International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

AI, Data Privacy, Security, Bias and Trust: A Systematic Review of Algorithmic Fairness and Robustness

Himanshu Dey
V Sidhartha, Syed Saoud, Sreeganesh Shelmohkar

ABSTRACT

This research paper examines the existing body of work related to the issues of algorithmic fairness, privacy, security, and model trustworthiness in adaptive AI algorithms. This research will discuss the issues related to fairness and privacy problems in learning models facing concept drift and continual learning problems. The key takeaways of the research include: (a) The definition of fairness (equalized opportunity) requires that true positive rates should be equal for each group which implies that fairness is context-specific; (b) Methods for achieving privacy (differential privacy) ensure privacy of individuals but add noise which decreases the accuracy of predictions; (c) Security mechanisms (adversarial training) may lead to increased robustness but lower nominal accuracy; (d) Continual learning mechanisms need to tackle the problem of catastrophic forgetting to keep being fair and useful. In each area of study there are some trade-offs between different approaches and the solution is focused on solving only one problem at a time. There is a gap between works which integrate fairness, privacy, and robustness.

Keywords: Artificial Intelligence, Algorithmic Bias, Fairness, Data Privacy, AI Security, Trustworthy AI, Explainable AI, Systematic Review, Concept Drift.

1. INTRODUCTION

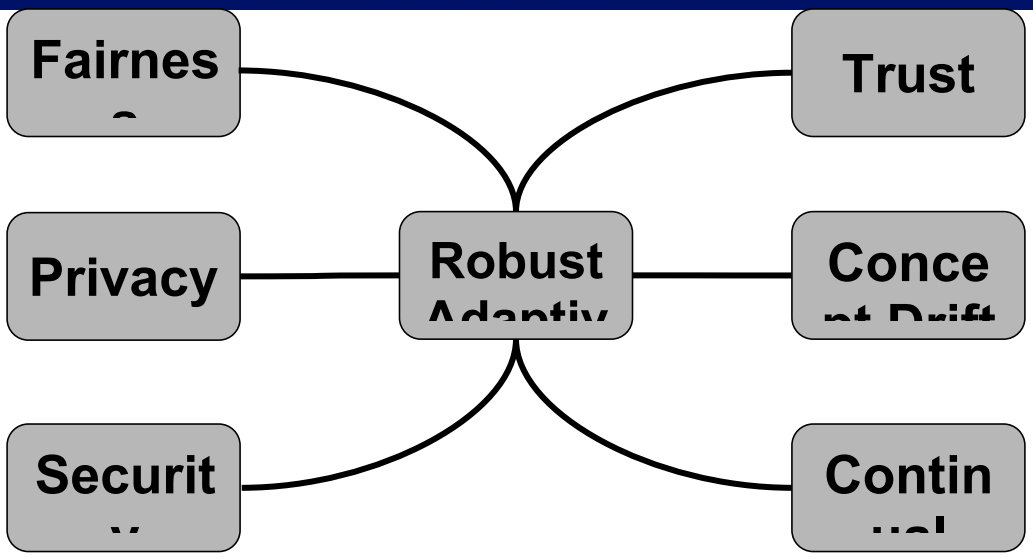
Machine learning models are now used in high-stakes fields, which leads to various issues related to fairness, privacy, security, and the reliability of models. The reason for those issues is that models trained on past data might be subject to social prejudices since “our historical examples…are imbued with our historical prejudices against certain social groups.” Consequently, a statistical rule that aims to maximize the overall accuracy might lead to unfair results. In order to solve the problem, scientists developed formal definitions of fairness. Thus, the Equality of Opportunity postulates that equally qualified people should be classified into positive categories with equal probability. In its turn, Demographic Parity stipulates that all social groups should get

the same share of positive outcomes regardless of their real labels (see Hardt et al. who prove that this definition heavily reduces utility).

The problem of data privacy is another critical concern. The concept of differential privacy (DP) is quite popular; the idea behind DP is the addition of noise to any query in order to make it impossible for the observer to determine the inclusion of a particular person's data. In DP, there is a basic tradeoff between the two properties: higher privacy (more noise) usually results in decreased accuracy. Adversarial security only complicates matters further. One of the well-known problems of neural networks is their vulnerability to adversarial examples: small changes in input cause misclassification with high confidence. In a similar manner, poisoning introduces malicious points into the training set in order to manipulate the behavior of the model. Biggio et al. (2012) described gradient-based attacks against support-vector machines, while other surveys cover the last decade of such attacks.

