International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Algorithmic Approaches to Trust and Safety in Real-Time Social Discovery

Venkata Karunakara Reddy Revunuru*

Senior Software Engineer , Search & AI Platforms Independent Technology Professional ,Formerly,Carvana (USA),Phoenix, Arizona, USA

Email:ven4happiness@gmail.com

Abstract

This article examines algorithmic approaches to trust and safety in real-time social discovery systems under conditions of increasing interaction speed, uncertainty, and accelerated transitions from online communication to offline encounters. The study is based on an analytical synthesis of contemporary empirical, computational, and architectural research in which trust and safety are interpreted not as auxiliary moderation functions, but as systemic properties of algorithmic organization. The analysis draws on studies addressing perceived risk, user value and loyalty, automated detection of harmful content, trust inference in social graphs, robustness under privacy constraints, and the deployment of production-scale safety systems. It is shown that user risk perception affects platform value and loyalty primarily through structural properties of algorithms, including ranking logic, predictability of recommendations, and transparency of decision-making, rather than through isolated negative incidents. The results demonstrate that effective trust and safety mechanisms depend on the integration of trust inference and risk assessment directly into search and ranking processes, as well as on the ability of models to remain robust under partial data availability. Particular attention is given to the trade-off between accuracy, interpretability, and scalability, highlighting the limitations of relying solely on complex models with higher formal accuracy. The study argues that sustainable real-time social discovery requires multi-level algorithmic architectures combining interpretable trust models, adaptive ranking, risk-aware visibility control, and human oversight. The article may be of interest to researchers and practitioners in social computing, recommender systems, and the design of trust- and safety-oriented digital platforms. The main contribution of this study lies in conceptualizing trust and safety as an intrinsic property of algorithmic architectures in real-time social discovery systems, rather than as an auxiliary moderation layer, and in outlining design principles for integrating trust inference and risk-aware ranking into production-scale platforms.

Keywords: trust and safety; real-time social discovery; perceived risk; trust inference; algorithmic ranking; interpretability; social platforms.

Received: 2/10/2026
Accepted: 4/10/2026
Published: 4/17/2026

---

* *Corresponding author.*

1. Introduction

Amid the rapid development of digital social platforms and the shift of interactions toward immediate response modes, processes of dating, contact matching, and trust formation increasingly occur under high uncertainty and accelerated transitions from online communication to personal meetings. This dynamic can be observed in production-grade AI-enhanced distributed search and discovery platforms, where real-time interaction speed, hybrid search logic, and large-scale profile indexing shape user perception of safety and behavioral stability. In operational environments, production-scale social discovery systems typically employ hybrid retrieval pipelines combining lexical BM25 ranking, vector embeddings, and reciprocal rank fusion, enabling low-latency query processing and stable relevance under high concurrent load. These characteristics have been discussed in prior research on production-scale search systems, where ranking predictability and reproducibility can be evaluated beyond controlled experimental settings. These production characteristics suggest that trust perception is influenced not only by moderation policies but also by measurable system properties such as ranking consistency, response predictability, and search result reproducibility. References to specific production platforms are used for illustrative purposes only and do not constitute empirical validation within this study. The analytical conclusions of this work are based exclusively on published academic sources. Such platforms integrate communication, joint activity, and business networking functions, while ranking and retrieval algorithms directly influence perceived risk, user safety, and the continuity of social interactions [8]. Empirical and operational observations indicate that threats in these environments emerge not only from individual malicious actions but also from structural algorithmic characteristics—matching speed, profile visibility logic, ranking opacity, and latency variability—which collectively increase vulnerability and reduce platform trust. Safety is determined not so much by the elimination of isolated incidents as by the sense of risk manageability, recommendation predictability, and the preservation of user autonomy; however, a holistic algorithmic approach to trust and safety in social discovery remains insufficiently developed.

Despite the active development of recommendation algorithms and automated moderation, existing research lacks a unified systemic framework describing trust and safety as intrinsic properties of the algorithmic architecture of real-time social discovery systems. Most works treat Trust & Safety either as a reactive mechanism for harm elimination or as an external constraint imposed upon recommendation algorithms, which fails to explain their impact on perceived platform value and behavioral sustainability.

