Towards User-Centered Metrics for Trustworthy AI in Immersive Cyberspace

Pengyuan Zhou[1*] , **Benjamin Finley**[2] , **Lik-Hang Lee**[3] , **Yong Liao**[1] , **Haiyong Xie**[1] , **Pan Hui**[2,4] ,

[1]University of Science and Technology of China, [2]University of Helsinki, [3]Korea Advanced Institute of Science and Technology, [4]The Hong Kong University of Science and Technology

Abstract

AI plays a key role in current cyberspace and future immersive ecosystems that pinpoint user experiences. Thus, the trustworthiness of such AI systems is vital as failures in these systems can cause serious user harm. Although there are related works on exploring trustworthy AI (TAI) metrics in the current cyberspace, ecosystems towards user-centered services, such as the metaverse, are much more complicated in terms of system performance and user experience assessment, thus posing challenges for the applicability of existing approaches. Thus, we give an overlook on fairness, privacy and robustness, across the historical path from existing approaches. Eventually, we propose a research agenda towards systematic yet user-centered TAI in immersive ecosystems.

1 Introduction

Today, artificial intelligence (AI) methods have shown state of the art performance in many fields and are becoming increasingly widespread in many areas of everyday life, including recommender systems [Cheng *et al.*, 2016], health care [Norgeot *et al.*, 2019], smart factory [Shiue *et al.*, 2018], financial modeling [Lin *et al.*, 2011], marketing [Cui *et al.*, 2006], education, science, and commerce [Jordan and Mitchell, 2015]. However, such integration also allows AI systems to access large datasets from countless users. AI systems can leverage these datasets along with significant networking and computing power to learn very granular and potentially sensitive user behavior. Additionally, to most users, AI systems appear as black-boxes that provide little insight into their internal decision-making process. This arouses moral concerns surrounding AI systems and especially the trustworthiness of AI for the sake of fairness, privacy, security and system reliability [Adadi and Berrada, 2018].

To address these concerns and strengthen human trust in AI systems, the area of Trustworthy AI (TAI) has recently seen significant attention from government entities, such as the European Commission [HLEGAI, 2019], United States Department of Defense [Heaven, 2021], China's Ministry of Science

---

*The Corresponding Author: Pengyuan Zhou, E-mail: (pyzhou@ustc.edu.cn)

and Technology, and numerous technology giants such as IBM, Google, Facebook. The major goal of TAI is to ensure the protection of people's fundamental rights while still allowing responsible competitiveness of businesses [European Union, 2012]. The term TAI has been around for years and was boosted by the well-known EU guidelines on TAI published in 2019 [HLEGAI, 2019]. The concept grew rapidly as the number of research papers in Google Scholar with the term in the title or abstract increased from 6 to 1,040 over 2017 – 2021.

In the TAI domain, TAI metrics are naturally a major issue and critical to accurately measuring the degree of system trustworthiness and the amount of protection offered by AI-enabled technologies. As a quantification of trustworthiness, a TAI metric can be a form of one or multiple system properties or system states to assess trustworthiness. The appropriate TAI metrics can vary in different domains due to various scenarios, user demand, processed data, adversaries, regulations, and laws. As integration of several related requirements, TAI should contain multiple 'dimensions', even though multi-criteria evaluation will lead to increased complexity. Despite the large number of case-by-case metrics used in current literature, a comprehensive and systematic outline of TAI focusing on metric selection has yet to be proposed, resulting in challenges for metric choices for non-experts and even professionals. Furthermore, future immersive ecosystems, such as the metaverse, are going to incorporate more complex systems that blend the virtual and physical worlds, and, more importantly, complicated definitions of system performances and user experiences. Therefore, current metrics and metric selection logic might need enhancements and evolution to fit the complexity.

Related Surveys and Our Scope. We acknowledge a number of existing surveys on ethical guidelines for AI [Smuha, 2019], big data [Mantelero, 2018] and robotics [Torresen, 2018]; along with surveys on specific domains (e.g., finance [Lin *et al.*, 2011]) and types of AI applications (e.g., recommender systems [Gunes *et al.*, 2014]). In contrast, this survey explores recent trends and advancements in TAI metric selection across two important domains, and sheds light on the metric selection logic for future user-centered cyberspace. Note that legal compliance (lawful AI) is beyond the scope of this survey. Specifically, we focus on the technical metrics for *fairness*, *privacy*, and, *robustness*. Other TAI requirements,

including transparency and accountability, do not have sufficient literature yet for survey purposes and thus are left for future work. Since AI is a socio-technical system instead of a mathematical abstraction, this work requires readers to personalize the metric selection on demand.

