Z-shaped membership functions, equal criterion weights, TOPSIS ranking, and dimension-normalized conical aggregation **Results.** The group indices were 0.693 for respondents aged eighteen to thirty-three, 0 for respondents with a bachelor's degree, and 0.307 for the 6,700–10,000 UAH income group. After dimension-normalized aggregation, the selected scenario produced 0.276, corresponding to high social risk **Scientific novelty.** Social risk is formalized as a formative fuzzy multicriteria construct integrating trust, concern, verification behavior, regulatory expectations, and socio-demographic grouping **Practical significance.** The model can support public information-security bodies and media-literacy program managers in prioritizing target groups, selecting interventions, and planning repeated monitoring; it is not an individual diagnostic instrument **Conclusions.** The model provides interpretable group-level screening, but its outputs require validation using the sample composition, subgroup intersection sizes, and sensitivity analysis.

Keywords: artificially created content; social risk; fuzzy model; decision support; information security; multicriteria assessment

***For citation*: Polishchuk V. V., Włoch Iwona, Matei A. A., Mlavets Yu. Yu. "A fuzzy multicriteria model for social risk assessment in the context of the spread of artificially generated content". *Herald of Advanced Information Technolog*. 2026; Vol.9 No.4: 736–751. DOI: https://doi.org/10.15276/hait.09.2026.45**

INTRODUCTION

Public responses to artificially generated content depend on both content characteristics and users' cognitive and socio-demographic profiles. Groh et al. [1] showed that people's ability to recognize political deepfake messages depends on the format of information presentation, with audio and video signals significantly affecting recognition accuracy.

Li and Yang [2] found that labeling AI-generated content helps users distinguish artificially generated content from human-generated content buthas a limited impact on perceptions of credibility and intentions to disseminate information.

In an Organisation for Economic Co-operation and Development (OECD) report, Lorenz et al. [3] emphasize that generative AI has significant transformative potential but also creates social, political, and economic risks, particularly related to disinformation and manipulated content. Lovato et al. [4] note that the recognition of deepfake content depends not only on the technical characteristics of the message but also on human cognitive biases. UNESCO [5] emphasizes the need for transparent, ethical, and responsible use of generative AI, especially in education, research, and the information environment.

Thus, the risks associated with the spread of artificially created content mainly focus on the detection of deepfake content, the labeling effect, and general political and ethical recommendations. A formalized approach that would combine respondents' social characteristics, linguistic assessments of their attitudes, and multi-criteria determination of the integral level of social risk remains underdeveloped.

© Polishchuk V. V., Włoch Iwona, Matei A. A., Mlavets Yu. Yu., 2026

This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/deed.uk)

In this study, social risk is defined as the potential for artificially generated content to undermine information trust, increase anxiety and susceptibility to manipulation, and induce unsafe information behaviour at the group level. The aim of the study is to develop and empirically test a fuzzy multicriteria model that converts survey responses into a comparable index of safe public perception and a corresponding social-risk level for socio-demographic groups. To achieve this aim, the study (i) formalizes social indicators and response criteria, (ii) maps heterogeneous linguistic responses onto a common [0,1] risk-oriented scale, (iii) ranks social groups using TOPSIS, and (iv) interprets group profiles for targeted information-security and media-literacy decisions. The research hypothesis is that socio-demographic groups differ in the integral safe-perception index and that lower index values indicate higher group-level social risk.

The task is to obtain a reproducible within-indicator coefficient of safe public perception for every group and then aggregate one selected group from each indicator into a scenario-level social-risk index. The inputs are respondent-level categorical answers and mutually exclusive categories within the same social indicator; the outputs are TOPSIS relative-closeness coefficients and the dimension-normalized scenario index. Groups with different social indicators are not compared as alternatives in a single TOPSIS matrix. The empirical example considers a fixed scenario composed of three preselected groups, one from each social indicator. The TOPSIS coefficients are relative to this fixed illustrative comparison set and are not interpreted as absolute or individual risk scores.

RELATED WORKS

Recent scientific research confirms that the spread of artificially created content creates a new type of social risk, related not only to the technical ability to generate texts, images, audio, or video, but also to the peculiarities of public perception of such content. Hwang et al. [6] experimentally showed that deepfake disinformation hurts the perception of information, and media literacy can reduce this impact, but does not eliminate the problem. Groh et al. [7] studied the ability of humans and machines to recognize deepfake content. They showed that human assessment remains limited, especially in conditions of high realism of artificially created materials. Groh et al. [1] later proved that the accuracy of recognizing political deepfake messages depends on the format of information presentation: text, audio, or video.

Diel et al. [8], based on a systematic review and meta-analysis of 56 papers, found that people's overall ability to recognize deepfake content is often close to random guessing, and accuracy varies significantly depending on the type of stimulus. This confirms that the social risk of artificially generated content cannot be assessed solely based on the technical presence of the fake; it is necessary to take into account people's reactions, trust, anxiety, and their ability to assess information critically. Lovato et al. [4] further showed that the recognition of deepfake content depends on human cognitive biases, i.e., users' reactions to artificially created content have socio-psychological heterogeneity.

A separate group of studies is devoted to labeling artificially generated content. Li and Yang [2] found that AI-generated content labels help users distinguish artificially generated content from human-generated content, but have only a limited impact on perceived credibility, message trust, and intention to share. Wittenberg et al. [9] also showed that labeling AI-generated media reduces belief in the presented claims, but simply informing users that the content is AI-generated does not always significantly change users' behavioral intentions. Williams-Ceci et al. [10] found that warning banners can have ambiguous effects and even elicit undesirable user reactions. Therefore, informing only about the artificial origin of content is not a sufficient mechanism for reducing social risk.

The socio-political dimension of the problem is explored in the works of Verma [11], Jacobsen and Simpson [12], and Hynek et al. [13]. Verma [11], based on qualitative interviews, showed that citizens perceive deepfake technologies as potentially dangerous for political communication, trust, and social stability. Jacobsen and Simpson [12] emphasized that deepfake content creates tensions between creative use of technology, information trust, security, and social responsibility. Hynek et al. [13] combined a systematic review with survey data in seven European countries and showed that attitudes towards deepfake technologies are heterogeneous, and the risks and potential benefits are perceived differently depending on the social context.

In studies focused on generative AI and disinformation, Park and Nan [14] summarized the role of generative AI in the creation, detection, mitigation, and impact of disinformation. Ma et al. [15] noted that generative AI changes the traditional model of fake news creation, enabling the rapid scaling of false or manipulative messages. Kumar et al. [16] considered the potential of large language

models for detecting disinformation, but their approach focused mainly on the technical aspect of detection. Moyo et al. [17] proposed a conceptual AI-oriented framework for detecting fake news and deepfake content. Still, the emphasis was also mainly on technical detection and explainability, rather than on quantifying public attitudes towards such content.

Recent studies also highlight that the problem of artificially generated content has not only a technical but also a behavioral dimension. Shin et al. [18] showed that the perception of disinformation generated by generative AI depends on the way the user cognitively processes the information, and not only on the presence or absence of signs of fakery. Bashardoust et al. [19] compared the willingness of users to spread fake news generated by humans and AI based on an experiment and found that even when AI-generated fake news is perceived as less credible, users can spread it with similar intensity. Ching et al. [20] in a critical scoping review, noted that deepfake content can influence beliefs, memory, and assessment of the reliability of information, but the strength of this influence depends on the context, type of message, and characteristics of the audience. This confirms the feasibility of assessing not only the fact of contact with AI-generated content, but also the level of trust, readiness for dissemination, critical perception, and social vulnerability of users. An additional important direction is the measurement of trust and risks associated with deepfake and AI-generated content. Plohl et al. [21] developed and validated the Perceived Deepfake Trustworthiness Questionnaire, which confirms the need for specialized tools for assessing subjective trust in artificially created media. Xu et al. [22], using topic modeling and sentiment analysis of social media, showed that public perception of deepfake technologies is formed through a combination of concern, distrust, interest, and expectations for regulation. Romero Moreno [23] examined deepfake content in the context of human rights and emphasized the need for a balanced approach that takes into account not only the fight against harmful content, but also the rights to privacy, freedom of expression, and the responsibility of digital platforms. Thus, the current literature confirms that the social risk of artificially created content should be assessed as a multidimensional phenomenon that combines technical realism, cognitive perception, trust, behavioral intentions, and normative and ethical context.