The aim of the study is to substantiate an architectural approach to designing trust and safety algorithms in real-time social discovery systems, considering production-scale platforms as illustrative reference points, where trust and safety mechanisms are validated through continuous live query streams, operational monitoring, and reproducible ranking behavior under varying load conditions, where Trust & Safety is viewed as an integral property of the algorithmic structure influencing perceived risk, user value, search predictability, and interaction stability. This perspective allows the transition from abstract recommendation models to operational algorithmic loops deployed in live distributed search environments handling hybrid retrieval, ranking fusion, and continuous user interaction streams. To achieve this goal, the work addresses the following tasks:

– identify structural sources of risk and vulnerability arising in social matching and recommendation algorithms;

– systematize the main classes of Trust & Safety algorithmic mechanisms, including harmful content detection, trust calculation, and risk-oriented ranking;

– establish the link between the algorithmic organization of trust and indicators of perceived user value, satisfaction, and loyalty;

– formulate architectural principles for integrating trust and safety into social discovery systems functioning in real-time.

The scientific contribution of the research lies in proposing an interpretation of trust and safety not as an auxiliary layer added to recommendation algorithms, but as an integral property of the social discovery algorithmic architecture. Unlike existing approaches emphasizing isolated classification models or reactive measures post-harm, the proposed approach views trust and safety as the result of coordinated interaction between user display algorithms, trust assessment, message analysis, and the maintenance of user autonomy, forming a managed and predictable social environment. Thus, the novelty lies not only in the conceptual rethinking of Trust & Safety but also in shifting the focus from the accuracy of individual models to the architectural consistency of social matching algorithmic loops.

The research is guided by the assumption that the effectiveness of ensuring trust and safety in real-time social discovery systems is determined by the architectural integration of preventive risk assessment, interpretable trust calculation, and adaptive ranking, rather than by maximizing the accuracy of individual algorithmic models.

The scope of the study is limited to digital social dating and contact matching platforms functioning in immediate response mode, including AI-enhanced distributed search and discovery environments, and does not cover offline social institutions or specialized law enforcement and medical practices. The analysis focuses exclusively on algorithmic design, hybrid search ranking, and trust inference mechanisms embedded into production-level real-time discovery infrastructures. The work does not consider issues of individual psychological diagnostics; safety is interpreted exclusively in the context of algorithmic design and the functioning of real-time social discovery systems.

2. Materials and Methods

The selection and analysis of scientific publications were performed as a targeted analytical synthesis with predefined relevance criteria focused on algorithmic approaches to ensuring trust and safety in real-time social discovery systems. The corpus included peer-reviewed journal and conference papers from 2022–2025 examining risks, trust, and safety in the context of digital social platforms, algorithmic contact matching, automated moderation, and recommendation mechanisms capable of formalization as computational models or architectural solutions. Studies lacking algorithmic or methodological elaboration of trust and safety mechanisms, as well as works focused exclusively on sociological or legal aspects without connection to algorithm functioning, were excluded.ource selection was carried out in two stages. Initial screening by title and abstract aimed to identify correspondence with the social discovery and safety domain, followed by full-text analysis excluding duplicate and conceptually overlapping works. Priority was given to studies containing either quantitative assessments of

algorithm effectiveness or detailed descriptions of computational mechanisms for trust, risk, and resilience to abuse. The final corpus comprised ten sources covering key levels of the investigated framework: user perceptions of safety, harmful content detection algorithms, trust calculation in social graphs, model robustness to privacy constraints, the impact of safety on user behavior, and industrial moderation system architectures.The study by Aljasim and his colleagues [1] identifies user requirements for safety and risk manageability in mobile social discovery systems based on a co-design method. An algorithmic approach to early cyberbullying detection in streaming social environments using fine-tuned language representations is presented by Gutiérrez-Batista and his colleagues [2]. The quantitative link between perceived risk, platform value, and user loyalty is demonstrated using structural modeling in the work of Huang and his colleagues [3]. The formalization of trust and social proximity as calculable ranking indicators in dynamic networks is proposed by Jung and his colleagues [4]. Prediction of trust link strength based on behavioral signals and machine learning methods is investigated by Kridera and Kanavos [5]. Interpretable trust inference under incomplete data and privacy constraints is implemented in a probabilistic graph model by Liu and Wang [6]. Analysis of algorithmic features of fraud and manipulative self-presentation in social platforms is performed by Lokanan [7]. The architecture of an industrial system for detecting unwanted content using taxonomy and active learning is presented by Markov and his colleagues [8]. Human-centric verification of automated safety decisions and discrepancies between model and user evaluations is conducted by Muralikumar and his colleagues [9]. Embedding trust directly into search and navigation algorithms in distributed networks is implemented in the study by Ye and his colleagues [10].

