КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

UDC 004.056:004.8:004.738.5
IRSTI 81.93.29; 28.23.20
https://doi.org/10.53364/24138614_2026_40_1_16

Ye. Makatov[1*], **Abdul Razaque**[2], **A. Ye. Makatova**[1]
[1]Shokan Ualikhanov Kokshetau University; Kokshetau, Kazakhstan
[2]Computer Science, Professor, Department of Computer Science and Mathematics, Seton Hall University, NJ USA

[*]*E-mail: m.yerkhan@list.ru*

COGNITIVE MODEL FOR PROTECTING SOCIAL MEDIA USERS: ARCHITECTURE AND PRINCIPLES

Abstract. *Social networks have become complex socio-digital ecosystems exposed to misinformation, phishing, and manipulative influence on users' affect. The topic is timely due to the acceleration of digital risk driven by artificial intelligence (AI) and large-scale automation. The subject of this study is a cognitive protection architecture for social media; the objective is to substantiate a model that integrates perception, interpretation, memory, decision-making, and an ethical filter. The tasks are to: (I) review approaches in behavioral analytics, affective computing, and explainable AI (XAI) relevant to social-media security; (II) develop an architectural framework capable of multimodal processing; (III) specify component roles and interconnections with emphasis on user interaction and transparency; and (IV) demonstrate novelty over rule-based and machine learning (ML) systems via integrated XAI, attention-based contextual modeling, and emotional-semantic analysis. Methods include hierarchical information processing; an integrative threat index $T=\alpha E+\beta B+\gamma S+\delta C$; Bayesian trust updating; ontological reasoning; softmax over the ethically admissible action set ethically admissible action set (A_{eth}); feedback-driven adaptation; and privacy-preserving mechanisms (federated learning, differential privacy* (*DP)). The main results demonstrate architectural coherence and functional feasibility, provide a mapping from threat levels to adaptive responses, and embed XAI interfaces while aligning with the General Data Protection Regulation (GDPR) and the EU Artificial Intelligence Act (EU AI Act) requirements. Conclusions and implications: by coupling cognitive depth with interpretability and privacy, the architecture enhances reliability, personalization, and trust; future work should extend toward multilingual and cross-cultural adaptation, neuro-symbolic integration, and participatory human-in-the-loop training; prototype verification will target phishing, manipulative content, and varied-trust sources with metrics spanning accuracy, false alarms, user trust, compute, and compliance.*

Keywords: *cognitive security architecture; multimodal perception; affective computing; explainable artificial intelligence; threat ontology; ethical decision-making; federated learning.*

Introduction.

Social networks have evolved into complex socio-digital environments facing a wide spectrum of cyber threats, including misinformation, phishing, and manipulative influence on users' emotional states. Recent telemetry shows sharp automation-driven growth (e.g., 36,000 s^{-1} in 2024); therefore, we focus on cognitively grounded, explainable protection [1].

Detecting such threats is challenging due to semantic ambiguity, emotional undertones, and contextual variability. Existing solutions are typically rule-based or rely on conventional machine

learning: rule-based approaches are rigid and difficult to scale, whereas ML models depend on large labeled datasets, are sensitive to noise and imbalance, and often lack transparency—rarely capturing affective or cognitive aspects of interaction, which undermines interpretability and trust [2].

The key research problem is the lack of cognitive depth in current digital security frameworks. By overlooking behavioral context, emotional interpretation, and semantic nuance, existing systems fail to provide adaptive, reliable decision-making under uncertainty. To address this limitation, we propose a cognitive model—an architecture that simulates human-like perception, interpretation, and ethically grounded decision-making in digital environments.

Objective. To design and substantiate a cognitive protection architecture for social networks that integrates perceptual, interpretive, and ethically adaptive modules, enabling transparent and context-aware threat assessment. The study pursues the following tasks: (I review approaches in behavioral analysis, affective computing, and explainable AI applied to social media security; (II) develop an architectural framework including perception, interpretation, memory, decision-making, and ethical filtering; (III) define the roles and interconnections of these components with emphasis on user interaction and transparency; and (IV) demonstrate the novelty of the approach versus rule-based and ML systems, (e.g., SHapley Additive exPlanations (SHAP)) attention-based contextual modeling, and emotional-semantic analysis; ethical adaptation is outlined conceptually and grounded in literature beyond anomaly detection and XAI surveys [3].

This research aims to close the gap between formal AI-driven systems and human-centered digital security. By embedding cognitive processing and interpretability into the architecture, it seeks to improve adaptability, trust, and ethical sustainability in social network protection.

Materials and methods.

The architecture of the cognitive protection model for social networks (Fig. 1) is grounded in the principles of hierarchical information processing. It is organized into three interdependent macro-blocks: perception, cognitive processing, and interpretation.