All the issues above relate to model adaptation. In practice, the data distribution changes with time (concept drift), and thus models need to adapt continuously by learning from new data. Drift left unmitigated results in model degradation. The field of continual learning (Parisi et al. 2019) explores how neural networks can be adapted to new tasks without suffering catastrophic forgetting. However, adaptation introduces new issues – how can we be sure that data change is genuine and not adversarially induced? Can adaptation happen without compromising privacy or fairness?

The present paper carries out an organized literature review with regard to the above questions. In Section 2, we review the basic literature on the issues of fairness, privacy, and security. In Section 3, we describe our searching procedure (as per the guidelines provided by PRISMA 2020). In Sections 4–7, we organize the results obtained from literature: we compare different methods in terms of their abilities to deal with drifting input data, adversarial attacks, and knowledge persistence while implementing fairness and privacy. Instead of giving a single recipe, we try to highlight the main criteria for developing a reliable AI system. First of all, it is the ability to distinguish between natural drift and malicious data, maintaining the correctness of the update data and balancing model adaptiveness with its stability.



2. REVIEW OF LITERATURE

2.1. Algorithmic Bias and Fairness Metrics

Algorithmic biases can occur through machine learning algorithms being trained on data where social biases, incompleteness, inaccuracies, or label biases exist. Even if the algorithm does not use protected characteristics such as gender, race, or age among its inputs, other variables might serve as proxies for those characteristics and allow for reproducing existing biases. According to Barocas et al. (2023), mathematical criteria of fairness are impossible to apply to machine learning algorithms because each definition implies different assumptions about the definition of a fair result. Consequently, selecting an appropriate fairness criterion should take into account the context of a particular decision and the nature of the data used.

Various measures have been used to measure how these disparities exist among the different demographic groups. For instance, Hardt et al. (2016) came up with equality of opportunity, in which individuals who meet the set criteria in different demographic groups are supposed to have equal true-positive rates. In their more general framework, they provide equalized odds, where true-positive rates and false-positive rates are required to be equal in different demographic groups. The above measures of fairness differ from demographic parity in that demographic parity

concentrates on the equality of positive prediction or selection rate rather than the equality of prediction accuracies. Hence, it is possible for a certain fairness measure to be met yet the other not.

It is becoming increasingly difficult to select and assess different fairness metrics as the underlying distribution of the data shifts. A model which was once conforming to a certain fairness metric may begin to demonstrate shifts in its performance for different groups due to shifts in the underlying demographic, behaviors, or data collection practices. This is why fairness cannot be seen as an attribute that can only be determined at the stage of model design.

2.2. Data Privacy and Preservation Techniques

Privacy-aware machine learning seeks to minimize the extent to which private information is revealed during the process of data gathering, training of the machine learning model, and deployment of the machine learning model. An example of a privacy-aware machine learning algorithm is federated learning whereby the training data stays on decentralized devices or company servers and only model updates are aggregated to form a common model. This technique makes it unnecessary for the training data to be transmitted to a centralized server, especially when the training data is from different sources (McMahan et al., 2017).

Decentralization of the training data alone, however, is not sufficient for ensuring privacy. Updates to the model, which occur during federated learning, carry information about the data used for training and may be subject to attacks such as membership inference and other information leakages. Thus, when considering privacy, one needs not only to take into account whether the raw data remains decentralized but also how much information can be revealed by the model parameters, gradients, updates, and its outputs (Shokri et al., 2017). This is significant as the system may avoid access to the raw data, but may reveal sensitive information nevertheless.

Differential privacy can be considered a more systematic approach to mitigate leakage of information through adding controlled noise to the data analysis or machine learning process. The privacy guarantee of differential privacy is usually defined using a privacy parameter that governs

the degree of privacy protection achieved. Stronger privacy protection usually means larger amounts of noise needed and possibly reduced utility of the model (Dwork & Roth, 2014). As a result, in privacy-preserving AI, there is a trade-off between the privacy protection that is achieved, the information that the adversary has access to, and the performance of the model.