Comparability was ensured through the use of quantitative metrics, model parameters, and formalized descriptions of computational processes presented in the primary sources. A methodological limitation of the study is the exclusive use of published results without independent experimental reproduction, which defines the analytical nature of the conclusions and the boundaries of their interpretation.

The analytical corpus includes both directly relevant studies focused on social discovery platforms and indirectly relevant works originating from adjacent domains such as networked trust inference and service discovery. The inclusion of the latter is based on the transferability of their computational mechanisms; however, such transfer should be interpreted with caution given contextual differences.

To ensure analytical transparency, it should be noted that the present study does not aim to provide a statistically representative systematic review, but rather a concept-driven analytical synthesis focused on identifying recurring algorithmic patterns. The relatively limited size of the corpus reflects the specificity of the research focus on computationally formalized trust and safety mechanisms in real-time systems. Accordingly, the conclusions should be interpreted as analytically grounded but not universally generalizable design propositions.

3. Results

This section presents the results of the analytical comparison of empirical and algorithmic studies, reflecting identified patterns in the influence of the algorithmic organization of trust and safety on perceived risk, user value, and interaction stability. The interpretation of the obtained results and their comparison with existing approaches are presented in the Discussion section.

The analysis suggests that perceived risk functions as a structural parameter shaping user evaluation of digital social discovery platforms.

Changes in risk levels are reflected not in isolated interaction aspects, but in the holistic perception of service utility, including the willingness to continue use and build stable social connections. Quantitative evidence indicates that increased perceived risk is associated with reduced perceived platform value, and this reduction subsequently translates into decreased satisfaction and weakened user loyalty [3].

It is important to note that in the analyzed data, risk does not act as a consequence of singular negative events. It forms as an integral result of the algorithmic organization of social interaction, including contact matching speed, user display logic, and recommendation predictability. In this context, user safety assessment is linked not to the expectation of total protection, but to the ability to control the situation and make informed decisions when interacting with strangers [1]. Consequently, perceived service value appears to depend on the extent to which the algorithmic system supports user agency. Figure 1 shows that perceived risk exerts a negative influence on perceived platform value, while the contribution of perceived value to satisfaction and loyalty exceeds the direct effects of individual benefit and cost components.

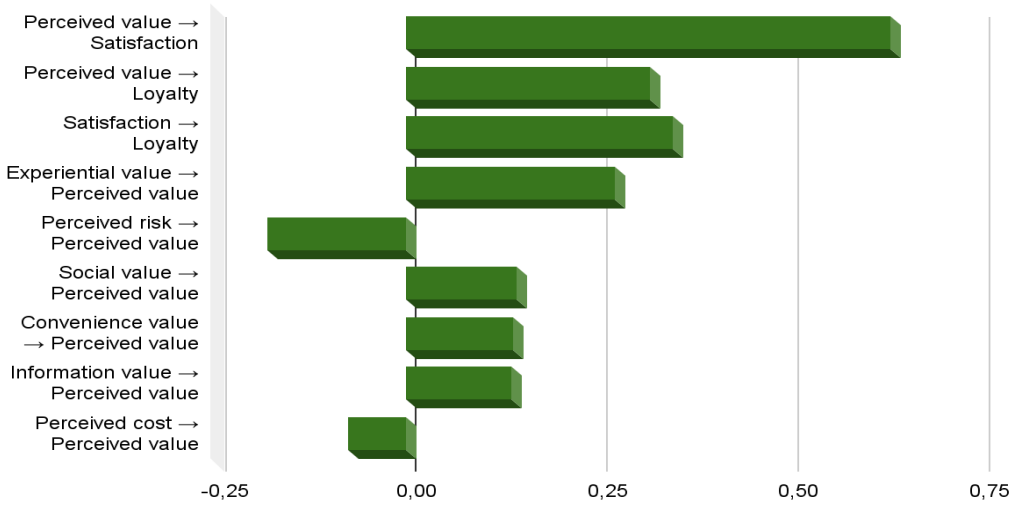

Figure 1: Impact of perceived value and risk factors on user satisfaction and loyalty (SEM, β coefficients)
(Compiled by the author based on source: [3])