Review Methodology. The selection of publications was conducted in four steps. First, we search in Google Scholar for "Trustworthy AI", and usecase names (e.g., recommendation system) plus TAI requirements (e.g., fairness) and find that relevant publications are spread across multiple scientific journals and conferences. Second, we choose papers primarily from three scientific repositories: ACM Digital Library, IEEE Explore, and Springer. Third, we try to select the top cited or top-venue (e.g., KDD and WWW) papers if proper papers exist. Finally, if no proper references are found from the second and third steps, we choose the most feasible references by the searching through Google Scholar with related keywords such as "AI in Networking" and filter by topics with relevant to TAI.

Contributions. This survey serves as a first effort to outline the critical TAI metrics in current domains, capture the general logic of metric selection, and call for advancing user-centered metrics for immersive ecosystems and autonomous metric selections. Specifically, our contributions are threefold. First, we describe the importance of the TAI requirements this article focuses on (Section 2). Second, we outline two most fundamental domains and dive into specific use cases therein to summarize TAI metrics in the current cyberspace (Section 3). Finally, we summarize the lessons learned from existing TAI metric/ selection methods, and discuss the research agenda for future ecosystems' with TAI as well as the related metric selection methodology (Section 4).

2 TAI: User-centered Requirements

As mentioned, we focus on **fairness**, **privacy**, **robustness**, and the key metrics to meet these requirements in different domains. In this section, we present an overview of these requirements and list some common metrics.

Fairness. AI, especially machine learning, often presents statistical discrimination due to non-identical data distribution or resources, which leaves certain privileged groups with performance advantages and others with disadvantages. The learning bias, regardless if generated on purpose or accidentally, exacerbates existing resource inequity and further harms fairness in society [Paluck *et al.*, 2019]. Therefore, the AI community has been putting efforts on measuring the fairness of AI systems to mitigate algorithmic bias.

Privacy. Privacy is fundamental yet hard to define explicitly. Nissenbaum [Nissenbaum, 2004] defines privacy in terms of contextual integrity and contextual information norms dictating how information may be used or shared. As agreed by most researchers, privacy is a multi-dimensional concept [Laufer, 1973] and thus is normally assessed via multiple metrics focusing on the exposure of private information.

Robustness. Due to natural variability or dynamic system conditions over time, predicting how future conditions will change might be hard or impossible. The term deep uncertainty, in this scenario, is more proper to describe the issue in AI. Deep uncertainty is defined as a situation in which parties to a decision do not know or cannot agree on how the system (or part of it) functions, the importance of the various outcomes of interest, and/or what the relevant exogenous inputs to the system are and how they might change in the future [Maier *et al.*, 2016]. Under deep uncertainty, typically, multiple environmental state factors, i.e., future conditions, jointly affect the decisions (e.g., policies, designs and plans) [Ben-Tal *et al.*, 2009], resulting in influences on the considered performance metric (e.g., cost, utility and reliability). Robustness metrics function as a transformation of the performance metrics under these future conditions.

General Rule for Metric Selection. In general, the system administrator can follow a series of steps to select the proper metrics: (1) Which requirements of trustworthiness should be assessed? (2) Who cares about the issue the most, e.g., the system administrator, the users, the regulators, society etc.? (3) Regarding each requirement, who are the major concerned entities for each party (because different parties may have different concerns, e.g., system admins care about performance while users care about privacy), e.g., consistent performance, protected data, equal performance? (4) What is the targeting or common adversary? (5) What are the available data resources to compute the selected metrics? (6) What is the difficulty and cost of the metric assessment? (7) Will the metrics stay valid over time?

Nevertheless, no matter how proper the selected metrics are, they are still only estimations and will not fully encompass all the desired TAI requirements accurately. Additionally, if maximizing the selected metrics became the major model optimization logic, the model may perform well in terms of the TAI metrics but fall short of the original model goal(s). Therefore, periodic user studies are recommended to continuously monitor the system compliance with the TAI metrics. Adaption of the metric selection or standard can be made accordingly to strike a balance between the original goal and trustworthiness.

3 TAI Metrics in the Existing Cyberspace

Although existing TAI guidelines provide assessment metrics, the metrics are often quite high level and thus for individual providers to find or create detailed definitions (the definition of a metric may vary in different fields) is not easy. Moreover, different domains have different performance priorities. Thus, the selection of TAI metrics should naturally also vary. Future ecosystems, with more complexity including in the user experience, only exacerbate these issues and pose more challenges for TAI metric selection. Therefore, we first examine TAI metrics in the current cyberspace and summarize lessons for future ecosystems. In this section, we outline TAI metrics in computing and networking, two basic components of cyberspace that leverage AI.