2.3. AI Security: Adversarial Threats and Data Poisoning

Not only is AI security concerned with the safeguarding of the infrastructure upon which the model relies but it includes the securing of the learning process, the model, and the output. It is possible for machine learning models to suffer from adversarial attacks where crafted input can cause the model to give inaccurate predictions. Goodfellow et al. (2015) have shown that small amounts of crafted input can make a neural network mis-classify the inputs with a lot of confidence. The fact is that high prediction performance on regular test data does not ensure that the model will be resistant to attacks.

Risks related to the security can also be identified at the stage of training. The data poisoning attacks are those cases when training samples are added or manipulated for the purpose of changing the behavior of the trained model. Biggio et al. (2012) showed that such poisoning samples increased the test error of support vector machines. Thus, it can be concluded that security threats are associated not only with the behavior of a model at the stage of adversarial inputs but also with the security of training data, especially in the case of systems that use data collected from several different sources (Biggio et al., 2012; Biggio & Roli, 2018).

There is another way in which the security and privacy threats can be inter-related and that involves information leakage from the outputs of the models. In particular, according to Shokri et al. (2017), there have been attacks related to membership inference where an attacker tries to determine whether a specific example belongs to the training set used for the model. The outcomes showed that under some circumstances, there is a possibility of revealing such information based on prediction outputs. As a result, an AI application can be perfectly secure in terms of its operations and still leak sensitive data via its interface.

2.4. Trustworthy AI, Continual Learning, and Explainability

Trustworthy AI entails technical, organizational, and human aspects relating to the development and implementation of AI technologies. Critical aspects of trustworthiness in AI include transparency, explainability, accountability, fairness, and the ability to detect and mitigate risks. Explainability entails the capability of giving understandable information on how the decision was made through a model or the way a decision was aided by the model. Transparency entails disclosing relevant information regarding the AI system, including the objective, limitations, and information used. Accountability, on the other hand, entails holding someone responsible for a decision made or system behavior and actions to be taken after that. According to Mittelstadt (2019), ethical principles at the high level alone are not sufficient as it is difficult to translate them into mechanisms.

The issue of explainability is especially significant when AI systems are applied to make decisions concerning individuals or organizations. Yet, offering an explanation alone does not ensure that a system is accurate, fair, or trustworthy. Thus, the trustworthiness of AI models involves analyzing both their performance and explanations, as well as possible limitations, risks, and corporate responsibilities for deploying a particular AI solution. This approach is necessary since AI systems may be technically accurate but still result in unexplainable decisions and be difficult to monitor.

Another factor which adds to the difficulty of creating trustable AI solutions is the possibility of adapting them based on new data and changes in the environment. The authors of Parise et al. (2019) define continual learning and lifelong learning as the ability of neural networks to continuously learn new things. At the same time, the authors discuss the issue of catastrophic forgetting which concerns the interference between new and old knowledge. In the case of dynamic AI systems, this means that along with evaluating the adaptation to new environment, one has to analyze how previous properties such as accuracy, performance, privacy, and security have been affected.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