The presented structural modeling coefficients show that perceived platform value mediates the influence of risk factors on user satisfaction and loyalty, with the magnitude of this mediating effect exceeding the direct impacts of individual benefit and cost components. Coefficient analysis indicates that perceived value performs the function of the main transmission mechanism between risk factors and final user loyalty, since its influence on satisfaction ($\beta = 0.635$) is nearly double the direct influence on loyalty ($\beta = 0.320$), whereas the contribution of satisfaction to loyalty remains comparable in magnitude ($\beta = 0.350$) [3]. Such a configuration indicates that loyalty is formed primarily indirectly, through accumulated user experience, rather than as an immediate reaction to individual platform properties. Significantly, the negative effect of perceived risk on value ($\beta = -0.182$) is

comparable in absolute magnitude to the total contribution of individual positive benefit components, each making a limited contribution (β in the range 0.138–0.273), indicating model asymmetry: a single source of uncertainty is capable of negating several independent sources of utility. In this context, the statistical insignificance of perceived cost ($\beta = -0.076$) suggests that user value in social discovery is determined not by economic parameters, but by risk structure and interaction predictability, which fundamentally shifts the optimization focus from price mechanisms to algorithmic trust and safety loops.

Comparison with trust modeling results allows for clarifying the mechanism of this effect. Interpretable algorithms for calculating trust links demonstrate that recommendation stability is maintained even under data incompleteness and privacy constraints, provided decision-making logic is consistent [6]. Under these conditions, user value is supported through the reproducibility of algorithmic behavior rather than through aggressive reduction of the interaction space. In production discovery systems, this reproducibility is achieved through deterministic ranking templates, hybrid relevance scoring, and operational validation procedures ensuring that repeated queries under similar conditions yield stable and interpretable results with low ranking variance and consistent response times. Such characteristics suggest that trust perception may be supported by measurable predictability rather than by isolated moderation events. Such predictability directly reduces perceived risk and strengthens user trust, demonstrating the practical alignment between theoretical trust models and live system behavior.

Simultaneously, analysis of user agreement with automated safety decisions shows that trust in the system increases in situations of clearly recognizable risk and decreases in zones of uncertainty [9]. This is consistent with the view that perceived value is formed by the result of filtering or blocking and the transparency of algorithmic logic. Collectively, the recorded effects show that Trust & Safety is built into the structure of user value as a risk regulation mechanism influencing the stability of user behavior [3].

Within the conducted analysis, the effectiveness of algorithmic trust models was evaluated via their ability to robustly identify trust links under conditions of data incompleteness and privacy constraints. The recall of trust relationship identification was considered a key indicator, as omissions of potentially reliable connections lead to user experience degradation and a reduction in the social discovery space [6]. The analyzed models are oriented not toward rigid filtering, but toward preserving system operability under partial unavailability of user features, which corresponds to the real functioning conditions of social platforms [6]. Table 1 presents quantitative indicators of trust inference algorithm robustness under various scenarios of data access restrictions and incompleteness, used to evaluate trust model effectiveness in real-time social discovery systems.

Table 1: Robustness Metrics of Trust Inference under Privacy Constraints (Compiled by the author based on source: [6])

Parameter / Metric	**Observed Value**	**Relevance for Trust & Safety**
Baseline accuracy (naive)	0.5000; 0.5014; 0.5269	Lower reference level
Train/test splits	50–50; 70–30; 80–20; 90–10; 0–40	Validation robustness
Feature removal (privacy)	20; 40; 60; 80	Stability under data loss
Training / inference limits	20 GD epochs; 10 LBP iterations	Practical feasibility
Mean recall (full data)	0.9327	Low miss rate of trust links
Mean recall (reduced data)	0.9409; 27/30 ≥ 0.90	Privacy robustness
Recall (baseline methods)	SVM 0.8239; DT 0.7950; RF 0.8382	Comparison benchmark
Feature contribution (accuracy / F1)	UGC +42.45–50.22%; Profile +14.92–20.87%	Trust signal hierarchy

Analysis of the presented data shows that the interpretable graph-based trust inference model maintains a high level of identification recall even with the sequential removal of significant feature groups. With a 20–80% reduction in available data volume, mean recall values remain higher than those of classical learning algorithms, indicating the structural robustness of the approach under incomplete observation conditions [6]. Significantly, this effect is achieved without aggressive aggregation of trust signals and is ensured through the explicit separation of trustor and trustee contributions within the probabilistic model.