We find that some metrics are widely used in varying ways across different domains, including regression model metrics, including mean absolute error (MAE), normalized MAE (NMAE), mean squared error (MSE) and root MSE (RMSE),

as well as ranking metrics, e.g., hit ratio, precision, recall, specificity (true negative rate), F-score, discounted cumulative gain (DCG), and their varieties. As they usually appear as a group, we refer to them as *regression model metrics* and *ranking metrics* throughout the survey.

3.1 Computing

This section focuses on the computing domain. We choose two representative usecases: search engine ranking (SER) and recommendation systems (RecSys), to illustrate the common TAI metrics. These usecases impact huge numbers of users (billions) thus are major concerns in terms of TAI.

Nowadays, SER algorithms consider many factors such as dwell time, page relevance, content quality, and so on. When a search engine presents results, it typically records or calculates such factors based on pre-defined policies, and treat these as implicit proof of user interest. Therefore, the user interaction with the ranking results is critical for training the learning to rank (LR) models. Similarly, RecSys incorporates a number of machine learning techniques and has been widely used in online media platforms, online shopping, and social networks. A RecSys normally collects users' historical choices for supervised learning (e.g., classifying items as recommended or not) or unsupervised learning (e.g., matrix factorization techniques common in collaborative filtering) to learn and predict the user interest in items.

SER and RecSys provide sorted results to users aiming to show results that match the user's search input or recommend items of user interest. However, when a user clicks on the top-ranked link on the result page determining whether the selection is simply because the result is top ranked/recommended or really the most relevant/interesting is difficult, as discovered by the researchers [Joachims *et al.*, 2017]. Since top results have higher chances of being selected thus potentially increasing revenues, it is important to guarantee the trustworthiness of the presented results for the benefit of users.

Fairness Fairness is a major concern for SER and RecSys. For instance, SERs sometimes are found to systematically favor certain sites over others in the results, thus distorting the objectiveness of the results and degrading user trust [Tavani, 2012]. Additionally, in RecSys, the number of recommendations is often fixed, therefore there are strong incentives to promote products with greater commercial benefit instead of fair recommendations based on ethical data mining.

Besides commercial incentives, implicit biases based on ethnicity, gender, age, community and so on, that widely exist in society, are often reflected in the big data collected from the Internet. Machine learning models thus often adopt these biases when being trained on the bias-embedded datasets. Technical flaws during data collection, sampling, and model design can further exacerbate unfairness by introducing edge cases, sampling bias, and temporal bias. Therefore, the measurement of fairness with proper TAI metrics is critical for both SER and RecSys to guarantee that the users receive neutral and impartial services.

A common strategy of metric selection is to focus on one factor and measure the deviation from the equality of that factor. For example, SER normally focuses on the (potential) attention items receive from users in terms of factors such as click-through rates, exposure, or inferences of the content relevance. The deviation from equality for these factors can then be quantified via disparate impact, disparate exposure, disparate treatment [Singh and Joachims, 2018; Castillo, 2019; Zehlike and Castillo, 2020], and inequity of attention [Biega *et al.*, 2018]. RecSys has also employed similar metrics for fairness, such as bias disparity, average disparity, and score disparity [Tsintzou *et al.*, 2018; Leonhardt *et al.*, 2018]. Kullback–Leibler (KL)-divergence has also been employed with ad-hoc adaptions to measure the fairness in SER [Geyik *et al.*, 2019].

Other issues also affect the definition and selection of metrics. For instance, fairness can refer to disparate treatment of individuals and of demographic groups, commonly termed as individual fairness [Rastegarpanah *et al.*, 2019] and group fairness [Kamishima *et al.*, 2012]. The former can be seen as a special case of the latter where the group size equals one. We further discuss more similar matters in Section 4.

Privacy Service providers are consistently improving the user experience by providing personalized service using machine learning models trained with data about users' personal behavior and interests. A common method to acquire such data is to request permission to collect data when associating the search engine or RecSys services with a user account, e.g., Google and Youtube can be associated with the user's Google account. While improving user experience, this also raises privacy issues as personal information is sent to a remote server [Xu *et al.*, 2007]. Considering the volume and detail of data current systems collect, privacy concerns should be taken seriously [Jeckmans *et al.*, 2013]. Additionally, companies often have financial incentives that conflict with protecting user privacy. For example, there has been a continuous series of privacy concerns over Google services. In 2012, Google changed its privacy policy to enable sharing data across a wide variety of services [1]. While in 2016, Google quietly dropped its ban on personally-identifiable information in its DoubleClick ad service [2].