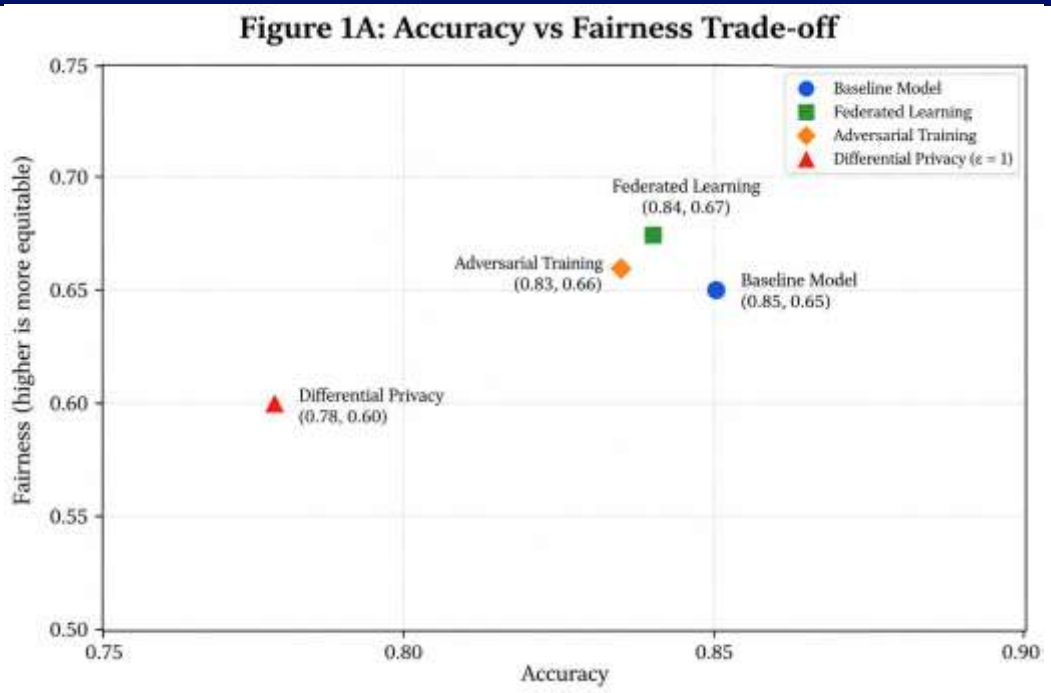


Figure 1A. Accuracy vs fairness trade-off for different approaches. Each point shows the model's accuracy (x-axis) and fairness score (y-axis, e.g., equal opportunity ratio, higher value indicates fairer outcomes).

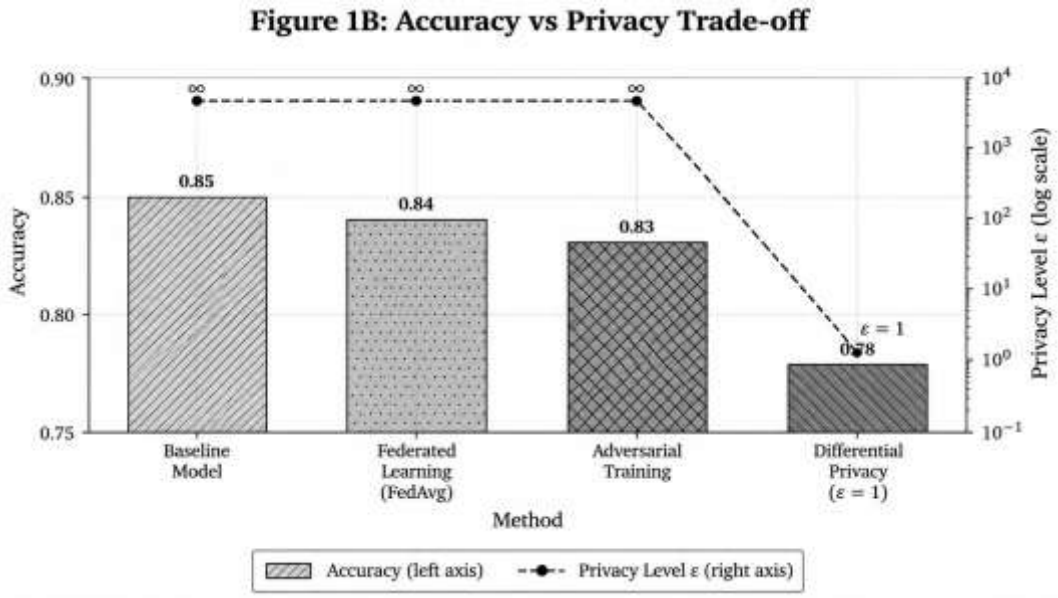


Figure 1B. Accuracy vs privacy trade-off for different approaches. Bars indicate model accuracy (left y-axis) and the dashed line shows the privacy level ε (right y-axis, log scale; smaller ε indicates stronger privacy). Baseline, federated learning, and adversarial training do not provide formal differential privacy (ε = ∞), while the differential privacy approach uses epsilon = 1.

3. RESEARCH METHODOLOGY

In this work, a systematic review of literature method is employed to evaluate research on algorithmic bias and fairness, privacy of data, AI security, and trustworthy artificial intelligence, especially in terms of concept drift and continual learning. The review process had been done according to predefined thematic areas and inclusion/exclusion criteria to ensure a consistent basis for identifying the literature that was considered relevant. The selected papers were then analyzed based on their research questions, methodology, evaluation metrics, applications, results, and limitations.