Comparison with alternative trust formalizations shows that relying solely on behavioral or solely on topological features leads to sharp quality degradation as the share of unavailable data increases. Conversely, combining user, content, and propagative signals ensures a smoother decline in metrics and reduces model sensitivity to local information losses [6]. A similar principle is used in ranking models, where trust is treated as a computable value resistant to manipulation.

Comparison with algorithms embedding trust directly into the search process shows that early exclusion of untrusted nodes reduces error accumulation in subsequent stages. This indicates that Trust & Safety effectiveness is determined by the algorithmic loop configuration, not by a separate accuracy metric. In this context, interpretable models offer an additional advantage by allowing signal contributions to be tracked and their influence adjusted without full system retuning.

Thus, the presented results indicate that the effectiveness of algorithmic models for ensuring trust and safety in real-time social discovery is determined by their robustness to data incompleteness, ability to maintain high

identification recall, and architectural integration of trust mechanisms into the contact matching process, rather than by maximizing accuracy on static and fully observable datasets.

4. Discussion

Analysis of algorithmic solutions for ensuring trust and safety in real-time social discovery systems shows that increasing formal model accuracy is not identical to improving decision quality. Under conditions of dynamic user interaction, the algorithm becomes a recognition tool and an active participant in shaping the social environment. Therefore, excessive focus on a single quality indicator inevitably leads to a shift in the balance between accuracy, interpretability, and practical applicability.

High-complexity models demonstrate gains in recognition accuracy, yet this gain comes at the cost of increased opacity of internal mechanisms and complicated control over error sources [7]. In safety tasks, this is of fundamental importance because errors are not distributed evenly: they are concentrated in rare, socially sensitive, and context-laden situations. As a result, increasing average accuracy may be accompanied by reduced decision manageability and increased risks of systematic skew. Figure 2 presents a comparison of the accuracy of algorithmic models of varying complexity applied to risk and violation detection tasks in digital social platforms, allowing for the comparison of accuracy gains against increased model complexity.

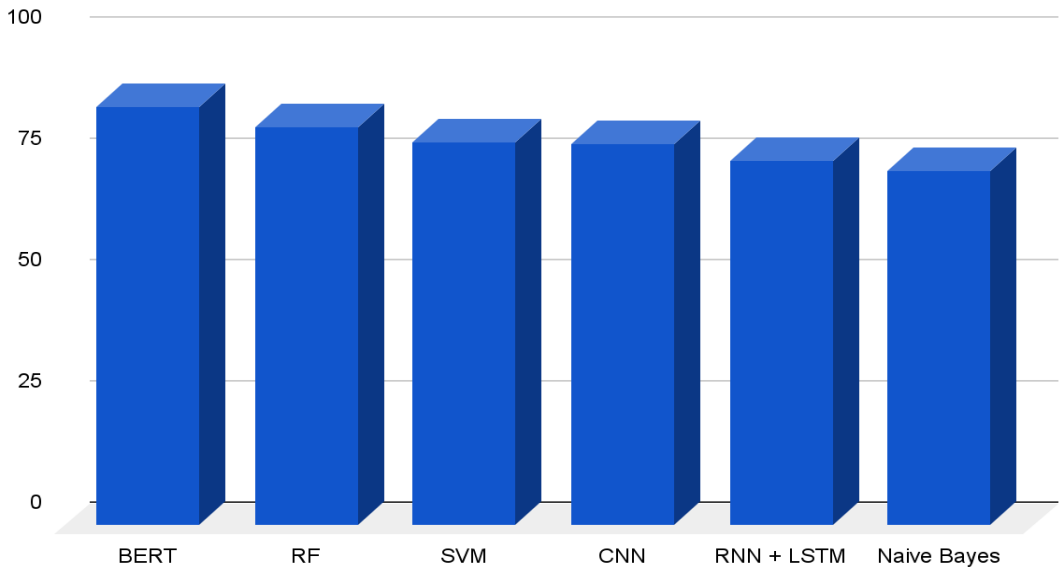

Figure 2: Comparison of the accuracy of machine learning models and deep neural models for safety analytics tasks (Compiled by the author based on source: [7])