Data collection→Behavior analysis→Risk assessment→User feedback→System adaptation

Figure 1- Architectural structure of the cognitive model for digital user protection.

At the input level, the system gathers multimodal data — texts, images, behavioral traces, social and biometric signals. The cognitive core handles uncertainty through memory, ontologies, and probabilistic risk models. The top layer ensures ethical filtering, adaptation, and explainable AI (XAI), securing transparency and compliance.

To operationalize the integrative threat-assessment architecture, we formalize risk as a weighted sum of factors:

$$T=\alpha\cdot E+\beta\cdot B+\gamma\cdot S+\delta\cdot C \quad (1)$$

where T denotes the final threat score, E is the emotional tone, B the degree of behavioral anomaly, S the trustworthiness of the source or social context, C the semantic risk, and α, β, γ, δ are the corresponding weights. This formulation consolidates heterogeneous modalities into a unified cognitive metric of risk and provides a foundation for adaptive personalization.

The weight intervals are defined heuristically but are grounded in established multimodal AI practices of adaptive fusion. Contemporary research demonstrates that modality weights should reflect both informativeness and signal quality: modality-adaptive transformers dynamically recalibrate weights [4], while AdaMoW networks apply modality-specific weighting during fusion [5]. Robustness is further enhanced by distribution-based feature recovery and fusion, which down-weights weaker modalities on the fly [6]. The concept of adaptive weight bounds is also demonstrated in evolutionary training of neural networks, as in DEAW, which dynamically adjusts weight limits during optimization. In the domain of digital security, multimodal approaches to

misinformation detection [8] further reinforce the relevance of integrative modeling for threat assessment.

The proposed baseline intervals are: $\alpha\in[0.2,0.5]$ for the emotional factor; $\beta\in[0.2,0.4]$ for behavioral anomalies; $\gamma\in[0.1,0.3]$ for source trust or social context; and $\delta\in[0.1,0.4]$ for semantic risk. Weights are normalized to sum to one (e.g., $\alpha=0.30$, $\beta=0.25$, $\gamma=0.20$, $\delta=0.25$) and vary with environment, user profile, and threat type [4, 6].

To couple principled constraints with responsiveness to context and user feedback, we calibrate the weights via a projected update:

$$w_i^{(t+1)} = \prod_{\Omega}\left(w_i^{(t)} + \eta * \Delta^{(t)}\right), \quad W = \{w: \sum_i w_i = 1,\ \ell_i \le w_i \le u_i\}. \qquad (2)$$

Here $w^{(t)}=(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}, \delta^{(t)})$, $\eta\in[0.01,0.1]$ is the adaptation rate, $\Delta^{(t)}$ encodes performance- and feedback-driven increments, and Π_W projects onto the simplex with box bounds $[\ell_i,u_i]$. The projection enforces normalization and bounded adaptation.

The perception block collects and structures user signals and interactions, corresponding to sensorimotor and perceptual layers in human cognition, thereby shaping higher-level representations. At the sensory tier, the system gathers explicit behavioral traces — such as clicks, scrolling, viewing time, and latency — alongside implicit biometric parameters, including facial expressions, vocal intonations, pulse, and galvanic skin response, when available and provided with user consent. Biometric capture is strictly opt-in, defaults to on-device processing, and follows minimization principles to remain compliant with GDPR and the AI Act. These inputs enable detection of engagement patterns, stress indicators, and behavioral anomalies [8, 9].

Building on these signals, the perceptual tier translates textual, visual, and behavioral inputs into semantically meaningful features. Natural language processing (NLP) supports sentiment detection, toxic content recognition, misinformation identification, and the tracing of implicit threats or aggressive speech patterns. In parallel, computer vision (CV) techniques analyze images and videos to uncover manipulative symbols or harmful content, while semantic analysis isolates key entities, concepts, and thematic relations. Cross-modal alignment enhances accuracy while degrading gracefully under missing modalities [7]. If a modality is unavailable or consent is withdrawn, the fusion pipeline re-weights the remaining modalities and imputes neutral affect for the missing channel.

At the final stage, the system performs emotional-semantic annotation, fusing sensory and perceptual streams. This enables recognition of both the user's current affective state (such as anxiety or excitement) and the emotional atmosphere of the content (e.g., hostility or alarm). Simultaneously, semantic object extraction identifies actors, actions, relationships, and contexts relevant to potential threats. These profiles feed the cognitive block, where they align with user patterns, threat memory, and ethical filters (see Table 1).