3.1. Search Strategy and Information Sources

The literature search process involved the use of four academic databases, including Google Scholar, IEEE Xplore, ACM Digital Library, and Scopus. The search strategy was developed through six thematic categories, which include: (1) Algorithmic bias and fairness, (2) Data privacy and privacy preserving machine learning, (3) AI security and adversarial attacks, (4) Trustworthy and explainable AI, (5) Concept drift and continual learning, and (6) Relationships between these thematic areas.

The search terms used for literature searches were chosen based on the concepts in the thematic categories described above and used Boolean operators where applicable within the specific database. The search targeted literature from 2014 to 2026; however, literature before 2014 was included only where it contributed significantly to the topic discussed.

The literature that was gathered was reviewed for relevancy in accordance with the criteria set out in Section 3.2. The relevant literature was subsequently classified based on its main research area, methodology, evaluation criteria, application area, findings, and limitations. This was done in an attempt to establish commonalities in findings, different methodologies, and gaps in research.

3.2. Eligibility Criteria

The studies were selected by taking into consideration certain inclusion and exclusion criteria, which are related to the relevancy of the topic, type of publication, the technical contribution of the study, and the period of publication. The inclusion and exclusion criteria were established in order to keep the relevancy to the topic of interest. The inclusion and exclusion criteria are summarized in Table 1.

Criterion	Inclusion Criteria	Exclusion Criteria
Topic Relevance	Studies addressing AI fairness, algorithmic bias, privacy, security, trustworthiness, concept drift, or continual learning.	Studies unrelated to the defined research themes or focused on unrelated applications.
Publication Type	Peer-reviewed journal articles, conference proceedings, and comprehensive scholarly reviews.	Editorials, short web content, non-scholarly material, and preprints without peer review.
Technical contribution	Empirical evaluations, methodological contributions, defensive architectures, theoretical analyses, or mathematical formulations relevant to the review objectives.	High-level discussions or speculative articles without sufficient technical or scholarly contribution.
Publication period	Studies published from 2014 to 2026, together with earlier foundational works directly relevant to the research topic.	Studies outside the defined period that do not provide a foundational contribution relevant to the review.

Table 1. Inclusion and Exclusion Criteria for Systematic Literature Review.

The chosen studies were analyzed comparatively to find out how the existing approaches deal with issues of fairness, privacy, security, and adaptability and how these areas might overlap or be underdeveloped in their integration.

4. OBJECTIVES OF THE REVIEW

The main purpose of this review is to examine how concept drift, continual learning, adversarial manipulation, and catastrophic forgetting jointly affect the security and operational robustness of ML-based systems. The review therefore focuses on the following objectives:

- To examine the main factors contributing to algorithmic bias and explore existing techniques for mitigating discriminatory outcomes in dynamic environments.
- To classify security threats to continual-learning methods, particularly data poisoning, backdoor attacks, and membership inference.
- To analyze how privacy-preserving techniques, such as differential privacy and federated learning, interact with bias-mitigation algorithms and whether these approaches work together or create conflicts.
- To design a security-aware evaluation framework that integrates drift detection, model adaptation, update-data security, and knowledge retention.

5. SYNTHESIS OF REVIEWED LITERATURE

The reviewed literature reveals that fairness, privacy, security, and adaptability are interconnected but considered separately from each other in different studies. Fairness algorithms are concerned mainly with disparity measurement between different groups, and privacy-related research centers on the reduction of exposure of sensitive data. Security algorithms protect machine learning models from being tampered with by adversaries, while continual learning algorithms are associated with the adaptation of models to new data without forgetting the old knowledge that was learned before. These areas demonstrate that improving one property does not entail improvement of the others.

5.1. Fairness and Algorithmic Bias

Studies of algorithmic fairness show that fairness cannot be characterized using a single universal notion. According to Barocas et al. (2023), what is important here is which type of fairness is suitable in each case and whether it depends on the context, data, and error types' effects. Hardt et al. (2016) developed the concept of Equality of Opportunity as a criterion for equal chances of having positive outcomes among equally qualified people of various demographic backgrounds.