Analysis of the diagram data indicates that the advantage of the most complex model is quantitatively expressed but not systemic. The maximum value is achieved by the BERT model (86.36%); however, the gap from the best classical algorithm is limited, amounting to only 4.36 percentage points compared to Random Forest (82%), indicating the absence of a manifold gain given the sharp increase in architecture complexity. Meanwhile, support vector methods demonstrate an accuracy of 79%, which is higher than convolutional neural network indicators

(78.53%) and comparable to results of more resource-intensive solutions, while recurrent models with long-term memory show an even lower value—75.27%. The lowest accuracy is recorded for the Naive Bayes classifier (73%); however, even in this case, the difference with a number of neural network approaches remains moderate. Consequently, the entire range of values falls within a relatively narrow interval from 73% to 86.36%, indicating saturation of the accuracy increase effect and reducing the practical significance of further model complication. In the context of trust and safety tasks, this means that accuracy growth is achieved not through sustainable recognition improvement, but due to isolated architectural effects, whereas the majority of models—regardless of complexity level—demonstrate close values that do not eliminate structural problems of interpretation and uneven recognition quality across classes.

Against this background, interpretable and structured trust models demonstrate a different efficiency profile. Their main advantage is the ability to explicitly manage the ratio between identification recall and precision, and the transparency of decision-making logic [6]. These models allow for local parameter tuning and adaptation to data changes without the need for full algorithmic loop reconfiguration, which is critical for scalable systems with data and computational resource constraints.

More complex models require significant resources for training and updating, reducing their adaptability in rapidly changing social environments. In contrast, less complex and interpretable approaches are easier to integrate into platform infrastructure and allow for maintaining decision stability as the audience grows and interaction scenarios become more complex. In this context, the choice of algorithmic approach is determined not by maximizing accuracy, but by the permissible level of opacity and manageability of the system as a whole.

Analysis of algorithmic solutions for real-time social discovery shows that trust cannot be viewed as an external filter or auxiliary layer connected after recommendation list formation.

Equally significant is the method of interaction between safety algorithms and the user. Effective visibility control and risk assessment loops do not replace user decisions but expand their awareness and situational manageability. When the algorithm acts as a decision-making partner rather than an automated protector, user agency is preserved, and the effect of a false sense of total security is reduced. This is particularly important in scenarios transitioning from online interaction to offline contacts, where the cost of error increases, and trust is formed based on an aggregate of weak signals.

From the perspective of industrial implementation, Trust & Safety architecture cannot be limited to model selection. Production-scale platforms deployed in live distributed search and discovery environments illustrate that operational stability may emerge from the coordination of ranking logic, hybrid retrieval pipelines, anomaly detection routines, and human-in-the-loop verification rather than from isolated model optimization. Cross-team usage involving engineering, site reliability, and data analytics units has been reported in prior studies as an important factor supporting practical adoption beyond experimental or prototype conditions. Cross-team adoption involving engineering, site reliability, and data analytics units further demonstrates that trust and safety mechanisms extend beyond theoretical algorithm design into continuous operational governance. Analysis of architectures of industrially applied Trust & Safety systems shows that solution stability is ensured not by selecting

a single model, but by a combination of a formalized risk taxonomy, controlled data update procedures, active learning, and human participation in the decision-making loop.

Tension arises when scaling such systems. Universal models without contextual calibration demonstrate limited portability between communities and interaction scenarios, intensifying the need for human verification and procedural decision adjustment [9]. Under these conditions, architectural priority shifts from maximizing accuracy to logic reproducibility and change manageability.

Overall, the discussion of results confirms that trust and safety in real-time social discovery systems form as emergent properties of the algorithmic architecture. Attempts to optimize individual components without considering their interaction do not ensure a sustainable effect and may lead to increased hidden risks when scaling systems.

At the same time, the findings should be interpreted in light of several limitations. The study is based on analytical synthesis rather than independent experimental validation, and the selected corpus remains limited in size. In addition, part of the framework relies on the transfer of computational principles from adjacent domains, which requires further empirical verification in real-time social discovery contexts.