A common measurement logic is to assess the unwilling exposure of private information. For unstructured data like browsing history and email, entropy can provide a measure of unique information and quantify the amount of exposed private information [Agrawal and Aggarwal, 2001]. Though, in practice, privacy preservation techniques can affect model performance. Therefore, privacy is commonly assessed as a multi-objective optimization problem. For example, cryptographic protocols, differential privacy, and anonymization approaches assess privacy preservation via trade-off measurements between privacy exposure (controlled by some characteristics of protocols or algorithms) and model performance using *regression model metrics* and *ranking metrics* [Xu *et al.*, 2007; McSherry and Mironov, 2009; Xin and Jaakkola, 2014].

Robustness As mentioned, enormous volumes of data are continuously generated online, thus models often require

---

[1]https://policies.google.com/privacy/archive/20120301
[2]https://thetechportal.com/2016/10/21/google-now-personal-web-tracking-ads/

quite significant time and resources for retraining or incremental learning and therefore cannot always be done timely. This slow retraining and learning can result in issues after sudden data shifts. In addition to real data shifts, the problem of spamdexing, namely when users enter fake ratings to manipulate the ranking results, still exists in SER and related attacks are also common in collaborative filtering RecSys nowadays, dubbed as shilling attacks or profile injection attacks [Lam and Riedl, 2004; Williams *et al.*, 2007]. Finally, malicious users may intentionally apply small targeted perturbations to datasets which can severely impact model performance, e.g., image classification tasks in multimedia recommender systems [Moosavi-Dezfooli *et al.*, 2017]. Overall, an SER or RecSys is considered robust if not significantly affected by attacks like spamdexing and perturbations or dramatically data shifts [Aggarwal, 2016].

A common measurement method is to assess the model performance in the face of varying attacks or data shifts using *ranking metrics* [Wang *et al.*, 2013; Li *et al.*, 2009; Bailey *et al.*, 2017; Tang *et al.*, 2019] and *regression model metrics* [O'Mahony *et al.*, 2004; Lam and Riedl, 2004; Mobasher *et al.*, 2006; Mobasher *et al.*, 2007; Cheng and Hurley, 2010; Gunes *et al.*, 2014].

3.2 Networking

This section focuses on the networking domain and illustrates several example networking problems where different TAI metrics are in use. These examples are relevant for both wireless and wired networks and cover several different network layers. In contrast to the computing domain, most networking research considers only a single TAI metric area (as others are often considered irrelevant to the problem or out of scope); thus, each example problem also describes only a single TAI metric area. Relatedly, the metrics are more heterogeneous and ad-hoc than in the computing domain because TAI is very new in the networking domain; therefore, more detail and context are provided for each metric.

Wireless networking fairness. Radio resource allocation in wireless networks is a well-known networking problem with significant AI research and a major fairness component. Specifically, in the LTE context, this problem manifests as the allocation of physical resource blocks to specific user equipment on the sector level. Fairness in this context means ensuring that certain users are not starved of resources. Overall, such a problem is a multi-objective optimization problem with fairness as one objective with others such as high overall system performance (throughput). As network complexity and application diversity have risen, simple analytic or heuristic scheduling solutions (e.g., round-robin, proportional fair, and best CQI) are seen as potentially insufficient, and research has turned to reinforcement learning (RL) to solve the problem.

The fairness aspect in RL is thus embedded in the reward function as this function directs the learning. In the simplest cases, this fairness component of the reward function can be a weighted version of a traditional network fairness metric, such as Jain's fairness index or entropy, or a custom fairness metric resulting from reward engineering for the specific problem. The basic Jain's fairness index is defined for a single

type of network quality of service (QoS) measure (typically user throughput). Thus, the index does not consider further QoS measures (such as delay or packet loss) or users with different applications (and thus different QoS requirements). This index can be analogized to the independence group of algorithmic fairness measures as the index is blind (to such characteristics) and thus independent.

In more complex cases, such as with multiple QoS measures, different approaches are possible. For example Comsa et al. [2019] considers fairness (within a group of users of the same app) through a measure of the total sum of user-QoS requirement combinations (e.g., user A with throughput threshold of X) met in a given period (one TTI). The authors then further generalizes this by considering many user groups with each group using another application. Specifically, the total fairness is a weighted sum of the intra-app-group fairnesses, with the weighting being a learnable (through RL) prioritization of the applications. As another example, Al-Tam et al. [2020] use a variant of discounted best-CQI (a common network fairness model) as the reward function with the discounting based on the well known min-max ratio fairness measure.