Thus, an analysis of current research shows that some aspects of the problem have already been studied quite thoroughly: technical recognition of deepfake content, the impact of labeling AI-generated content, the role of media literacy, users' cognitive biases, political risks, and the general social consequences of generative AI. At the same time, existing approaches are mainly focused either on technical detection, or on experimental study of individual perception factors, or on a general description of risks.

An approach that would simultaneously take into account the social characteristics of respondents, the linguistic nature of their responses, the uncertainty of public assessments, the multi-criteria nature of risk factors, and the need to obtain an integral index of public attitudes remains insufficiently developed. That is why the proposed fuzzy multi-criteria model is relevant and necessary, as it allows us to move from disparate citizen responses to a formalized quantitative-linguistic result. Such a result can be used to compare social groups, identify groups with higher levels of distrust or concern, and develop scenarios to support decision-making in the field of society's information security.

Related journal studies have addressed machine-learning analysis of public sentiment [27], emotion recognition in text [28], and multilingual political-leaning classification [29], confirming the relevance of automated analysis of public information responses.

Accordingly, this study treats social risk as a multidimensional formative construct rather than as a single psychological state. Distrust, concern, verification behaviour, and regulatory expectations are complementary criteria through which exposure to artificially generated content may threaten information trust and social stability; they are not separate outcome categories. Existing studies usually examine these components individually, whereas the proposed model integrates them into a transparent group-level decision-support index.

RESEARCH METHODOLOGY

Let the assessment of the level of public attitude towards artificially created content be carried out based on data obtained because of the analysis of the opinions of citizens of a certain region, where the set of respondents is given as $C = \{c_1; c_2; \dots; c_n\}$. For each citizen, the social indicators SC_l, are considered, reflecting the respondents' characteristics. Each of these indicators, in turn, can be detailed through separate groups

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

of sub-indicators: $SC_1 = \{sc_{11};\ sc_{12}; \dots; sc_{1r_1}\}; \dots;$ $SC_l = \{sc_{l1};\ sc_{l2}; \dots; sc_{lr_l}\}$. addition, a set of criteria $K = \{k_1; k_2; \dots; k_s\}$, is formed, according to which the assessment of public attitudes towards artificially created content is carried out. A set of social indicators, criteria for assessing public attitudes, and a data processing system obtained based on citizens' opinions form a developed information method for assessing social risk in the context of the spread of artificially created content, which we will denote as K_{SR}.

It is advisable to present a fuzzy multi-criteria model for assessing social risk in the context of the spread of artificially generated content, in the form of an operator that formalizes the combination of input social indicators, evaluation criteria, and fuzzy aggregation procedures:

$$M_{SR}: (C, SC, K, K_{SR}) \to Y(m_{SR}, D). \quad (1)$$

M_{SR} is a fuzzy multi-criteria model operator that transforms input data C, SC, K and K_{SR} into a set of output values Y. The results of the operator are: m_{SR} is an integral index of safe public perception of artificially created content; D is a generalized level of social risk that is formed as a result of the distribution and safe perception of artificially created content in society.

The conducted research is based on a combination of data obtained from respondents and expert information. In this regard, within the proposed approach, it is advisable to distinguish the following management entities:

- respondents – citizens whose opinions are used to identify expectations, levels of trust, concerns, and general attitudes towards artificial intelligence and artificially generated content;
- systems analyst – a specialist who provides setup for procedures for collecting, structuring, processing, and interpreting data within a multi-criteria fuzzy evaluation model;
- decision maker (DM) – an entity or group of entities that, based on the research results obtained, form management conclusions, response scenarios, and recommendations to support decision-making in the context of information security of society.

In this study, the decision maker may be a public authority responsible for information security, a media-literacy programme administrator, a regulator, or a public-communication unit. The output supports group-level screening and prioritization: it indicates which population segments require deeper diagnosis and which measures-targeted media-literacy training, fact-checking prompts, clear labelling of artificially generated content, or explanatory campaigns-should be considered. The index is not intended to diagnose individuals or justify restrictive measures without additional evidence.

Next, an information method for assessing social risk in the context of the distribution of artificially created content is presented – K_{SR}.

Let each citizen be assigned a set of social indicators, divided into separate classification groups. For a formalized representation of the information method, an open set of social indicators is defined.

SC_1 – respondent's age.

It is advisable to classify this social indicator by the following age groups: $sc_{11} = [18; 34)$; $sc_{12} = [35; 44)$; $sc_{13} = [45; 54)$; $sc_{14} = [55; 65)$; $sc_{15} = [66; 99]$.

SC_2 – highest level of completed education.

According to this social indicator, citizens can be classified according to the following educational groups: sc_{21} ={ complete general secondary education}; sc_{22} ={ vocational and technical education}; sc_{23} ={ higher education bachelor's degree}; sc_{24} ={ higher education master's degree}.

SC_3 – average net monthly income.

The classification groups of the specified indicator are formed according to the level of monetary income of respondents, expressed in hryvnias, and have the following form: sc_{31} = {up to 6700 UAH}; sc_{32} = {6700 – 10000UAH}; sc_{33} = {10000 – 14000 UAH}; sc_{34} = {14000 – 18000 UAH}; sc_{35} = {18000 – 22000 UAH}; sc_{36} = {22000 – 26000 UAH}; sc_{37} = {26000 – 30000 UAH}; sc_{38} = {30000 – 34000 UAH}; sc_{39} = { more than 34000 UAH}.

For citizens from the set C scores are formed on the set of criteria F, which reflect their attitude towards artificially created content. Each criterion is interpreted as a separate question of the questionnaire, to which the respondent chooses one of the proposed answers. At the same time, a different number of possible answer options can be provided for different criteria.

To formally represent the method of responding, it is advisable to represent respondents through terms of a linguistic variable that form a term set: $T = \{t_1;\ t_2;\ \dots\ t_k\}$. The elements of this term set reflect the gradation of attitudes or levels of trust toward artificial intelligence and artificially created content. For example, $t_1$can correspond to the value: t_1 is “absolute trust in AI”; t_2 is «high trust in AI»; …; t_k is “absolute distrust in AI”.

The following open set of criteria is proposed:

k_1 – Can the mass distribution of content created by artificial intelligence influence the formation of public opinion? The logical conclusion for the answers to the questions is as follows:

IF "Will not have a significant impact on public opinion" THEN$\tau(f_1) = t_{11}$;

IF "May be partially affected, but depends on control" THEN $\tau(f_1) = t_{12}$;

IF "Significantly influences and can shape manipulative narratives" THEN $\tau(f_1) = t_{13}$.

k_2 – What is your opinion on using artificial intelligence to create political or social news?

IF "It is acceptable if the content is reviewed by experts" THEN $\tau(f_2) = t_{21}$;

IF "This requires strict state control" THEN $\tau(f_2) = t_{22}$;

IF "This is dangerous and should be banned" THEN $\tau(f_2) = t_{23}$.

k_3 – If you found out that most of the news you read was created by artificial intelligence, how would you react?

IF "I hadn't changed my attitude" THEN$\tau(f_3) = t_{31}$;

IF "I would check the information more carefully" THEN $\tau(f_3) = t_{32}$;

IF "I would lose trust in the source of information" THEN$\tau(f_3) = t_{33}$.

k_4 – How do you assess the potential risk of using AI to create fake videos or "deepfake" content?

IF "This is a minor risk, technology quickly detects fakes" THEN$\tau(f_4) = t_{41}$;

IF "This may pose some risks, but control is possible" THEN $\tau(f_4) = t_{42}$;

IF "This is a serious threat to national security" THEN $\tau(f_4) = t_{43}$.

k_5 – How much do you trust government agencies to counter disinformation created by AI?

IF "Fully trust" THEN $\tau(f_5) = t_{51}$;

IF "Partially trust, but have doubts" THEN $\tau(f_5) = t_{52}$;

IF "I don't trust at all" THEN $\tau(f_5) = t_{53}$.

k_6 – How do you think society should respond to the spread of artificially created content?