Table 1 – Functional structure of the perception block

Level	Description	Example Technologies	Analysis Results	Output
Sensory	Collection of microbehavioral and biometric data	Web-tracking, eye-tracking, wearables	Engagement model, stress metrics, interaction patterns	Raw streams s_t, stress metrics
Perceptual (Analytical)	Feature processing of texts and images	NLP (BERT, RoBERTa), CV (YOLO, CLIP)	Sentiment, toxicity, fake content, visual risks	Feature vector p_t, σ_p
Emotional & Semantic	Emotional state + semantic extraction	Affective computing, sentiment fusion	Emotional profile, semantic threat map	$\{e_t,g_t\}$

Note – compiled by the authors

Beyond content/network baselines, we leverage multimodal inputs—textual, behavioral, and (with explicit consent) on-device, privacy-preserving affective cues—to detect latent threats in a context-aware manner, degrading gracefully under missing modalities.

The cognitive block interprets inputs, estimates risk under uncertainty, and selects adaptive strategies (see Table 2).

At its foundation lies a memory model, where behavioral patterns accumulate over time, forming persistent user profiles. This model stores typical response dynamics, a history of threat encounters, and stable user habits, often implemented with temporal database (DB) structures, thereby supporting recognition of recurring threats and semi-automated responses. Probabilistic judgments at this stage follow Bayes' theorem:

$$\mathrm{P}(\,\mathrm{Threat} \mid \mathrm{X}, \tau\,) = \frac{\mathrm{P}(\mathrm{X}|\mathrm{Threat},\tau)\,\mathrm{P}(\mathrm{Threat}|\tau)}{\mathrm{P}(\mathrm{X}|\tau)},\ P(X/\tau)=\sum_{y\in\{Threat,\overline{Threat}\}} P(X \mid y,\tau)P(y \mid \tau). \quad (3)$$

where X denotes observed features and τ is the trust prior. The denominator P(X|τ) follows from the law of total probability. This mechanism is conceptually aligned with Bayesian trust modeling approaches, where prior trust is iteratively updated based on accumulated evidence.

At the cognitive level, incoming situations are interpreted with contextual awareness. Indicators in user behavior or content are detected, and hypotheses about intentions of other participants are formed. Attention mechanisms, categorization, and probabilistic inference drive this stage. To prioritize actions, the system applies softmax normalization over the ethically admissible set A_{eth}:

$$\mathrm{P}(a_i)=\frac{exp\,(z_i/k)}{\sum_{a_j\in A_{eth}} exp\,(z_j/k)},\ k>0 \quad (4)$$

Here, a_i is an admissible action, A_{eth} denotes the ethically filtered action set, z is the vector of action scores (logits), z_i is the score of action a_i, and κ is the softmax temperature controlling stochasticity. Lower κ makes the distribution sharper (near-deterministic), higher κ makes it more uniform (exploratory).

At the ontological level, semantic relations are structured to support analogies and precedent-based reasoning. By integrating ontologies — for example, the Web Ontology Language (OWL) and the Resource Description Framework (RDF) — along with graph-based reasoning, the system supports analogical inference and contextual expansion through external knowledge.

At the social level, trust in a source is formalized as:

$$\mathrm{Trust}(u,s)=\theta_1 R_s+\theta_2 F_s+\theta_3 H_s,\ \theta_k\geq 0,\ \sum_k \theta_k=1 \quad (5)$$

where R_s is source reputation, F_s community confirmation, and H_s historical reliability. Weights θ are determined empirically or adaptively.

Table 2 – Functional Components of the Cognitive Block

Level	Functions	Example Technologies	Inputs	Results
Memory Model	Store behavioral patterns	Recurrent profiles, Temporal DB	User history m_t	Personal threat context, response history
Cognitive	Contextual interpretation	Cognitive modeling, affect mining	Feature vector z_t	Intention prediction, risk level $\hat{r}_t$
Ontological	Semantic understanding	OWL, RDF, Graph Reasoning	Semantic graph g_t	Formalized threats, analogical reasoning
Social	Trust evaluation	Trust modeling, graph analysis	Reputation/ feedback	Trust level, group relevance

Note – compiled by the authors

The final block — interpretation, adaptation, and ethics — transforms analytical outputs into transparent, user-centered responses, ensuring personalization and compliance.

Explainable AI (XAI) interfaces are embedded throughout processing. The interpretability score is defined as:

$$\text{XAI\_Score} = \frac{\sum_{i=1}^{n}|\phi_i|\cdot I_i}{\sum_{i=1}^{n}|\phi_i|},\ I_i=1\{|\phi_i|>\text{Tcrit}\}. \quad (6)$$

where ϕ_i is the attribution weight of feature i, and I_i is an indicator. T_{crit} regulates granularity, while A_t sets overall exposition depth.

Adaptation personalizes responses according to user traits, emotional states, and prior feedback, consistent with emotionally adaptive tutoring systems that modulate support based on learners' affect. It follows an inertia mechanism:

$$At=\lambda A_{t-1}+(1-\lambda)\ Rt,\ \lambda\in[0,1],\ At\in[0,1] \quad (7)$$

where R_t is current feedback. If users ignore advice, notification intensity drops; if they follow, detail increases.