Another fact highlighted by these papers is that applying different notions of fairness results in various assessments of the same algorithm. The three discussed measures of fairness (demographic parity, Equality of Opportunity, and Equalized Odds) represent different dimensions of group-level disparity, which means that they assume different kinds of fairness in terms of their implications. Therefore, relying on just one measure for evaluation is inappropriate.

Moreover, changes in data distributions make fairness assessment even more difficult. As shown by studies, an algorithm that is compliant with some fairness criterion in some situation can fail to do so when there are changes in the population's features, behaviors, or something else that can affect data.

5.2. Privacy and Model Utility

In the context of literature on privacy preserving AI, there is a clear conflict between the necessity to protect information and the usefulness of the model. The idea of differential privacy, introduced mathematically by Dwork and Roth (2014), is based on the limitation of the effect that any individual data point could have on the output of computations. Federated learning can be considered another way of training the models that can allow decentralized computation of the model without transferring all training data to the centralized place (McMahan et al., 2017).

However, decentralized computation is not sufficient for ensuring full privacy of the data. Information about the training data can be revealed through model updates and output. Moreover, membership inference is the proof that machine-learning models could leak information on whether any record was used for training (Shokri et al., 2017). Thus, privacy is to be understood in the context of a particular threat model, privacy tool, and available information.

Privacy preserving techniques could affect fairness analysis as well. For example, if some sensitive attributes and other information necessary for auditing is not available or intentionally protected, then it could be difficult to establish whether the model complies with certain fairness criterion.

5.3. Security and Robustness

Adversarial machine learning research reveals that the protection of AI systems involves more than securing their infrastructure. Specifically, Goodfellow et al. (2015) found that strategically

designed input manipulations could cause neural networks to make incorrect predictions while Biggio et al. (2012) found that manipulating training samples could affect the operation of learning algorithms. Both of these results indicate that inputs to a model and its training or update data could be turned into an attack surface.

An adaptive AI system should therefore take into account not only its accuracy at processing clean test data but also its performance with previously unknown attack patterns, modified training data, and changed inputs after being deployed since this becomes crucial in case of continual learning where inputs could impact updates of a model.

Finally, the privacy of AI and the security of its data are closely connected since the former can leak data-related information in case of membership inference attacks (Shokri et al., 2017). In other words, an AI system must be protected with respect to both confidentiality of its data and robustness of its learning and inference processes.

5.4. Continual Learning, Concept Drift, and Knowledge Retention

Research on concept drift and continual learning identifies another problem, where AI systems in real-world application cannot always take for granted that the data distribution will be stationary. Concept drift is referred to as the change in the underlying data-generating process by Gama et al. (2014), leading to the decrease of efficiency of already trained models. Research on continual learning also focuses on difficulties associated with learning new information while preserving the acquired one (Parisi et al., 2019).

This brings up additional problems of fairness, privacy, and security. Changes in the behavior of a model due to new data can lead to the appearance of differences between the groups which were observed earlier. Meanwhile, new data can carry sensitive information or be intentionally poisoned. Adaptation of the model is thus associated with new risks, despite the initial evaluation of the model on criteria of fairness, privacy, and security.

It is thus clear that adaptability should not be considered as a separate criterion of model performance. Evaluation of a continually learning system should include the test of whether the

important information is preserved and whether the model still meets the necessary criteria of fairness, privacy, and security after adaptation.

5.5. Cross-Domain Synthesis

In all the discussed areas of research, there is a recurring issue in which the properties of AI such as accuracy, fairness, privacy, security, and adaptability can conflict with each other. The improvement of privacy can impact the performance of the model or make it difficult to check for fairness. Adaptation to evolving data can boost the predictive performance while changing fairness properties and adding security risks. In the same way, security measures can decrease the adaptability of the model and add computational complexity.

Thus, the reviewed literature supports the need to move from the assessment of AI systems using one performance metric to multi-dimensional evaluation. In particular, in case of evolving and adversarial environments, the performance of AI systems should be assessed based on their predictive performance, fairness, privacy, security and robustness, and adaptability without losing the previous knowledge.

Overall, the reviewed literature offers sufficient research foundations in terms of the discussed topics individually, while comparatively little research is done on the joint properties of the systems in changing environments. Such fragmentation is a research gap which is addressed in the following chapters.