IF "No need to intervene, the market will regulate itself" THEN $\tau(f_6) = t_{61}$;

IF "Strengthen government regulation and sanctions" THEN $\tau(f_6) = t_{62}$;

IF "Implement digital education and media literacy" THEN $\tau(f_6) = t_{63}$.

k_7 – If an AI-generated news story containing provocative content appears in your information environment, what will you do?:

IF "I probably won't react, this is normal" THEN$\tau(f_7) = t_{71}$;

IF "I will share the news with a warning to others" THEN $\tau(f_7) = t_{72}$;

IF "I will verify the authenticity through other sources" THEN$\tau(f_7) = t_{73}$.

k_8 – If you had to choose between free newsletters created by artificial intelligence or written by professional journalists, which would you prefer?:

IF "Definitely news created by artificial intelligence or algorithms" THEN $\tau(f_8) = t_{81}$;

IF "Rather news created by artificial intelligence or algorithms" THEN $\tau(f_8) = t_{82}$;

IF "Rather news created by human journalists" THEN $\tau(f_8) = t_{83}$;

IF "Definitely news created by human journalists" THEN$\tau(f_8) = t_{84}$.

k_9 – Should the reader be informed if a news story is created by artificial intelligence, that is, a computer algorithm without the participation of a human journalist?:

IF "Definitely not" THEN$\tau(f_8) = t_{91}$;

IF "Probably not" THEN $\tau(f_8) = t_{92}$;

IF "More like this" THEN $\tau(f_8) = t_{93}$;

IF "Definitely yes" THEN $\tau(f_8) = t_{94}$.

Thus, for the input data received from citizens $C = \{c_1; c_2; \dots; c_n\}$, it is advisable to structure them according to the social indicators SC_l and their classification groups. Each social indicator can be presented as a set of separate groups: $SC_1 = \{sc_{11}; sc_{12}; \dots; sc_{1r_1}\}; \dots; SC_l = \{sc_{l1}; sc_{l2}; \dots; sc_{lr_l}\}$. For each social indicator, an ordered set of input data is formed, in which citizens are distributed across the corresponding social groups, and their answers, according to the criteria K are presented as linguistic variables. For example, for the social indicator SC_1 citizens can be divided between groups $sc_{11}, sc_{12}, \dots, sc_{1r_1}$, and for each citizen c_i the answer $\tau(k_{ui})$ according to the corresponding criterion k_u recorded.

In this case $\tau(k_{ui})$ denotes the response of the i-th citizen according to the criterion k_u, where $i = \overline{1,n}$ and $u = \overline{1,s}$, and k_u is one of the terms in the term set of linguistic variables T. Thus, for each social group, a set of linguistic assessments is accumulated according to all criteria, which allows for their further fuzzy processing, aggregation, and comparative analysis of public attitudes towards artificially created content in different social categories.

The following are calculations according to a multi-criteria model for assessing social risk in the context of the spread of artificially created content.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

The study considers a set of citizens C, who are survey respondents and provided answers in accordance with the procedure defined by the proposed information method.

At the first stage, for each criterion of public attitude to artificially created content, the resulting term score T_u^* is determined within the framework of the individual groups of social indicators sc_{pr_p}, belonging to the sets $SC_1, SC_2, \dots, SC_l$. Such a score reflects the generalized position of citizens of a certain social group according to a specific criterion k_u.

Formally, this can be represented as follows:

$$T_u^*\left(sc_{pr_p}\right) = L\left(\tau(k_{u1}), \tau(k_{u2}), \dots, \tau(k_{u\bar{m}})\right), \qquad (2)$$

where $T_u^*(sc_{pr_p})$ is the resulting term score for the criterion f_u within the social group sc_{pr_p}; $\tau(k_{ui})$ is the individual linguistic score of the i-th citizen according to the criterion k_u; $\bar{m}$ is the number of citizens in the corresponding social group; L – the operator of logical generalization of individual responses into the resulting term $u = \overline{1,s}$, $p = \overline{1,l}$.

Since different criteria may yield different numbers of possible answers, it is advisable to determine the resulting term score based on the percentage of answers within each group. To do this, a rule base of the "If – Then – Else" type is formed, which establishes a correspondence between the proportion of certain linguistic scores and the final term $T_u^*(sc_{pr_p})$.

For example, if, for a certain social group, according to the criterion k_u at least 60% of the responses belong to the term t_1, then the resulting score takes the value $T_u^* = t_1^*$. If at least 60% of the responses belong to the term t_2, or the total share of responses $t_1 + t_2$ is at least 80%, then the resulting score can be defined as $T_u^* = t_2^*$. Similarly, rules are formed for the intermediate terms $t_3, \dots, t_{k-1}$. For the limiting term t_k, which may correspond, for example, to "absolute distrust of AI", the resulting score is determined by the rule of the predominance of responses with this term or by the condition of the dominance of negative assessments in the corresponding social group.

For reproducibility, the rules are applied in the ordinal order of the response terms. The first term t_h satisfying either $q_h \geq 0.60$ or the cumulative share $Q_h \geq 0.80$ is selected. Because $Q_b = 1$, the procedure always returns one term. If two terms have equal support, the higher-risk term is selected as a conservative tie-breaking rule.

Thus, the proposed approach allows moving from individual responses of citizens to a generalized term assessment for each social group and each criterion. All rules of logical generalization can be further formalized in the form of a knowledge base, which provides transparent fuzzy processing of responses and preparation of data for the next stage of social risk assessment.

The questionnaire is used here as a formative multicriteria instrument: its items capture complementary facets of social risk and are not assumed to be interchangeable manifestations of one latent psychological scale. Therefore, a single Cronbach's alpha is not used as a criterion of validity, because low inter-item correlation may reflect the intended multidimensionality rather than measurement error. The robustness of the composite should instead be examined through explicit risk-oriented coding, criterion-level diagnostics, sensitivity to weights and membership functions, and, where theoretically justified, reliability analysis of homogeneous subscales.

After forming the resulting term estimates T^* in terms of groups of social indicators $SC_1, SC_2, \dots, SC_l$, the data of the second stage are presented in the form of matrices, the elements of which are the corresponding generalized term estimates:

$$SR\left(SC_p\right) = \begin{pmatrix} T_{1p1}^* & T_{1p2}^* & \dots & T_{1pr_p}^* \\ T_{2p1}^* & T_{2p2}^* & \dots & T_{2pr_p}^* \\ \dots & \dots & \dots & \dots \\ T_{sp1}^* & T_{sp2}^* & \dots & T_{spr_p}^* \end{pmatrix}. \qquad (3)$$

The obtained matrices $SR(SC_p)$, $p = \overline{1,l}$, have the dimension $s \times r_p$, where s is the number of criteria for assessing the public attitude to artificially created content, and r_p is the number of groups of the social indicator SC_p. Their elements are the resulting term estimates $T_{upr_p}^*$, obtained based on the rules of logical inference.

Next, a multi-criteria evaluation of groups of social indicators is performed to construct the corresponding ranking series. This allows us to determine the level of public attitude towards artificially created content within individual social groups.

For further aggregation, linguistic assessments need to be transformed into normalized quantitative values. For this purpose, fuzzy logic membership functions are proposed. Within the study's framework, it is advisable to use Z-shaped functions, as they reflect a gradual decrease in safe public

perception with increasing social risk. For example, a quadratic Z-spline can be used:

$$\mu_{upr_p} = \begin{cases} 1, & f \le 1; \\ 1 - 2\left(\frac{f-1}{b-1}\right)^2, & 1 < f \le \frac{1+b}{2}; \\ 2\left(\frac{b-f}{b-1}\right)^2, & \frac{b+1}{2} < f < b, \\ 0, & f \ge b. \end{cases} \quad (4)$$

where f is the ordinal number of the linguistic response in the term set; b is the number of possible responses according to the corresponding criterion; μ_{upr_p} is the normalized value of the membership function for the criterion f_u within the group of the social indicator sc_{pr_p}, $u = \overline{1,s}$; $p = \overline{1,l}$.

The study's logic drives the use of the Z-shaped membership function, as it reflects a gradual decrease in safe public perception with increasing distrust, anxiety, or social risk.