Privacy and compliance rely on federated learning, DP, and edge computing:

$$\Pr[M(D_1)\in S]\leq e^{\varepsilon}\Pr[M(D2)\in S]+\delta \quad (8)$$

supporting GDPR-oriented privacy compliance [10, 11].

The ethical layer targets fairness, autonomy, and rights, aligning with widely cited AI-ethics principles [12]. Ethical filters prune forbidden strategies before decision normalization (see Table 3).

Table 3 – Functional Role of the Components of the Final Block

Component	Purpose	Methods	Expected Outcomes	User-visible Artifacts
XAI Interface	Transparency	SHAP, LIME, attention	Trust, reduced anxiety	Explanation chart/report
Adaptation	Personalization	Reinforcement learning	Empathetic interaction	Customized detail depth
Privacy	Minimize risk	Federated Learning, DP	Compliance, protection	Privacy notice
Ethical Filter	Safeguard fairness	Bias mitigation	Fairness, autonomy	Action set A_{eth}

Note – compiled by the authors

Formally, the architecture is:

$$\text{Model}=\{L_i=(F_i,M_i)\mid i=1..n\} \quad (9)$$

where each level L_i is defined by functions F_i and methods M_i.

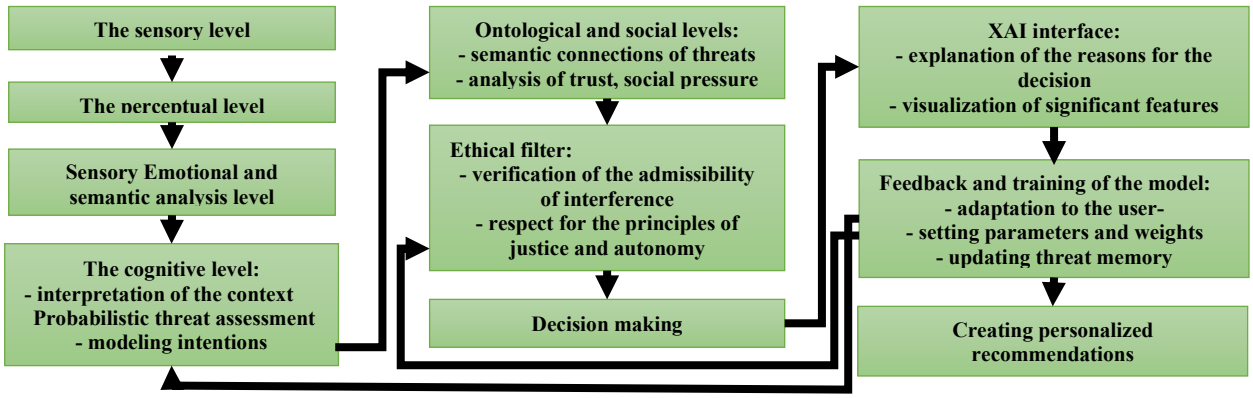

Figure 2 – Stages of Cognitive Threat Processing

From sensory input to ethical reasoning, all stages form a closed cognitive loop. Each stage — from perceptual filtering and emotional-semantic annotation to ontological reasoning, ethical evaluation, and XAI-supported feedback — completes a closed cognitive loop in which user feedback updates A_t, memory state m_t, and trust priors (see Table 4).

Table 4 – Mapping of Threat Levels to Adaptive Responses

Threat Level	Example Threat	Response	Fallback if uncertain
Behavioral	Atypical activity	Notification, monitoring	Passive monitoring
Emotional	Stress, anxiety	Gentle recommendations	Delay further alerts
Semantic	Manipulative phrasing	Content filtering	Human review
Social	Peer pressure	Source trust evaluation	Conservative scoring
Ethical	Potential harm	Blocking, XAI report	Escalate to human oversight
Note – compiled by the authors			

Threat levels are mapped to corresponding adaptive mechanisms, while functional structure by layers provides implementation detail (see Table 5).

Table 5 – Functional Structure by Layer

Model Layer	Functions	Examples	Governance hooks
Technical	Data collection, ML, NLP	DNNs, CLIP, BERT	Monitoring, audit logging
Cognitive	Interpretation, memory	Graph reasoning	Model calibration
Socio-psychologic	Trust, group influence	Social graphs	Group trust metrics
Organizational	Security, risk management	Policy configs	Policy configs, audit trails
Legal	Compliance, explanation	XAI, opt-out	Opt-out, GDPR/AI Act checks
Note – compiled by the authors			

Results and Discussion.