Network traffic privacy. Network traffic classification is a major networking area, especially for mobile network operators as knowing the specific network traffic mixture supports many network tasks. For example, traffic mixture knowledge enables network optimizations including optimizing user QoS and QoE and better traffic volume prediction. However, more nefariously, traffic classification can also impinge on user privacy as entities (such as companies and governments) can use such classification to identify, for example, users of specific apps or websites (like those used by political dissidents). These network traffic classifiers often use a variety of ML and AI methods, including RF, SVM, and DNN, with and without handcrafted traffic features. Thus, to counteract network traffic classification, researchers are applying both AI methods such as generative adversarial networks (GANs) [Li *et al.*, 2019; Fathi-Kazerooni and Rojas-Cessa, 2020; Hou *et al.*, 2020] and non-AI methods, such as adaptive packet padding [Pinheiro *et al.*, 2020] and optimized dummy packet injection [Shan *et al.*, 2021], to intelligently obfuscate the network traffic. These research studies primarily use related accuracy-based metrics to assess privacy improvement.

Some works use the misclassification rate of the classifiers on the obfuscated traffic [Shan *et al.*, 2021; Hou *et al.*, 2020]. Relatedly, other works use the differences in the classification accuracies of the classifier on the original and obfuscated traffic [Pinheiro *et al.*, 2020]. Goal-wise, certain studies analyze the classifier accuracy, recall, and precision on traffic of one type disguised (by a GAN) to look like traffic of a different (target) type (rather than a goal of just general obfuscation) [Fathi-Kazerooni and Rojas-Cessa, 2020; Hou *et al.*, 2020]. Finally, Li et al. [2019] uses both indistinguishability under Classification Attack (IND-CA) and the differences in AUC of the ROC curve of the classifier on the original and obfuscated traffic. IND-CA quantifies how distinguishable the traffic is when considering two traffic types

with equal shares (each representing 50% of traffic). Specifically, IND-CA represents the normalized benefit of the classification with zero meaning a random guess and, in contrast, one meaning full certainty.

Networking congestion control robustness. Similar to the resource allocation problem, the networking congestion control problem also lends well to an RL approach. The issue also deals with significant multi-layer network complexity (where simpler heuristics are insufficient). Specifically, the major target is TCP congestion control with RL adapting the receive window size. A potential benefit to the RL approach is its robustness to different network conditions.

However, this is sometimes only the case if those network conditions were part of the training regime (for offline RL). Thus, robustness testing with conditions that both span and extend beyond the training regime is important. Some of the TCP RL works use regret-based metrics with a baseline scenario containing only network conditions from the training range. The metric is then defined as the performance gap between this baseline and scenarios that include conditions beyond the training regime [He *et al.*, 2021].

In other cases where robustness to different conditions is built-in to the RL approach, the metric is simply the performance gap in the diverse conditions of this approach from baseline approaches. Du et al. [2021], for example, use a hybrid approach with traditional (heuristic) and RL parts to improve robustness in both diverse wired and wireless network situations. The approach proves better (in terms of mean throughput and delay) than either approach alone in such situations. Other works use statistical dispersion metrics such as standard deviation and confidence intervals to illustrate the general robustness of the results (for example, where stability is important). For instance, an RL approach [Xiao *et al.*, 2019] illustrates a 95% confidence interval that is smaller than all the baselines for a specific performance measure (a fairness index); thus illustrating robustness to more significant swings.

4 Lessons Learned and Research Agenda

By examining the TAI metric selection across computing and networking domains, we can see that the definition and selection of TAI metrics for computing are more straightforward than for networking. A significant reason is that the outputs of many computing systems like recommendation systems and search engines are more oriented towards end-users, e.g., the results of ranking algorithms well match the needs of SER which uses user data and show results directly to users.

In contrast, the learning algorithms in the networking context usually result in intermediate metrics that serve to adapt protocols or algorithms that eventually affect the target metrics. In other words, the output of the learning algorithms in computing can often be directly used to assess user experience, while in networking there is normally an intermediate model to transfer networking performance (controlled by AI) to user experience. Thus, TAI metrics in networking require more ad-hoc designs, definitions of usage and user context, and targets of networking systems.