The number of response options varies because the questionnaire uses category sets appropriate to the content of each item. Before fuzzification, every item is recoded in a common risk direction: the least-risk response receives ordinal position q=1 and the highest-risk response receive $q = b_u$, with intermediate responses retaining their order. The Z-shaped function is then parameterized by the item-specific number of categories b_u. Consequently, a four-option item is represented by four membership positions and is not reduced to a three-level scale.

Criteria k_6 and k_7 describe the reported intensity and type of social or behavioural response rather than an independently validated measure of protective effectiveness. Their coding in the present example should therefore not be interpreted as normative evidence that non-intervention or non-reaction is safer than media literacy or source verification. Alternative risk-oriented coding of these criteria will be examined in further validation.

After converting linguistic assessments into normalized quantitative values, a decision matrix is formed for each social indicator:

$$\overline{SR(SC_p)} = \begin{pmatrix} \mu_{1p1} & \mu_{1p2} & \dots & \mu_{1pr_p} \\ \mu_{2p1} & \mu_{2p2} & \dots & \mu_{2pr_p} \\ \dots & \dots & \dots & \dots \\ \mu_{sp1} & \mu_{sp2} & \dots & \mu_{spr_p} \end{pmatrix}. \quad (5)$$

In this matrix, the rows correspond to the criteria for assessing public attitudes towards artificially created content, and the columns correspond to the groups of the corresponding social indicator SC_p. The elements μ_{upr_p} are the normalized values of the membership functions that reflect the level of safe public perception or, conversely, potential social risk.

Further multi-criteria evaluation of alternatives presented in matrices $\overline{SR(SC_p)}$, can be carried out using multi-criteria selection methods. In particular, for ranking groups of social indicators, the method of hierarchy analysis, TOPSIS, VIKOR, Best-Worst Method, matrix multiplication method, and other approaches suitable for comparing alternatives by a set of criteria can be used.

For the multi-criteria evaluation of groups of social indicators SC_p the TOPSIS method is proposed. The choice of the TOPSIS method is justified by its ability to compare alternatives based on their relative proximity to the positive and negative ideal solutions, which is appropriate for ranking social groups by the level of public attitude towards artificially created content. Similar logic has already been used in information risk studies: Kaul and Guaba [24] used Fuzzy TOPSIS to rank social media platforms according to the probability of spreading infodemics and disinformation, and Gaeta et al. [25] combined TOPSIS with Fuzzy Rough Sets for the explanatory analysis of information disorder phenomena.

The initial input for the TOPSIS method is the normalized decision matrix $\overline{SR(SC_p)}$, the elements of which are the values of μ_{upr_p}, where $u = \overline{1,s}$; $p = \overline{1,l}$. In this matrix, the rows correspond to the criteria k_u, and the columns correspond to the groups of the social indicator SC_p. Therefore, each column of the matrix characterizes a separate alternative, i.e., the corresponding group of the social indicator sc_{pr_p}.

At the first stage, a vector of criteria weights is introduced:

$$W = \{w_1, w_2, \dots, w_s\}, \sum_{u=1}^{s} w_u = 1, \quad (6)$$

where w_u is the weight coefficient of the criterion k_u. in the absence of a priori expert preferences, the weight coefficients of the criteria are assumed to be equivalent, which reduces the subjectivity of the initial model setup and ensures its reproducibility.

Next, a weighted normalized decision matrix is formed:

$$\overline{V(SC_p)} = \begin{pmatrix} v_{1p1} & v_{1p2} & \dots & v_{1pr_p} \\ v_{2p1} & v_{2p2} & \dots & v_{2pr_p} \\ \dots & \dots & \dots & \dots \\ v_{sp1} & v_{sp2} & \dots & v_{spr_p} \end{pmatrix}, \quad (7)$$

whose elements are determined by the formula:

$$v_{upr_p} = w_u \cdot \mu_{upr_p}, u = \overline{1,s}. \quad (8)$$

With equal criterion weights, the common factor 1/9 rescales the TOPSIS distances but does not change the relative-closeness coefficients or the ranking. It is retained to preserve the general weighted-matrix notation and to allow unequal expert weights in future applications. A formal sensitivity analysis of alternative weights is outside the scope of this illustrative validation and is included among the study limitations.

This stage allows us to consider not only the normalized assessments of social groups according to individual criteria, but also the importance of each criterion for the overall assessment of public attitudes towards artificially created content.

After applying the Z-shaped membership function, larger values of μ_{upr_p} correspond to lower social risk and a safer public perception, so all criteria in this problem are considered utility criteria. Therefore, the positive ideal solution for the social indicator SC_p is defined as:

$$A_p^+ = \left\{\max_{r_p} v_{1pr_p}, \max_{r_p} v_{2pr_p}, \dots, \max_{r_p} v_{spr_p}\right\}, \quad (9)$$

and the negative ideal solution:

$$A_p^- = \left\{\min_{r_p} v_{1pr_p}, \min_{r_p} v_{2pr_p}, \dots, \min_{r_p} v_{spr_p}\right\}. \quad (10)$$

The positive ideal solution A_p^+ corresponds to the best profile of public attitude towards artificially created content within the social indicator SC_p. The negative ideal solution A_p^-, on the contrary, characterizes the worst profile, i.e., the state with the highest manifestation of social risk.

At the next stage, for each group of social indicator sc_{pr_p} the distance to the positive ideal solution is calculated:

$$d_{pr_p}^+ = \sqrt{\sum_{u=1}^{s}\left(v_{upr_p} - A_{up}^+\right)^2}, \quad (11)$$

and the distance to the negative ideal solution:

$$d_{pr_p}^- = \sqrt{\sum_{u=1}^{s}\left(v_{upr_p} - A_{up}^-\right)^2}, \quad (12)$$

where A_{up}^+ and A_{up}^- are respectively the positive and negative ideal values according to the criterion f_u for the social indicator SC_p.

The integral index of safe public perception of artificially created content for the group sc_{pr_p} is defined as the relative proximity to the positive ideal solution:

$$m_{SR}\left(sc_{pr_p}\right) = \frac{d_{pr_p}^-}{d_{pr_p}^+ + d_{pr_p}^-}, \quad 0 \le m_{SR}(sc_{pr_p}) \le 1. \quad (13)$$

The value of $m_{SR}(sc_{pr_p})$ is an integral index of the safe public perception of artificially created content, for the corresponding group of social indicators. If the index value approaches 1, it indicates a safer, more trusting, or more controlled attitude towards artificially created content. If the value approaches 0, this indicates an increased level of social risk, distrust, or anxiety.

The results obtained characterize the level of public attitude towards artificially created content across individual social indicators. However, in practice, DM often needs to assess this level for different combinations of social characteristics, such as age, education, and income.

Next, the integral index of safe public perception of artificially created content, m_{SR} is derived. For this purpose, it is proposed to use multidimensional membership functions that allow aggregating normalized estimates in the $[0;1]$ interval. In particular, cone-shaped or pyramidal functions can be used, since their values approach 1 as all input estimates approach the maximum. The choice of a specific type of membership function is carried out by the system analyst during model tuning.

The conical membership function has the following form:

$$m_{SR} = \begin{cases} 1-\xi, & if\ \xi < 1 \\ 0, \text{otherwise.} \end{cases}$$

$$\xi = \frac{1}{\sqrt{l}} \cdot \sqrt{\left(m_{SR}(sc_{1r_p}) - 1\right)^2 + \dots + \left(m_{SR}(sc_{lr_p}) - 1\right)^2}. \quad (14)$$

To make the aggregation independent of the number of selected social dimensions, the distance is normalized by its maximum value $\sqrt{l}$. This normalization guarantees $\xi \in [0,1]$ for any l and removes the need for an unexplained fixed coefficient. For the three values used in the example (0.693, 0, and 0.307), the corrected aggregate is m_{SR}=0.276, which remains within the high-risk interval (0.2;0.4].

In the present study, the number of social indicators is fixed at l=3 and comprises age, education, and income. Accordingly, equation (14)

and the D1–D5 scale are applied only to this three-indicator scenario; direct comparison with configurations containing a different number of indicators is outside the scope of the study. Further validation of the aggregation scale for variable l is considered a direction for future research.