The results of the present study are expressed not in numerical benchmarks alone, but in the demonstration of the architectural consistency and functional feasibility of the proposed cognitive model for digital user protection. A prototype implementation integrating NLP and behavioral analysis modules has been developed to verify architectural feasibility and will be used for forthcoming experimental evaluation. The architecture embodies a closed cognitive loop, extending from sensory data acquisition to perceptual and semantic analysis, cognitive interpretation, ethical filtering, and adaptive feedback. This full-cycle design constitutes one of the central contributions of the work.

A further essential outcome is the formalization of an integrative threat index (see Eq. 1), which unifies emotional (E), behavioral (B), semantic (S), and social (C) parameters into a single cognitive metric of digital risk. This formulation enables dynamic, context-driven reallocation of weights, moving beyond rigid rules or static classifiers.

The novelty of the model becomes clearer when contrasted with conventional approaches. Table 6 provides an overview of how the proposed architecture extends beyond rule-based or machine learning systems.

Table 6 – Comparison of architectural layers in the proposed model versus traditional approaches

Architecture Level	Functions in Cognitive Model	Support in Rule-based / ML Systems	Distinctive Contribution
Sensory	Micro-behavioral and biometric data collection	Limited (content only)	Deep behavioral signals integrated
Perceptual	Text and image analysis (NLP, CV)	Partial (surface-level, text only)	True multimodal interpretation
Emotional-Semantic	Affective annotation (joy, anxiety, etc.)	Absent	Emotions treated as analytical signal
Ontological	Semantic relations, analogies, formal ontology	Limited via embeddings	Full formal threat ontology
Cognitive	Contextual understanding of intent and behavior	Not modeled	Human-like interpretation
Social	Trust, reputation, group influence	Absent	Captures dynamics of social proof
Ethical & Interpretive	Explainability, GDPR/AI Act compliance	Absent	Integrated XAI and legal safeguards

Note – compiled by the authors

As the table shows, the proposed model uniquely integrates emotion, cognition, and ethics into the threat detection process, providing interpretability and transparency at every stage.

Beyond structural comparison, practical scenarios further highlight the architecture's value. Table 7 illustrates three representative use cases.

Table 7 – Application scenarios of the cognitive model

Scenario	Situation	System Actions	Outcome
1. Reaction to phishing	Suspicious message with external link	Sensory layer registers delay; perceptual module detects social engineering markers; social layer assigns low trust	Gentle warning with explanation
2. XAI-based explanation	User requests justification of blocked content	Factors highlighted: aggressive tone, suspicious URL, absence of validation	Builds trust, option to contest decision
3. Privacy preservation	Interaction with emotionally sensitive content	Analysis performed locally; federated learning applied	Maintains confidentiality, GDPR/AI Act compliance

Note – compiled by the authors

These scenarios demonstrate the model's practical effectiveness. The architecture is deployment-ready due to its modular design. Each layer can be executed independently and scaled horizontally across distributed infrastructure. Edge-level inference minimizes latency and privacy risks, while centralized components handle ontology updates and model calibration. This design supports integration into existing social platforms without requiring full system replacement.

To further underline its distinctiveness, the proposed cognitive model surpasses traditional systems across several dimensions. While rule-based models offer only partial interpretability and ML/deep learning (DL) approaches largely lack transparency, the cognitive architecture embeds XAI throughout, ensuring explainability. In contrast to the absence or fragmentation of emotional analysis in earlier systems, it provides full affective integration. Unlike static rules or limited fine-tuning, it enables feedback-driven personalization, adapting dynamically to user behavior. Instead of relying on embeddings alone, it incorporates a formal threat ontology with reasoning. Finally,

where conventional models neglect or only partly address privacy and compliance, the proposed model integrates them fully through local processing, federated learning, and DP, aligning with GDPR and the AI Act.

In addition to conceptual analysis, the study also outlines the planned verification of the prototype, which will include a web interface integrating NLP and CV modules, XAI visualization, and a local profiling database. The prototype will be tested in scenarios such as phishing attacks, manipulative content, and trusted source messages. Evaluation will rely on metrics including classification accuracy, false positive and false negative rates, user trust, computational load, and compliance with GDPR/AI Act requirements. Together, these scenarios provide a coherent picture of the research results, showing that the cognitive architecture not only conceptually surpasses existing approaches but also demonstrates practical adaptability, ethical resilience, and regulatory compliance.

The proposed architecture improves context sensitivity and governance of decisions in social-media security. A unified risk index (see Eq. 1), coupled with trust modeling and adaptive weighting, lowers false alarms and yields interpretable rationales for users and operators. Privacy-preserving mechanisms (on-device processing, federated learning, and DP) enable compliance-oriented deployment while retaining analytic utility. Additionally, the pipeline supports graceful degradation under missing modalities and calibrated mapping from threat levels to responses.