5. Conclusion

The conducted analysis shows that trust and safety in real-time social discovery systems cannot be reduced to the accuracy of individual algorithms and cannot be ensured through isolated filtering or moderation measures. They are formed as the result of the architectural organization of the algorithmic loop, in which risk assessment, trust calculation, ranking, and user interaction act continuously and in coordination. The obtained results are consistent with the proposed hypothesis that the effectiveness of ensuring trust and safety is determined by the presence of a multi-level algorithmic architecture combining interpretable trust calculation, preventive risk assessment, and adaptive ranking, rather than by maximizing formal accuracy metrics of individual algorithms.

The main stability factor of such systems is not the maximization of formal quality metrics, but the ability of algorithms to maintain behavior predictability and decision manageability under conditions of data incompleteness, privacy constraints, and high interaction dynamics. In this context, interpretable models and multi-level algorithmic structures possess a fundamental advantage, as they allow for controlling error sources and adapting to changes without user experience degradation.

The results suggest that the perceived value of social platforms is determined by risk structure and algorithmic interaction logic, rather than by economic or interface parameters. Consequently, the effectiveness of social discovery systems directly depends on the extent to which algorithms support user autonomy and decision-making transparency.

Collectively, the findings confirm the hypothesis that Trust & Safety in real-time social discovery must be viewed as an integral property of the algorithmic architecture, rather than as an auxiliary functional layer added to already formed recommendation mechanisms. Future research perspectives involve empirical validation of the proposed

architectural principles on production platform data, as well as studying trust dynamics in hybrid algorithmic loops combining automated and human-centric decision-making mechanisms. Additional interest lies in analyzing the impact of various levels of algorithm transparency on user behavior in scenarios transitioning from online interaction to offline contacts.

References

[1]. Aljasim, H. K., & Zytko, D. (2022). Foregrounding women's safety in mobile social matching and dating apps: A participatory design study. Proceedings of the ACM on Human-Computer Interaction, 7(GROUP), Article 9, 1–25. https://doi.org/10.1145/3567559

[2]. Gutiérrez-Batista, K., Gómez-Sánchez, J., & Fernandez-Basso, C. (2024). Improving automatic cyberbullying detection in social network environments by fine-tuning a pre-trained sentence transformer language model. Social Network Analysis and Mining, 14, 136. https://doi.org/10.1007/s13278-024-01291-0

[3]. Huang, Q., Zhang, R., Lee, H., Xu, H., & Pan, Y. (2024). A study on customer behavior in online dating platforms: Analyzing the impact of perceived value on enhancing customer loyalty. Behavioral Sciences, 14(10), 973. https://doi.org/10.3390/bs14100973

[4]. Jung, J., & Weon, I. (2025). The social side of Internet of Things: Introducing trust-augmented social strengths for IoT service composition. Sensors, 25(15), 4794. https://doi.org/10.3390/s25154794

[5]. Kridera, S., & Kanavos, A. (2024). Exploring trust dynamics in online social networks: A social network analysis perspective. Mathematical and Computational Applications, 29(3), 37. https://doi.org/10.3390/mca29030037

[6]. Liu, Y., & Wang, B. (2022). User trust inference in online social networks: A message passing perspective. Applied Sciences, 12(10), 5186. https://doi.org/10.3390/app12105186

[7]. Lokanan, M. E. (2023). The Tinder Swindler: Analyzing public sentiments of romance fraud using machine learning and artificial intelligence. Journal of Economic Criminology, 2, 100023. https://doi.org/10.1016/j.jeconc.2023.100023

[8]. Markov, T., Zhang, C., Agarwal, S., Eloundou, T., Lee, T., Adler, S., Jiang, A., & Weng, L. (2023). A holistic approach to undesired content detection in the real world [Conference presentation]. arXiv. https://doi.org/10.48550/arXiv.2208.03274

[9]. Muralikumar, M. D., Yang, Y. S., & McDonald, D. W. (2023). A human-centered evaluation of a toxicity detection API: Testing transferability and unpacking latent attributes. ACM Transactions on Social Computing, 6(1–2), Article 4, 1–38. https://doi.org/10.1145/3582568

[10]. Ye, Z., Sheng, H., & Zou, H. (2025). Trusted web service discovery based on a swarm intelligence algorithm. Mathematics, 13(9), 1402. https://doi.org/10.3390/math13091402