Here, $m_{SR}(sc_{pr_p})$ is the integral index of safe public perception of artificially created content for the selected group sc_{pr_p} of social indicator SC_p, $p = \overline{1,l}$ which can cover different social indicators. In this case, the scaling of coordinates is carried out using the vector $(2; 2; \dots; 2)$, and the center of the basis of the membership function is given by the unit vector$(1; 1; \dots; 1)$. Therefore, the closer m_{SR} is to 1, the safer the public perception is; the closer it is to 0, the higher the social risk is.

D – generalized level of social risk that is formed because of the distribution and perception of artificially created content in society.

Based on the integral index m_{SR} the generalized level of social risk D is determined. Since $m_{SR} \in [0; 1]$ characterizes the level of safe or positive social attitude, it is advisable to interpret D on this same scale through the corresponding risk intervals.

$$D = \begin{cases} D_1, & m_{SR} \in (0.8; 1]; \\ D_2, & m_{SR} \in (0.6; 0.8]; \\ D_3, & m_{SR} \in (0.4; 0.6]; \\ D_4, & m_{SR} \in (0.2; 0.4]; \\ D_5, & m_{SR} \in [0; 0.2]. \end{cases} \quad (15)$$

where D_1 is very low level of social risk; D_2 is low level of social risk; D_3 medium level of social risk; D_4 is high level of social risk; D_5 is very high level of social risk.

The linguistic labels describe a group-level decision-support profile and must not be interpreted as clinical or individual diagnoses. Very low and low risk indicate predominantly safe perception and routine monitoring; medium risk denotes a mixed profile requiring continued observation; high and very high risk indicate a concentration of unfavourable criteria and justify targeted follow-up assessment and preventive information measures. Distrust and concern are model inputs, whereas social risk is their multicriteria aggregate together with verification behaviour and regulatory expectations.

This interpretation is logical, since higher values of m_{SR} indicate a safer public attitude towards artificially created content, while lower values indicate increased distrust, anxiety, and social risk. The scale for interpreting social risk levels is based on an even division of the normalized interval $[0; 1]$ into five linguistic categories. This approach provides a transparent and reproducible interpretation of the integral index and allows for comparing results across different social groups.

Thus, the proposed fuzzy multi-criteria model M_{SR} provides a consistent transition from individual responses of respondents to the integral index m_{SR} and the generalized level of social risk D. The model combines social indicators, evaluation criteria, linguistic variables, membership functions, and the TOPSIS method, which allows for a formal assessment of public attitudes towards artificially created content in terms of individual social groups.

RESULTS AND DISCUSSION

This section presents the results of experimental testing of the proposed fuzzy multi-criteria model M_{SR}, designed to assess public attitudes towards artificially created content and determine the generalized level of social risk. The testing was performed on data from a questionnaire survey of 381 respondents collected in Ukraine in April 2026, using three social indicators and criteria reflecting trust, concern, willingness to verify information, and perceptions of risks associated with artificially created content [26]. The survey was conducted as an anonymous questionnaire, and the responses were used only for general scientific analysis. The sample included respondents with different socio-demographic characteristics, particularly age, level of education, and average net monthly income, enabling a comparative assessment of social risk within individual groups.

The goal of this stage is not only to demonstrate the computational procedure, but also to obtain comparable quantitative and linguistic estimates for different social groups. This allows us to reveal the variability in public attitudes towards artificially created content by age, education level, and income, as well as to identify groups with higher or lower levels of social risk.

The formed sample is sufficient for experimental testing of the model, as it includes respondents with different socio-demographic characteristics. This made it possible to obtain a general idea of the public attitude towards artificially created content among the studied group of citizens.

Table 1 summarizes the composition of the sample. Within the modelled scenario, the selected groups comprised 327 (85.8 %) respondents aged 18–34, 215 (56.4 %) respondents with a bachelor's degree, and 97 (25.5 %) respondents in the UAH 6,700–10,000 income category. Because age, education, and income are separate social indicators

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

and each respondent contributes to one category within each indicator, percentages are calculated separately for each indicator and must not be summed across indicators. In the full dataset (N=381), age was M=25.38 years (SD=10.61), with a median of 22 years and an IQR of 3 years. Education and income were collected as categorical variables; therefore, category frequencies and percentages are reported, whereas arithmetic means are not interpreted. One record was coded as age 17 and is reported separately in Table 1; its compliance with the inclusion criterion should be verified before the final analysis.

Because the 18–34 age group comprises 85.8% of the sample and the older subgroups are small, age-related findings are exploratory and should not be interpreted as population-representative comparisons of all age strata.

Experimental calculations were performed on the entire data sample using three social indicators and nine criteria to assess attitudes towards artificially created content. To demonstrate the logic of the calculations and ensure the reproducibility of the proposed approach, an example of assessing public attitudes towards artificially created content for individual combinations of social indicators, based on fragments of input data, is given below.

The DM defined the groups of social indicators as follows: for SC_1 (age of the respondent) – $sc_{11} = [18; 34)$; for SC_2 (highest level of completed education) – $sc_{23} =$ {higher education bachelor's degree}; for SC_3 (average net monthly income) – $sc_{32} =$ {6700 – 10000 UAH}. According to the real sample, we have: for sc_{11} – 327citizens; for sc_{23} – 215 citizens; for sc_{32} – 97 citizens.

For citizens from the set C, who belong to certain groups of social indicators, scores were obtained on the set of criteria K, that characterize their attitude toward artificially created content. Fragments of these scores, formed because of data processing using the information method of social risk assessment K_{SR}. To avoid presenting an incomplete illustrative fragment, the respondent-level table has been removed. The complete anonymized input records and the coding scheme used in all calculations are available in the open dataset [26], whereas the numerical example below begins with the aggregated term vectors derived from the full sample.

The following calculations are based on the multi-criteria model of social risk assessment in the context of the spread of artificially created content. At the first stage, for each assessment criterion, the resulting term scores $T^*$are determined within the corresponding social indicator groups. Their formation is based on the constructed rules of belonging, which allow generalizing citizens' individual responses into final linguistic scores:

Table 1. **Fragments of input linguistic data**

Social indicator	Category	n	%
Age, years	17 (outside the prespecified 18–99 range)	1	0.3
Age, years	18–34	327	85.8
Age, years	35–44	18	4.7
Age, years	45–54	23	6.0
Age, years	55–65	6	1.6
Age, years	66–99	6	1.6
Age, years	**Total**	**381**	**100.0**
Education	Complete general secondary education	97	25.5
Education	Vocational and technical education	31	8.1
Education	Bachelor's degree	215	56.4
Education	Master's degree	38	10.0
Education	**Total**	**381**	**100.0**
Monthly income	Up to UAH 6,700	142	37.3
Monthly income	UAH 6,700–10,000	97	25.5
Monthly income	UAH 10,000–14,000	46	12.1
Monthly income	UAH 14,000–18,000	28	7.3
Monthly income	UAH 18,000–22,000	20	5.2
Monthly income	UAH 22,000–26,000	14	3.7
Monthly income	UAH 26,000–30,000	10	2.6
Monthly income	UAH 30,000–34,000	10	2.6
Monthly income	More than UAH 34,000	14	3.7
Monthly income	**Total**	**381**	**100.0**

***Source*: compiled by the authors**

$T_1^*(sc_{11}) =$
$(t_{12}^*;\ t_{22}^*;\ t_{32}^*;\ t_{43}^*;\ t_{52}^*;\ t_{63}^*;\ t_{73}^*;\ t_{83}^*;\ t_{94}^*);$
$T_2^*(sc_{23}) =$
$(t_{13}^*;\ t_{22}^*;\ t_{32}^*;\ t_{43}^*;\ t_{52}^*;\ t_{63}^*;\ t_{73}^*;\ t_{83}^*;\ t_{94}^*);$
$T_3^*(sc_{32}) =$
$(t_{13}^*;\ t_{22}^*;\ t_{32}^*;\ t_{43}^*;\ t_{52}^*;\ t_{63}^*;\ t_{73}^*;\ t_{83}^*;\ t_{93}^*).$

After the formation of the resulting term estimates T^* the decision matrix SR is formed.