Currently, most computing and networking systems use "functionality-driven design", which we use as a contrast to "user-centered design", in the sense that the former focuses more on pre-defined systematic performance metrics though also sometimes considers user-related metrics, such as QoS and QoE. Relatedly, TAI design and metric selection in such systems also often focus on the most functionalities during specific life-cycle phases. Ideally, TAI metric selection demands more thorough considerations to guarantee trustworthiness through the life cycle of usage. Furthermore, the current mindset of TAI design and metric selection, restricted by the aforementioned design philosophies, takes into consideration only part of human cognition, specifically the conscious and concrete areas that can be more easily measured and quantified, such as pattern recognition, language, attention, perception, and action. These are widely explored by AI communities.

However, the exploration of the unconscious and abstract areas of cognition, e.g., mental health and emotion, is just beginning. Methodological limits is a key reason for this, e.g., lack of devices and theories to accurately capture bioelectrical signals and convert these signals to emotional statuses.

Trustworthiness itself consists of cognitive, emotional and behavioral factors, since trustworthiness is a user-oriented term. This aspect will play an increasingly important role when "user-centered design" dominates future cyberspace, replacing the current "functionality-driven design". In the future, considering the unconscious and abstract cognition areas will be vital to guarantee TAI. These parts are hard to quantify and might remain so even with advanced techniques in sensor-enabled immersive cyberspace. Therefore, other assessment methods may be required for TAI in such immersive cyberspaces. This is discussed in the remaining paragraphs.

From ad-hoc to systematic metric selection Although we presented an outlook on how TAI metrics are selected from the surveyed literature in computing and networking domains, it can be more complicated when system developers try to select or define TAI metrics in reality. Because achieving trustworthiness highly relies on the contextual environment besides the AI system itself, the selection of metrics requires a holistic and systemic consideration encompassing all processes within the system's socio-technical context throughout its entire life cycle.

Therefore, every system, even if it is an identical copy of another, may require different TAI metrics or some metric adaptions due to the differences in the deployed contextual environments and life cycles. Moreover, the selection of metrics also depends on the granularity of the concerned target (pointwise, pairwise, listwise) and the operation phase (pre-processing, in-processing, post-processing). Thus, even for straightforward computing tasks, developers need delicate considerations about the whole context and potential scenarios to fully justify the selection of TAI metrics.

TAI metric selection for immersive cyberspace Current cyberspace is evolving as new technologies develop, and the advent of immersive cyberspace will be enabled by AI in a greater extent. For example, since 2021, the metaverse, also known as *the immersive Internet*, has risen to public attention,

thanks to Facebook's rebranding, and worldwide academic and industrial promotion. In the metaverse, AI continues to play a core role as the foundation of several key technologies, namely, computer vision, augmented & virtual reality (AR/VR), data mining, and robotics [Lee *et al.*, 2021]. The most critical difference between the metaverse and current cyberspace is that in the metaverse, human users are absorbed into the projected blended virtual-physical world without explicit exit points instead of just standing by as external interactors. As such, any consequences caused by AI misbehavior could be significantly worse and hence severely impact human users' welfare.

For example, VR technologies are capable of recording richer personal data, such as eye movements and emotional reactions, which could be deployed in threatening ways powered by AI techniques to manipulate users' beliefs, emotions, and behaviors [Spiegel, 2018]. In contrast, AR requires strong contextual awareness, in terms of users, their adjacent environments, and social interactions, to augment the physical world [Lam *et al.*, 2021]. As such, users have to share egocentric views of various contexts, e.g., Project EGO4D[3].

In other words, users have to build trust with virtual-physical blended cyberspace mediated by computing systems, as the users' daily interactions with physical worlds and people are recorded in an unprecedentedly massive scale. Meanwhile, users will require an easy-to-interpret score, e.g., AI Trust Score [Wang and Moulden, 2021], to judge AI trustworthiness over time (instead of a single snapshot), while AI is resilient to the occurrence of glitches and afterwards able to recover trusts with users. Furthermore, the "user-centered" features of the metaverse may bring important changes to TAI and TAI metric selection. In current cyberspace, AI-enabled applications and human users interact but are still significantly and explicitly separated. Hence the measure of TAI mainly focuses on system performance and technical metrics.

In the metaverse, however, user-representative avatars, cognitive emotional-interactive products, and other similar humanoids will play vital roles in improving a user's feeling of involvement to seamlessly experience the blended virtual-physical world. More importantly, the aforementioned avatars and humanoids will collaborate with human users. Thus, user-centered TAI metrics, such as those focusing on cognition, sentiment, and psychology, with sensor-enabled monitoring in the metaverse, will become a new driver for understanding robustness, privacy and fairness.