Rule-based frameworks are transparent yet brittle and scale poorly as threat patterns evolve. Conventional ML/DL approaches scale better but require large labeled corpora, are sensitive to label noise and class imbalance, struggle under distribution shifts, and often lack transparency; affective dimensions are rarely modeled [2]. In contrast, our approach (I) treats emotion and behavior as first-class analytical signals [2]; (II) employs a formal threat ontology for analogical reasoning rather than embeddings alone (see Table 2); (III) embeds XAI across the pipeline [3]; and (IV) adopts privacy-by-design principles (on-device inference, federated learning, DP) to satisfy regulatory constraints without centralizing sensitive data [3]. Similar conclusions were reported by [3], who emphasized that multimodal fusion reduces false positives under domain shift; however, unlike their embedding-based pipelines, our model integrates an explicit ontology and end-to-end interpretability, thereby enhancing both robustness and compliance.

At the same time, several limitations must be acknowledged. First, the current validation is limited in scale, and broader experiments with diverse platforms and adversarial scenarios are needed. Second, while our framework accounts for ethical boundaries and explainability, issues of fairness across demographic groups and long-term robustness under adaptive adversarial attacks remain unresolved. Third, reliance on heuristic weight bounds and modular hyperparameters introduces sensitivity that requires further tuning in real-world deployments. Despite these constraints, the architecture demonstrates that combining behavioral, affective, and textual cues with ontology-driven reasoning provides a viable path toward trustworthy and human-centered cybersecurity. These findings not only align with current trends in explainable AI and privacy-preserving design but also highlight practical implications for compliance with GDPR, the EU AI Act, and national digital governance strategies. Large-scale empirical validation with human-in-the-loop explanations represents a critical next step in bridging technical performance with societal trust.

Future work will include large-scale empirical validation across heterogeneous social platforms, incorporating real-world user interaction data and adversarial testing scenarios. Particular emphasis will be placed on evaluating robustness under distribution shift, fairness across demographic groups (e.g., age, gender, language, and cultural background), and longitudinal adaptation of cognitive parameters.

Conclusion.

We presented and substantiated a cognitive protection architecture that integrates multimodal evidence, trust modeling, ontological reasoning, and explainability. The design enables context-aware detection of latent threats, graceful degradation under partial observability,

and compliance-oriented deployment via privacy-preserving learning. It surpasses rule-based rigidity and ML opacity through interpretable adaptive control. Demonstration scenarios indicate feasibility for typical social-media risks (e.g., phishing and manipulative content). Future work will pursue large-scale evaluations, multilingual and cross-cultural adaptation, neuro-symbolic extensions, fairness audits and human-in-the-loop studies of XAI's effects, alongside profiling latency and energy on edge devices. The practical objective is a production-grade prototype with jointly optimized metrics of quality, cost, trust, and compliance.

References

1. Fortinet. (2025). Global Threat Landscape Report. FortiGuard Labs. https://www.fortinet.com/content/dam/fortinet/assets/threat-reports/threat-landscape-report-2025.pdf

2. Oueslati, H., Oujaoura, I., & Benkhelifa, E. (2024). A systematic review of deceptive activity detection on social media: Methods, challenges, and perspectives. Computers & Security, 136, 103589. https://doi.org/10.1016/j.cose.2023.103589

3. Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEe Access, 10, 93104-93139.

4. Wang, Y., He, J., Wang, D., Wang, Q., Wan, B., & Luo, X. (2024). Multimodal transformer with adaptive modality weighting for multimodal sentiment analysis. *Neurocomputing, 572*, 127181. https://doi.org/10.1016/j.neucom.2023.127181

5. Zhang, J., Wu, X., & Huang, C. (2023). AdaMoW: Multimodal sentiment analysis based on adaptive modality-specific weight fusion network. *IEEE Access, 11*, 48410–48420. https://doi.org/10.1109/ACCESS.2023.3276932

6. Wu, D., et al. (2024). Robust multimodal sentiment analysis of image–text pairs by distribution-based feature recovery and fusion. (preprint/PDF).

7. Kumari, S., & Singh, M. P. (2024). A Deep Learning Multimodal Framework for Fake News Detection. *Engineering, Technology & Applied Science Research*, *14*(5), 16527-16533.

8. Verma, A., Moghaddam, V., & Anwar, A. (2022). Data-driven behavioural biometrics for continuous and adaptive user verification using Smartphone and Smartwatch. Sustainability, 14(12), 7362.