After determining the resulting term estimates, they are transformed into normalized quantitative values using the membership function (4). This ensures the comparability of the obtained estimates and allows forming the corresponding decision matrix:

$$\overline{SR(SC_p)} = \begin{pmatrix} 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \\ 0 & 0 & 0 \\ 0.5 & 0.5 & 0.5 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.222 & 0.222 & 0.222 \\ 0 & 0 & 0.222 \end{pmatrix}.$$

Criteria that have identical values for all groups in the illustrative matrix do not contribute to the TOPSIS distances. They are retained in the general model because their discriminatory contribution may differ in other samples; therefore, the numerical result is interpreted as scenario-specific.

Further multi-criteria evaluation of alternatives, represented in the matrix $\overline{SR(SC_p)}$, can be carried out using the TOPSIS method. In this case, it is assumed that the weight coefficients of the criteria are equivalent: $w_u = \frac{1}{9} = 0.111, u = \overline{1,9}$.

1. Weighted normalized matrix. After multiplying each element by $w_u = 0.111$ we get:

$$\overline{V(SC_p)} = \begin{pmatrix} 0.056 & 0 & 0 \\ 0.056 & 0.056 & 0.056 \\ 0.056 & 0.056 & 0.056 \\ 0 & 0 & 0 \\ 0.056 & 0.056 & 0.056 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.025 & 0.025 & 0.025 \\ 0 & 0 & 0.025 \end{pmatrix}.$$

2. Positive and negative ideal solutions, by (9) and (10): $A_p^+ = (0.056;\ 0.056;\ 0.056;\ 0;\ 0.056;\ 0;\ 0;\ 0.025; 0.025), A_p^- = (0;\ 0.056;\ 0.056;\ 0;\ 0.056;\ 0;\ 0;\ 0.025;\ 0).$

3. Distances to ideal solutions. For three groups of social indicators, obtained by (11) and (12):

$$d_{p1}^+ = 0.025, d_{p1}^- = 0.056;$$
$$d_{p2}^+ = 0.061, d_{p2}^- = 0;$$
$$d_{p3}^+ = 0.056, d_{p3}^- = 0.025.$$

4. The integral index of safe public perception of artificially created content for social groups is determined by (13):

$$m_{SR}(sc_{11}) = \frac{0.056}{0.025 + 0.056} = 0.693,$$
$$m_{SR}(sc_{23}) = \frac{0}{0.061 + 0} = 0,$$
$$m_{SR}(sc_{32}) = \frac{0.025}{0.056 + 0.025} = 0.307.$$

Thus, based on the results of applying the TOPSIS method, integral indices of safe public perception of artificially created content were obtained for three groups of social indicators.

Next, the integral index of safe public perception of artificially created content m_{SR} is derived by (14): m_{SR} =0.276.

Based on the integral index m_{SR} the generalized level of social risk D is determined: $m_{SR} \in (0.2; 0.4] \rightarrow D_4$ – high level of social risk.

The selected age, education, and income categories are not mutually exclusive and may overlap within the sample. The values 0.693, 0, and 0.307 are relative coefficients obtained for the fixed illustrative comparison set; in particular, 0 denotes proximity to the negative ideal within this set and not an absolute absence of safe perception. Therefore, m_{SR}=0.276 characterizes only the selected scenario and should be interpreted as an exploratory group-level result.

Aggregation of the obtained values for the combination of social characteristics sc_{11}, sc_{23}, sc_{32} gave an integral index $m_{SR} = 0.276$. According to the proposed interpretation scale, this value falls within the interval $(0.2; 0.4]$, corresponding to level D_4– high social risk. Thus, the studied combination of social groups is characterized by heightened concern about the spread of artificially generated content, confirming the practical feasibility of using the model to identify groups with potentially greater vulnerability to information risks.

The results demonstrate the computational feasibility and interpretability of using a fuzzy multi-criteria model to assess social risk in the context of the spread of artificially generated content. The proposed approach allows us to move from

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

individual responses of respondents to a formalized quantitative-linguistic result that reflects the levels of public attitude, trust, concern, and potential information vulnerability across different social groups.

Experimental verification of the model was conducted using questionnaire data from 381 respondents collected in Ukraine in April 2026. The study considered three social indicators: age, level of education, and average net monthly income. For applied testing, a combination of social groups was considered: age group – 327 citizens; group of respondents with a bachelor's degree – 215 citizens; and group with an income of 6700–10000 UAH – 97 citizens. According to the aggregation results, an integral index of public attitude towards artificially created content m_{SR}=0.276, was obtained, indicating a high level of social risk.

The obtained value indicates a noticeable concern among respondents about the spread of content created by artificial intelligence, especially in the fields of political and social news, deepfakes, and fake videos, and their influence on the formation of public opinion. The value m_{SR}= 0.276 shows that the studied combination of social characteristics is characterized by increased distrust, the need to verify information, and the expectation of more effective mechanisms of public or state response.

The results are consistent with recent studies emphasizing the social nature of artificially generated content. Groh et al. [1] showed that recognition of political deepfakes depends on the format of information presentation, while Diel et al. [8] found that human detection of deepfake content is often limited and may approach random guessing. This confirms that social risk should be assessed not only by the technical presence of deepfakes, but also by citizens' trust, concern, and behavioral responses.

Studies by Li and Yang [2] and Wittenberg et al. [9] show that labeling AI-generated content is useful but insufficient for reducing risk, as it does not always significantly affect trust or sharing intentions. The proposed model complements these approaches by quantifying how different social groups perceive artificially generated content and identifying groups with higher social risk.

Methodologically, the approach is close to the works of Kaul and Guaba [24] and Gaeta et al. [25], which apply multi-criteria and fuzzy methods to information risks and disinformation. However, the proposed model combines linguistic responses, fuzzy membership functions, TOPSIS evaluation, and integral interpretation of social risk, increasing the transparency and practical usefulness of the results.

The main advantage of the model is its ability to process heterogeneous and subjective data. Respondents' answers reflect trust, fear, caution, and critical perception, which are difficult to express as exact numbers. Fuzzy logic makes it possible to formalize these assessments and convert them into normalized values suitable for multi-criteria analysis.

The model is also flexible and can be adapted to different social indicators, criteria, response types, and respondent groups. Additional criteria, such as digital literacy, social media use, media trust, or experience with AI-generated content, can be included if needed.

The practical value of the model lies in identifying social groups with higher levels of distrust, anxiety, or information vulnerability. Its results can support targeted media literacy programs, explanatory campaigns, information verification tools, and mechanisms for labeling artificially generated materials.

The practical interpretation is group-specific. In the present example, the age group sc_{11} (0.693; low risk) requires routine monitoring and general awareness measures; the income group sc_{32} (0.307; high risk) justifies targeted media-literacy and verification support; and the bachelor's-degree group sc_{23} (0; very high risk) is a priority for a follow-up diagnostic survey before intervention. Public information-security bodies and media-literacy programme managers can use these outputs to allocate outreach, select messages, and schedule repeated measurement. The group index does not justify conclusions or restrictions at the individual level.

At the same time, the study has limitations. The testing was based on respondents from Ukraine, so the results cannot be directly generalized to other socio-cultural contexts. In addition, questionnaire data reflect self-assessments, which may differ from real behavior. The proposed risk interpretation scale also requires further validation on larger samples.

The experiment demonstrates the computational feasibility and interpretability of the proposed model rather than predictive accuracy. Further validation should test sensitivity to criterion weights, membership-function parameters, aggregation rules, and risk-interval boundaries, as well as replication on an independent sample.

Thus, the proposed fuzzy multi-criteria model has scientific and practical value, as it represents public

attitudes toward artificially generated content through an integral index and a linguistic level of social risk. Further research should expand the sample, refine criteria weights, conduct cross-country comparisons, and test the model under real information influence.

CONCLUSIONS

The study addresses the scientific problem of formalizing the assessment of social risk arising from the spread of artificially created content. For this purpose, a fuzzy multi-criteria model has been developed that combines respondents' social characteristics, linguistic assessments of their attitudes, membership functions, and the TOPSIS ranking procedure. The proposed approach provides a transition from disparate questionnaire responses to a generalized quantitative-linguistic indicator suitable for interpreting the level of safe public perception and the social risk associated with it.