Emerging techniques may have the potential to tackle the challenge through understanding the internal states of users. For example, the state-of-the-art Brain-Computer Interface (BCI) can estimate the user's current emotion, attention, or fatigue level to some extent by monitoring the bioelectrical signals that reflect brain activity. These signals can be recorded by a device like an electroencephalogram [Shatilov *et al.*, 2021]. These emerging techniques may allow for quantitatively measuring the abstract metrics that currently rely on limited-scale qualitative experiments (often based on user interviews).

Nevertheless, these techniques are still immature. More-

over, the practicality and applicability of the techniques maybe be limited, as they normally require the users to wear additional devices, which are often inconvenient. Nowadays, immersive headsets (AR/VR) have similar issues [Lee *et al.*, 2022]. Therefore, requirements for qualitative measures like user studies for TAI assessment, e.g., eliciting user requirements for trust-guarantee AI services, might be reasonable. However, the current approaches for understanding users, to a large extent, are costly and time-consuming. Thus, this timeliness issue is a current challenge to be solved.

TAI governance Currently, the governance of the trustworthiness of AI systems is mostly left in the hands of the service providers themselves and some third-party companies and government institutes. As user awareness of TAI increases, a key challenge for TAI governance is how to assess and guarantee fairness, privacy, and robustness in a more standardized, transparent, and systematic way. Moreover, the coming metaverse will integrate more AI services into daily life on an unprecedentedly massive scale. Governance by each individual service provider creates coordination and standardization problems and thus provides no guarantee of equity in trustworthiness standards. Additionally, the burden might be too large for a limited number of third party companies or government institutes. Accordingly, building an autonomous governance platform, "meta-TAI", to govern TAI performance, might be worth exploring. The meta-TAI platform could be collaboratively governed by a number of trustworthy institutes authorized by the involved countries, as well as offering TAI scores for various providers of AI services and their individual solutions. As such, the platform can save manual effort while ensuring the equity of standards across various phases of a user-AI interaction cycle.

5 Conclusion

This survey discussed metrics for TAI and further examined the aspects of fairness, privacy and robustness in the computing and networking domains. The existing metrics are mainly driven by system functionalities and efficacy with less emphasis on user-centered factors. Meanwhile, the ad-hoc metric selection causes sub-optimal results in building trustworthiness with users. We revisited the TAI domain to lay out a research agenda that will assist researchers working on TAI and immersive cyberspace to contextualize and focus their efforts. We note that AI will become an indispensable driver of immersive cyberspace and that users will interact with AI-enabled services in this virtual-physical blended world under the premise that user trust is essential to the wide adoption of such services. Therefore, we call for a user-centered paradigm of building trustworthiness beyond sole system measurements and considering cognitive and affective factors.

References

[Adadi and Berrada, 2018] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[Aggarwal, 2016] C. C. Aggarwal. *Recommender systems*, volume 1. Springer, 2016.

---

[3]https://ego4d-data.org/

[Agrawal and Aggarwal, 2001] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proc. of the 20th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of database systems*, pages 247–255, 2001.

[Al-Tam *et al.*, 2020] F. Al-Tam, N. Correia, and J. Rodriguez. Learn to schedule (leasch): A deep reinforcement learning approach for radio resource scheduling in the 5g mac layer. *IEEE Access*, 8:108088–108101, 2020.

[Bailey *et al.*, 2017] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. Retrieval consistency in the presence of query variations. In *Proc. of the 40th Inter. ACM SIGIR*, pages 395–404, 2017.

[Ben-Tal *et al.*, 2009] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton Uni. press, 2009.

[Biega *et al.*, 2018] A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st Inter. ACM SIGIR Conf.*, pages 405–414, 2018.

[Castillo, 2019] C. Castillo. Fairness and transparency in ranking. In *ACM SIGIR Forum*, volume 52, pages 64–71. ACM USA, 2019.

[Cheng and Hurley, 2010] Z. Cheng and N. Hurley. Robust collaborative recommendation by least trimmed squares matrix factorization. In *2010 22nd IEEE Inter. Conf. on Tools with AI*, volume 2, pages 105–112. IEEE, 2010.

[Cheng *et al.*, 2016] H.-T. Cheng, L. Koc, J. Harmsen, et al. Wide & deep learning for recommender systems. In *Proc. of the 1st WKSP on deep learning for recommender systems*, pages 7–10, 2016.