6. RESEARCH GAPS AND FUTURE DIRECTIONS

The reviewed literature lays strong groundwork for fairness, privacy preservation, AI security, and continual learning, yet each of these topics tends to be researched separately. Some of the identified gaps restrict the development and assessment of AI models that work in evolving environments.

6.1. Integrated Evaluation of Fairness, Privacy, and Security

One of the key problems is the lack of consideration of multiple properties – namely, fairness, privacy, and security – within one evaluation approach. Current research usually considers only one property at a time. At the same time, privacy-preserving approaches may influence the amount of information needed for auditing of the model fairness, and security and continual learning may affect the behavior of the model in different demographics. Therefore, future research should focus on the approaches to evaluation of these properties simultaneously rather than separately.

6.2. Fairness Under Changing Data Distributions

Fairness is usually assessed on the static datasets and fixed distributions of features. In practice, however, characteristics of the population, user behavior, and other properties of data may evolve. It may lead to different outcomes of fairness assessment even with no changes in the model itself. Therefore, further research should assess how such properties as concept drift affect the process of evaluation of fairness.

6.3. Security of Continual-Learning Systems

However, continual-learning architectures provide additional attack surfaces since the incoming data might impact future model states. Distinguishing legitimate concept drift from manipulation of incoming data represents one of the critical problems. Further research is needed in developing secure updating procedures that will include detecting the drift, validating data, detecting anomalies, protecting against poisoning and backdoor attacks, and enabling adaptation to these attacks.

6.4. Knowledge Retention and Robust Adaptation

Adaptation to the new information is the core of continual learning. However, continual learning involves retaining already obtained knowledge as well. Forgetting of previous knowledge leads to catastrophic forgetting and affects the performance of the model in the older domain. On the other hand, too strong retention of the knowledge prevents adaptation. Therefore, further research is needed in this area.

6.5. Privacy-Preserving and Auditable AI

The problem of privacy can become challenging for auditing a model when sensitive attributes or even individuals' information is needed to measure some of its properties like fairness or presence of backdoors.

6.6. Standardized Evaluation Under Realistic Conditions

One such gap relates to the lack of use of evaluation frameworks that at once take into consideration time variation, adversarial scenarios, privacy issues, and fair outcomes. Fixed or randomly partitioned datasets might not necessarily reflect the actual situations after the implementation phase. Future work would benefit from taking into consideration dynamic evaluations where predictive performance, fairness, privacy, adversarial situations, security, and adaptability can be considered.

6.7. Future Research Direction

In general, future work in this field should tend toward a multi-criteria approach for evaluating AI systems, and not just focus on improving prediction accuracy. An ideal evaluation method should be one that evaluates predictive performance along with relevant considerations for fairness, privacy issues, adversarial scenarios, and the system’s adaptability without significant loss of prior learned data.

7. DISCUSSION

The synthesis of the discussed literature suggests that fairness, privacy, security, and adaptability are linked aspects of AI algorithms, which are nevertheless assessed independently. However, such an approach becomes increasingly challenging when it comes to continuously operating systems since any changes in the nature of data, model parameters, and deployment environment affect multiple aspects at once.

An important aspect to notice here is the inadequacy of assessing AI systems deployed in dynamic environment by their performance alone. A model can demonstrate high accuracy without ensuring fairness, violating privacy or being vulnerable to new attacks. The research on fairness has shown that different criteria lead to different results for the same model. At the same time, the literature

on privacy and security proves that different types of threat need to be considered in order to protect data and models.

The trade-off between privacy and fairness is especially relevant to note. Privacy-enhancing methods can hinder access to sensitive data that could be helpful in auditing disparities in demographics. On the other hand, the removal or protection of sensitive attributes does not guarantee the absence of biased results since other attributes might serve as a proxy for such. It means that privacy protection and bias detection are not mutually exclusive tasks; future systems require ways to safeguard sensitive data but allow for discrimination detection at the same time.

Concept drift vs. deliberate manipulation of data is another problem that researchers emphasize as being crucial. The former phenomenon is expected in many real-life environments, while the latter one can be used for poisoning the model in purposeful ways. Treating all types of distributional changes in data as innocent may lead to constant exposure of the adaptive system to malicious changes, while treating all changes as malicious may deprive the system of the opportunity to adapt to changing conditions. This problem necessitates the development of adaptive security measures.