The results demonstrate the computational feasibility of the model for an illustrative within-sample comparison of selected social groups. Based on survey data from 381 respondents, the integral index of safe public perception of artificially created content was determined to be $m_{SR} = 0.276$, indicating an insufficient level of safe perception. The corresponding social risk index corresponds to a high level of social risk. This indicates the presence of increased distrust, anxiety, and the need for more active mechanisms for the verification and regulation of artificially created information content. The obtained numerical values should be interpreted as exploratory scenario estimates rather than as evidence of predictive validity or population-level representativeness.

The scientific novelty of the study lies in formalizing the social risk of the spread of AI-generated content as a multi-criteria fuzzy problem, in which respondents' subjective linguistic responses are converted into normalized quantitative assessments and aggregated into an integral indicator. Unlike approaches focused mainly on technical detection of deepfakes or labeling of artificially created content, the proposed model allows us to assess the social perception, level of trust, and potential information vulnerability of individual population groups.

The practical value of the model lies in its use as a decision-support tool in information security, digital literacy, and communication policy. Within the analysed example, the model distinguished a low-risk age group, a high-risk income group, and a very-high-risk education group. These findings demonstrate how group priorities can be derived from the index and linked to routine monitoring, targeted media-literacy support, or follow-up diagnosis. The results are screening outputs for population groups and should be validated using subgroup sizes, sensitivity analysis, and repeated measurement before policy decisions are made.

Prospects for further research include expanding the empirical sample, conducting interregional and intercountry comparisons, refining the weights of the criteria using expert methods, and testing the model on behavioral data. A separate promising direction is the creation of a software module or an intellectual and analytical platform for automated monitoring of social risk associated with the spread of artificially created content.

Conflict of interests: Authors declare that they do not have conflict of interests, in particular financial, personal, copyright or any of a different nature, which could have influenced the research, as well as the results published in this articles

Financing: Research was carried out without financial support

Accessibility data: The anonymized dataset containing responses from 381 participants and the coded variables used in the calculations is openly available in the Zenodo repository [26]: https://doi.org/10.5281/zenodo.20056389. The model, its parameters, and the computational procedure are described in the manuscript.

Using artificial intelligence tools: Authors confirm that they did not use instrumental AI tools for writing this one work

REFERENCES

1. Groh, M., Epstein, Z., Firestone, C. & Picard, R. “Human detection of political speech deepfakes across transcripts, audio, and video”. *Nature Communications*. 2024; 15: 8128. DOI: https://doi.org/10.1038/s41467-024-51998-z.

2. Li, F. & Yang, Y. “Impact of artificial intelligence-generated content labels on perceived accuracy, message credibility, and sharing intentions for misinformation: Web-based, randomized, controlled experiment”. *JMIR Formative Research*. 2024; 8: e60024. DOI: https://doi.org/10.2196/60024.

3. Lorenz, P., Perset, K. & Berryhill, J. "Initial policy considerations for generative artificial intelligence". *OECD Artificial Intelligence Papers*. 2023; 1. DOI: https://doi.org/10.1787/fae2d1e6-en.

4. Lovato, J., Zhao, Y., Wan, A., Pasquetto, I. V. & Hancock, J. T. "Diverse misinformation: Impacts of human biases on detection of deepfakes on networks". *npj Complexity*. 2024; 1: 4. DOI: https://doi.org/10.1038/s44260-024-00006-y.

5. Miao, F. & Holmes, W. "Guidance for generative AI in education and research". *UNESCO*. 2023. – Available from: https://unesdoc.unesco.org/ark:/48223/pf0000386693 – [Accessed: May 7, 2026].

6. Hwang, Y., Ryu, J. Y. & Jeong, S.-H. "Effects of disinformation using deepfake: The protective effect of media literacy education". *Cyberpsychology, Behavior, and Social Networking*. 2021; 24 (3): 188–193. DOI: https://doi.org/10.1089/cyber.2020.0174.

7. Groh, M., Epstein, Z., Firestone, C. & Picard, R. "Deepfake detection by human crowds, machines, and machine-informed crowds". *Proceedings of the National Academy of Sciences*. 2022; 119 (1): e2110013119. DOI: https://doi.org/10.1073/pnas.2110013119.

8. Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M. & Bäuerle, A. "Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers". *Computers in Human Behavior Reports*. 2024; 16: 100538. DOI: https://doi.org/10.1016/j.chbr.2024.100538.

9. Wittenberg, C., Epstein, Z., Peloquin-Skulski, G., Berinsky, A. J. & Rand, D. G. "Labeling AI-generated media online". *PNAS Nexus*. 2025; 4 (6): pgaf170. DOI: https://doi.org/10.1093/pnasnexus/pgaf170.

10. Williams-Ceci, S., Macy, M. W. & Naaman, M. "Misinformation does not reduce trust in accurate search results, but warning banners may backfire". *Scientific Reports*. 2024; 14: 10977. DOI: https://doi.org/10.1038/s41598-024-61645-8.

11. Verma, N. "'One video could start a war': A qualitative interview study of public perceptions of deepfake technology". *Proceedings of the Association for Information Science and Technology*. 2024; 61 (1): 374–385. DOI: https://doi.org/10.1002/pra2.1035.

12. Jacobsen, B. N. & Simpson, J. "The tensions of deepfakes". *Information, Communication & Society*. 2024; 27 (6): 1095–1109. DOI: https://doi.org/10.1080/1369118X.2023.2234980.

13. Hynek, N., Gavurova, B. & Kubak, M. "Risks and benefits of artificial intelligence deepfakes: Systematic review and comparison of public attitudes in seven European countries". *Journal of Innovation & Knowledge*. 2025; 10 (5): 100782. DOI: https://doi.org/10.1016/j.jik.2025.100782.

14. Park, S. & Nan, X. "Generative AI and misinformation: A scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation". *AI & Society*. 2025. DOI: https://doi.org/10.1007/s00146-025-02620-3.

15. Ma, R., Zhang, X. & Li, Y. "Fighting fake news in the age of generative AI: Strategic responses and governance implications". *Technological Forecasting and Social Change*. 2025. DOI: https://doi.org/10.1016/j.techfore.2025.123996.

16. Kumar, R., Goddu, B., Saha, S. & Jatowt, A. "Silver lining in the fake news cloud: Can large language models help detect misinformation?" *IEEE Transactions on Artificial Intelligence*. 2025; 6 (1): 14–24. DOI: https://doi.org/10.1109/TAI.2024.3440248.

17. Moyo, B. V. C., Tuyikeze, T., Matsebula, F. & Obagbuwa, I. C. "An AI-driven conceptual framework for detecting fake news and deepfake content: A systematic review". *Frontiers in Artificial Intelligence*. 2026; 9: 1737790. DOI: https://doi.org/10.3389/frai.2026.1737790.

18. Shin, D., Koerber, A. & Lim, J. S. "Impact of misinformation from generative AI on user information processing: How people understand misinformation from generative AI". *New Media & Society*. 2025; 27 (7): 4017–4047. DOI: https://doi.org/10.1177/14614448241234040.

19. Bashardoust, A., Feuerriegel, S. & Shrestha, Y. R. "Comparing the willingness to share for human-generated vs. AI-generated fake news". *Proceedings of the ACM on Human-Computer Interaction*. 2024; 8 (CSCW2): 1–21. DOI: https://doi.org/10.1145/3687028.

20. Ching, D., Twomey, J., Aylett, M. P., Quayle, M., Linehan, C. & Murphy, G. "Can deepfakes manipulate us? Assessing the evidence via a critical scoping review". *PLOS ONE*. 2025; 20 (5): e0320124. DOI: https://doi.org/10.1371/journal.pone.0320124.

21. Plohl, N., Mlakar, I., Aquilino, L., Bisconti, P. & Smrke, U. "Development and validation of the Perceived Deepfake Trustworthiness Questionnaire (PDTQ) in three languages". *International Journal of*

Human–Computer Interaction. 2025; 41 (11): 6786–6803. DOI: https://doi.org/10.1080/10447318.2024.2384821.

22. Xu, Z., Wen, X., Zhong, G. & Fang, Q. "Public perception towards deepfake through topic modelling and sentiment analysis of social media data". *Social Network Analysis and Mining*. 2025; 15: 16. DOI: https://doi.org/10.1007/s13278-025-01445-8.

23. Romero Moreno, F. "Generative AI and deepfakes: A human rights approach to tackling harmful content". *International Review of Law, Computers & Technology*. 2024; 38 (3): 297–326. DOI: https://doi.org/10.1080/13600869.2024.2324540.