9. Yang, P., Liu, N., Liu, X., Shu, Y., Ji, W., Ren, Z., Sheng, J., Yu, M., Yi, R., & Zhang, D. (2024). A multimodal dataset for mixed emotion recognition. *Scientific Data, 11,* 847. https://doi.org/10.1038/s41597-024-03676-4

10. Truong, N. B., Sun, K., Lee, G. M., & Guo, Y. (2021). Privacy preservation in federated learning: An insightful survey from the GDPR perspective. Computers & Security, 110, 102402. https://doi.org/10.1016/j.cose.2021.102402

11. Cummings, R. (2024). Differential privacy in practice: Emerging challenges and future directions. Journal of Privacy and Confidentiality, 14(1), 1–25. https://doi.org/10.29012/jpc.796

12. Laine, H., Lähteenmäki, R., & Saariluoma, P. (2024). Trust, fairness and transparency in AI: A human-centered perspective. AI & Society, 39(2), 567–580. https://doi.org/10.1007/s00146-023-01592-8

ӘЛЕУМЕТТІК МЕДИА ПАЙДАЛАНУШЫЛАРЫН ҚОРҒАУҒА АРНАЛҒАН КОГНИТИВТІК МОДЕЛЬ: АРХИТЕКТУРА ЖӘНЕ ҚАҒИДАТТАР

Аңдатпа. *Әлеуметтік желілер күрделі социо-цифрлық экожүйеге айналып, дезинформация, фишинг және пайдаланушылардың эмоциялық күйіне манипуляциялық ықпал сияқты қатерлермен ұштасуда. Өзектілік жасанды интеллект (ЖИ) пен ауқымды автоматтандыру әсерінен цифрлық тәуекелдердің үдеуімен негізделеді. Зерттеу пәні – әлеуметтік желілерде пайдаланушыны қорғауға арналған когнитивтік архитектура;*

мақсаты – қабылдау, интерпретация, жад, шешім қабылдау және этикалық сүзгі модульдерін біріктіретін үлгіні теориялық-тәжірибелік тұрғыда негіздеу. Міндеттер: (I) мінез-құлықтық талдау, аффективті есептеу және түсіндірілетін жасанды интеллектке қатысты тәсілдерді шолу; (II) мультимодальды деректерді өңдейтін архитектуралық қаңқа құру; (III) компоненттердің рөлдері мен байланыстарын, әсіресе пайдаланушымен өзара әрекеттесу мен ашықтық механизмдерін анықтау; (IV) түсіндірілетін жасанды интеллект (XAI), назар механизмдері және эмоциялық-семантикалық талдауды кіріктіру арқылы жаңалықты көрсету. Әдістер: иерархиялық өңдеу; интегралды қатер индексі $T=\alpha\cdot E+\beta\cdot B+\gamma\cdot S+\delta\cdot C$; Байес сенімін жаңарту; онтологиялық пайымдау; этикалық шектеулі softmax функциясы; кері байланысқа негізделген бейімделу; федеративті оқыту және дифференциалды құпиялық. Негізгі нәтижелер: архитектураның тұтастығы мен жүзеге асырымдылығы көрсетілді; қатер деңгейлерін бейімделген жауаптармен салыстыру картасы ұсынылды; XAI интерфейстері бүкіл цикл бойына енгізілген; жалпы деректерді қорғау регламенті (GDPR) және ЕО-ның жасанды интеллект туралы заңы (EU AI Act) талаптарының сақталуы расталған. Архитектура жабық когнитивтік цикл қағидасымен жұмыс істейді: сенсорлық жинақтау, перцептивтік талдау, эмоциялық-семантикалық аннотация, когнитивтік интерпретация, онтологиялық және әлеуметтік пайымдау, этикалық сүзгі, түсіндірме және кері байланыс. Қорытынды: когнитивтік тереңдік, интерпретацияланғыштық және құпиялық сенімділік пен дербестендіруді арттырады; болашақта мультитілді бейімдеу, нейросимволдық интеграция және human-in-the-loop оқыту бағытында кеңейту ұсынылады. Бағалау өлшемшарттары дәлдік, жалған ескертулер, пайдаланушы сенімі, есептеу шығыны және комплаенспен сипатталды; сенім моделінде репутация, қауымдастық растауы және тарихи сенімділік салмақталады, нәтижелер бейімделген әрекетпен ұштасады. Қолданбалы сценарийлер фишинг пен манипулятивті контентті қамтиды.