An additional aspect to consider in this case is adaptability versus knowledge retention. While continual learning enables models to react to new information, overly strong adaptation leads to forgetting of already learned knowledge. In case there are any system properties (such as fairness or security) established during previous phases of learning, they might be affected by modifications of the model. Consequently, the effectiveness of continual learning should be assessed by both the capability to learn new information and retain the required previous knowledge.

Based on all discussed issues, it can be said that trustworthy AI is essentially an ongoing process of evaluation. The properties of fairness, privacy, security, and robustness may vary with time as data and models change.

Nonetheless, the examined literature base remains highly fragmented in nature. Robust grounds have been laid for separate topics such as fairness metrics, differential privacy, adversarial machine learning, concept drift, and continual learning; yet, little attention has been paid to the relations of the mentioned aspects within the same evaluation environment.

On the whole, the discussion reveals that any future trustworthy AI system will need to go beyond the traditional goal of improving predictive accuracy alone. The more holistic approach is to take into account accuracy, fairness, privacy, security, adaptability, and knowledge persistence all at once, acknowledging that gains in one might come at a cost in another.

8. SUGGESTIONS

Based on the findings and discussion, several practical recommendations can be made for researchers and developers working with AI systems in dynamic environments.

- **Security-aware drift detection:** Drift-detection mechanisms should distinguish between legitimate changes in data distributions and potentially malicious changes. This can help prevent adversarially manipulated data from being incorporated into continual-learning updates.
- **Validation and monitoring of model updates:** Incoming data should be evaluated before being incorporated into model updates. Data validation, anomaly detection, and appropriate defenses against poisoning and backdoor attacks should be incorporated into the update process, particularly for systems that continuously learn from incoming data.
- **Multi-dimensional evaluation:** AI systems should not be evaluated using predictive performance alone. Evaluation should consider relevant fairness measures, privacy guarantees such as differential-privacy parameters, adversarial robustness, and the ability of the model to adapt while retaining important previously learned knowledge.
- **Continuous monitoring after deployment:** Fairness, privacy, security, and model performance should be reassessed as data distributions and operating conditions change. This is particularly important for continually learning systems because adaptation can alter model behavior over time.
- **Cross-disciplinary research:** Researchers working on algorithmic fairness, privacy-preserving machine learning, adversarial machine learning, and continual learning should collaborate more closely. Integrating these perspectives can help address interactions between properties that are often studied independently.

These recommendations emphasize that trustworthy AI should be approached as an ongoing process of evaluation and monitoring rather than as a one-time assessment performed before deployment.

9. CONCLUSION

In this review, the linkages between fairness of algorithms, privacy of data, security of the AI systems, and trustworthiness of artificial intelligence were analyzed, especially in relation to the systems functioning under dynamically changing data distributions and in conditions of continual learning.

The analysis reveals the strong interconnection of these properties; however, these properties are typically investigated separately from each other.

As follows from the literature synthesis, the improvement of any of these characteristics does not lead to improvement of the other ones. Data privacy techniques can create difficulties in the audit of fairness of the models, model adaptation can influence the previously discovered properties of fairness and security, and continual learning requires balancing adaptation to new data and maintaining previous knowledge. The example of the security issues illustrates the dependency of the security of AI system not only on its surrounding environment but also on data, learning process, and model itself.

Thus, the most important conclusion made from the literature is the necessity of multi-dimensional and continual assessment of AI systems. Predictive performance should be evaluated together with the proper fairness, privacy, security, and adaptation without loss of knowledge.

Even though there is a significant body of literature in each of these areas separately, it lacks coherence and cohesion when taken together as a whole. Future studies should concentrate their efforts on developing evaluation methodologies and realistic dynamic environments where all of these issues could be tested at once.

In essence, building trustworthy AI goes well beyond just striving for maximum model accuracy. When designing AI that is meant to operate for an extended period of time, more criteria than just model accuracy need to be considered.

10. REFERENCES

[1] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[2] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012, pp. 1467–1474.

[3] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[4] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, nos. 3–4, pp. 211–407, 2014.

[5] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, Art. 44, 2014.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[7] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, vol. 29, pp. 3315–3323, 2016.

[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54, pp. 1273–1282, 2017.

[9] B. D. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019.

[10] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[11] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 3–18.