24. Kaul, A. & Guaba, R. "Communication through social media: Fake or reality". In *Communication Management*. London, UK: IntechOpen. 2021. DOI: https://doi.org/10.5772/intechopen.99719.

25. Gaeta, A., Loia, V. & Orciuoli, F. "An explainable prediction method based on Fuzzy Rough Sets, TOPSIS and hexagons of opposition: Applications to the analysis of Information Disorder". *Information Sciences*. 2024; 659: 120050. DOI: https://doi.org/10.1016/j.ins.2023.120050.

26. Polishchuk, V., Matei, A. & Polishchuk, I. "Data from 381 respondents for assessing social risk in the context of the spread of artificially created content (Ukraine)". *Zenodo*. 2026. – Available from: https://doi.org/10.5281/zenodo.20056389. – [Accessed: 6 August 2026].

27. Uhryn, D. I., Karachevtsev, A. O., Tomka, Y. Ya., Zakharov, M. M. & Troianovska, Y. L. "Information system for analyzing public sentiment in web platforms based on machine learning". *Herald of Advanced Information Technology*. 2024; 7(2): 199–212. DOI: https://doi.org/10.15276/hait.07.2024.14.

28. Tereikovskyi, I. A. & Korovii, O. S. "A model for constructing neural network systems for recognizing emotions of text fragments". *Herald of Advanced Information Technology*. 2025; 8 (2): 197–208. DOI: https://doi.org/10.15276/hait.08.2025.12.

29. Kosiv, Y. A. & Yakovyna, V. S. "Three language political leaning text classification using natural language processing methods". *Applied Aspects of Information Technology*. 2022; 5 (4): 359–370. DOI: https://doi.org/10.15276/aait.05.2022.24.

Received 28.09.2026
Received after revision 23.10.2026
Accepted 30.10.2026

DOI: https://doi.org/10.15276/hait.09.2026.45
УДК 004.8:316.4:005.52

Нечітка багатокритеріальна модель оцінювання соціального ризику на основі суспільного ставлення до штучно створеного контенту

Поліщук Володимир Володимирович[1,2)]
ORCID: https://orcid.org/0000-0003-4586-1333; volodymyr.polishchuk@tuke.sk. Scopus Author ID: 57204564304
Влох Івона[3)]
ORCID: https://orcid.org/0000-0002-9969-0827; iwloch@prz.edu.pl. Scopus Author ID: 6506235489
Матей Андрій Андрійович[2)]
ORCID: https://orcid.org/0009-0001-0280-1763; andrii.matei@uzhnu.edu.ua.
Млавець Юрій Юрійович[2)]
ORCID: https://orcid.org/0000-0002-1480-9017; yurii.mlavets@uzhnu.edu.ua. Scopus Author ID: 55123275500
[1)] Технічний університет в Кошицях, вул. Летна, 1/9. Кошице, 042 00, Словаччина
[2)] ДВНЗ «Ужгородський національний університет», пл. Народна, 3. Ужгород, 88000, Україна
[3)] Жешувська політехніка, алея Повстанців Варшави, 12, 35 029 Жешув, Польща

АНОТАЦІЯ

Штучно створений контент може підривати довіру до інформації, підвищувати вразливість до маніпуляцій і формувати групові соціальні ризики, які складно описати чіткими показниками **Мета.** Розробити та емпірично перевірити нечітку багатокритеріальну модель, яка перетворює неоднорідні анкетні відповіді на групову оцінку соціального ризику **Завдання.** Формалізувати соціальні показники й критерії, зорієнтувати лінгвістичні відповіді за напрямом ризику, порівняти соціально-демографічні групи та визначити правила інтерпретації індексу **Методи.** Дані анонімного анкетування трьохсот вісімдесяти

одного респондента з України опрацьовано із застосуванням порядкового перекодування, Z-подібних функцій належності, рівних ваг критеріїв, методу TOPSIS і нормованого за розмірністю конічного агрегування **Результати.** Групові індекси становили 0,693 для респондентів віком від вісімнадцяти до тридцяти трьох років, 0 для респондентів із дипломом бакалавра та 0,307 для групи з доходом 6700–10000 грн. Після нормованого агрегування отримано 0,276, що відповідає високому соціальному ризику **Наукова новизна.** Соціальний ризик формалізовано як формативний нечіткий багатокритеріальний конструкт, що інтегрує довіру, занепокоєння, поведінку щодо перевірки інформації, регуляторні очікування та соціально-демографічне групування **Практична значимість.** Модель може допомагати органам інформаційної безпеки та керівникам програм медіаграмотності визначати пріоритетні групи, добирати заходи й планувати повторний моніторинг; вона не є інструментом індивідуальної діагностики **Висновки.** Модель забезпечує інтерпретоване групове скринінгове оцінювання, однак результати потребують перевірки з урахуванням складу вибірки, чисельності перетинів підгруп і аналізу чутливості.

Ключові слова: штучно створений контент; соціальний ризик; нечітка модель; підтримка рішень; інформаційна безпека; багатокритеріальне оцінювання

ABOUT THE AUTHORS



Volodymyr V. Polishchuk - Doctor of Engineering Sciences, Professor, Department of Flight training. Technical University of Košice, 1/9 Letná, 042 00, Slovakia. Department of Software Systems, Uzhhorod National University, 3, Narodna Square, 88 000, Ukraine
ORCID: https://orcid.org/0000-0003-4586-1333; volodymyr.polishchuk@tuke.sk. Scopus Author ID: 57204564304
***Research field*:** Information technologies; hybrid decision support models; fuzzy knowledge modeling methods

Поліщук Володимир Володимирович - доктор технічних наук, професор. Кафедра льотної підготовки. Технічний університет в Кошицях, вул. Летна, 1/9. Кошице, 042 00, Словаччина. Кафедра програмного забезпечення систем. ДВНЗ "Ужгородський національний університет", пл. Народна, 3. Ужгород, 88000, Україна
Галузь досліджень: інформаційні технології; гібридні моделі підтримки рішень; методи нечіткого моделювання знань



Iwona Włoch - Associate Professor, Department of Discrete Mathematics, Faculty of Mathematics and Applied Physics. Rzeszow University of Technology, 12, al. Powstańców Warszawy, 35 029, Poland
ORCID: https://orcid.org/0000-0002-9969-0827; iwloch@prz.edu.pl. Scopus Author ID: 6506235489
***Research field*:** Discrete mathematics; Graph theory; Combinatorics; Fibonacci sequences; Mathematical modelling

Влох Івона – доктор габілітований, професор, кафедра Дискретної математики. Жешувська політехніка, алея Повстанців Варшави, 12, 35 029 Жешув, Польща
Галузь досліджень: Дискретна математика; Теорія графів; Комбінаторика; Послідовності Фібоначчі; Математичне моделювання



Andrii A. Matei – PhD Student, Department of Software Systems. Uzhhorod National University, 3, Narodna Square, 88 000, Ukraine
ORCID: https://orcid.org/0009-0001-0280-1763; andrii.matei@uzhnu.edu.ua
***Research field*:** Software development; information technology; risks

Матей Андрій Андрійович – аспірант кафедри Програмного забезпечення систем. ДВНЗ "Ужгородський національний університет", пл. Народна, 3. Ужгород, 88000, Україна
Галузь досліджень: розробка програмного забезпечення; інформаційні технології; ризики



Yurii Yu. Mlavets - Candidate of Physical and Mathematical Sciences, Associate Professor, Department of Cybernetics and Applied Mathematics. Uzhhorod National University, 3, Narodna Square, 88 000, Ukraine
ORCID: https://orcid.org/0000-0002-1480-9017; yurii.mlavets@uzhnu.edu.ua. Scopus Author ID: 55123275500
Research field: Stochastic processes and fields; banach spaces; stochastic modelling; intelligent data analysis

Млавець Юрій Юрійович - кандидат фізико-математичних наук, доцент кафедри Кібернетики і прикладної математики. ДВНЗ "Ужгородський національний університет", пл. Народна, 3. Ужгород, 88000, Україна
Галузь досліджень: розробка програмного забезпечення; інформаційні технології; ризики