Түйін сөздер: *когнитивтік қауіпсіздік архитектурасы, мультимодальды қабылдау, аффективті есептеу, түсіндірілетін жасанды интеллект, қатер онтологиясы, этикалық шешім қабылдау, федеративті оқыту.*

КОГНИТИВНАЯ МОДЕЛЬ ЗАЩИТЫ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ МЕДИА: АРХИТЕКТУРА И ПРИНЦИПЫ

Аннотация. *Социальные сети превратились в сложные социо-цифровые экосистемы, подверженные дезинформации, фишингу и манипулятивному воздействию на эмоциональные состояния пользователей. Актуальность исследования обусловлена ускорением цифровых рисков на фоне широкого применения искусственный интеллект (ИИ) и масштабной автоматизации. Предмет — когнитивная архитектура защиты в социальных сетях; цель — теоретически и практически обосновать модель, интегрирующую модули восприятия, интерпретации, памяти, принятия решений и этической фильтрации. Задачи включают: (I) обзор подходов поведенческого анализа, аффективных вычислений и объяснимого ИИ; (II) разработку архитектурного каркаса для мультимодальной обработки; (III) спецификацию ролей и связей компонентов с акцентом на взаимодействие с пользователем и механизмы прозрачности; (IV) демонстрацию новизны за счёт интеграции объяснимого искусственного интеллекта (XAI), механизмов внимания и эмоционально-семантического анализа. Методы: иерархическая обработка; интегральный индекс риска $T=\alpha\cdot E+\beta\cdot B+\gamma\cdot S+\delta\cdot C$; байесовское обновление доверия; онтологическое рассуждение; softmax, ограниченный этической допустимостью; адаптация по обратной связи; федеративное обучение и дифференциальная приватность. Основные результаты: показана целостность и реализуемость архитектуры; представлено сопоставление уровней угроз с адаптивными ответами; Интерфейсы XAI встроены сквозь весь цикл; соблюдение Общего регламента по защите данных (GDPR) и*

Закон ЕС об ИИ (EU AI Act) подтверждено. Архитектура работает как замкнутый когнитивный цикл: сенсорный сбор, перцептивный анализ, эмоционально-семантическая аннотация, когнитивная интерпретация, онтологическое и социальное рассуждение, этический фильтр, объяснение и обратная связь. Выводы: когнитивная глубина, интерпретируемость и приватность повышают надёжность и персонализацию; дальнейшее развитие предполагает мультиязычную адаптацию, нейросимволическую интеграцию и human-in-the-loop обучение; валидация охватывает фишинг, манипулятивный контент и источники разной доверенности с метриками точности, ложных срабатываний, доверия и комплаенса. Оценивание учитывает вычислительные затраты и управленческие аспекты, а доверие моделируется с учётом репутации, подтверждения сообществом и исторической надёжности источников.

Ключевые слова: *когнитивная архитектура безопасности, мультимодальное восприятие, аффективные вычисления, объяснимый искусственный интеллект, онтология угроз, этическое принятие решений, федеративное обучение.*

Сведение об авторах

Мақатов Ерхан Каиржанович	магистр педагогических наук, лектор, кафедра «Информационно-коммуникационных технологий», Кокшетауский Университет имени Шокана Уалиханова, Кокшетау, Қазақстан, m.yerkhan@list.ru, 87071406577
Abdul Razaque	Доктор компьютерных наук, профессор, факультет компьютерных наук и математики, Университет Сетон Холл, Нью-Джерси, США, arazaque@my.bridgeport.edu
Макатова Асия Еншлесовна	магистр технических наук, лектор, кафедра «Информационно-коммуникационных технологий», Кокшетауский Университет имени Шокана Уалиханова, Кокшетау, Қазақстан, Asiya_kokshe_84@mail.ru

Авторлар туралы мәлімет

Макатов Ерхан Каиржанович	педагогика ғылымдарының магистрі, лектор, "Ақпараттық-коммуникациялық технологиялар" кафедрасы, Шоқан Уәлиханов атындағы Көкшетау университеті, Көкшетау, Казахстан, m.yerkhan@list.ru , 87071406577
Abdul Razaque	Информатика ғылымдарының кандидаты, Сетон Холл Университетінің Информатика Және Математика Кафедрасының Профессоры, НЬЮ-ДЖЕРСИ, АҚШ, arazaque@my.bridgeport.edu
Макатова Асия Еншлесовна	техника ғылымдарының магистрі, лектор, "Ақпараттық-коммуникациялық технологиялар" кафедрасы, Шоқан Уәлиханов атындағы Көкшетау университеті, Көкшетау, Казахстан, Asiya_kokshe_84@mail.ru

Information about the authors

Makatov Yerkhan Kairzhanovich	Master of Pedagogical Sciences, lecturer, Department of Information and Communication Technologies, Shokan Ualikhanov Kokshetau University, Kokshetau, Kazakhstan, m.yerkhan@list.ru , 87071406577
Abdul Razaque	PhD in Computer Science, Professor, Department of Computer Science and Mathematics, Seton Hall University, NJ USA, arazaque@my.bridgeport.edu
Makatova Asia Yenshlesovna	Master of Science, lecturer, Department of Information and Communication Technologies, Shokan Ualikhanov Kokshetau University, Kokshetau, Kazakhstan, Asiya_kokshe_84@mail.ru