[Comșa *et al.*, 2019] I.-S. Comșa, R. Trestian, G.-M. Muntean, and G. Ghinea. 5mart: A 5g smart scheduling framework for optimizing qos through reinforcement learning. *IEEE Trans. on Network and Service Mgt.*, 17(2):1110–1124, 2019.

[Cui *et al.*, 2006] G. Cui, M.-L. Wong, and H.-K. Lui. Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Mgt. Sci.*, 52(4):597–612, 2006.

[Du *et al.*, 2021] Z. Du, J. Zheng, H. Yu, L. Kong, and G. Chen. A unified congestion control framework for diverse application preferences and network conditions. In *Proc. of the 17th Inter. Conf. on emerging Networking Experiments and Tech.*, pages 282–296, 2021.

[European Union, 2012] European Union. Eu charter of fundamental rights. 2012.

[Fathi-Kazerooni and Rojas-Cessa, 2020] S. Fathi-Kazerooni and R. Rojas-Cessa. Gan tunnel: network traffic steganography by using gans to counter internet traffic classifiers. *IEEE Access*, 8:125345–125359, 2020.

[Geyik *et al.*, 2019] S. C. Geyik, S. Ambler, and K. Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proc. of the 25th ACM SIGKDD*, pages 2221–2231, 2019.

[Gunes *et al.*, 2014] I. Gunes, C. Kaleli, A. Bilge, and H. Polat. Shilling attacks against recommender systems: a comprehensive survey. *AI Review*, 42(4):767–799, 2014.

[He *et al.*, 2021] B. He, J. Wang, Q. Qi, et al. Deepcc: Multi-agent deep reinforcement learning congestion control for multi-path tcp based on self-attention. *IEEE Trans. on Network and Service Mgt.*, 18(4):4770–4788, 2021.

[Heaven, 2021] W. D. Heaven. The department of defense is issuing ai ethics guidelines for tech contractors. 2021.

[HLEGAI, 2019] HLEGAI. Ethics guidelines for trustworthy ai. *B-1049 Brussels*, 2019.

[Hou *et al.*, 2020] C. Hou, G. Gou, et al. Wf-gan: Fighting back against website fingerprinting attack using adversarial learning. In *2020 IEEE Symp.on Comp. and Commun. (ISCC)*, pages 1–7. IEEE, 2020.

[Jeckmans *et al.*, 2013] A. JP Jeckmans, M. Beye, Z. Erkin, et al. Privacy in recommender systems. In *Social media retrieval*, pages 263–281. Springer, 2013.

[Joachims *et al.*, 2017] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting click-through data as implicit feedback. In *ACM SIGIR Forum*, volume 51, pages 4–11. ACM USA, 2017.

[Jordan and Mitchell, 2015] M. I. Jordan and T. M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[Kamishima *et al.*, 2012] T. Kamishima, S. Akaho, H. Asoh, et al. Enhancement of the neutrality in recommendation. In *Decisions@ RecSys*, pages 8–14. Citeseer, 2012.

[Lam and Riedl, 2004] S. K. Lam and J. Riedl. Shilling recommender systems for fun and profit. In *Proc. of the 13th Inter. Conf. on WWW*, pages 393–402, 2004.

[Lam *et al.*, 2021] K. Y. Lam, L. H. Lee, and P. Hui. *A2W: Context-Aware Recommendation System for Mobile Augmented Reality Web Browser*, page 2447–2455. ACM, USA, 2021.

[Laufer, 1973] R. S. Laufer. Some analytic dimensions of priacy. In *Architectural psychology, Proc. of the Lund Conf., Lund: Studentlitteratur*, 1973.

[Lee *et al.*, 2021] L.-H. Lee, T. Braud, P. Zhou, et al. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *arXiv:2110.05352*, 2021.

[Lee *et al.*, 2022] L.-H. Lee, T. Braud, S. Hosio, and P. Hui. Towards augmented reality driven human-city interaction: Current research on mobile headsets and future challenges. *ACM CSUR*, 54:1 – 38, 2022.

[Leonhardt *et al.*, 2018] J. Leonhardt, A. Anand, and M. Khosla. User fairness in recommender systems. In *Compan. Proc. of the The Web Conf. 2018*, pages 101–102, 2018.

[Li *et al.*, 2009] X. Li, F. Li, S. Ji, Z. Zheng, et al. Incorporating robustness into web ranking evaluation. In *Proc. of the 18th ACM Conf. on Info. and Knowledge Mgt.*, pages 2007–2010, 2009.

[Li *et al.*, 2019] J. Li, L. Zhou, H. Li, L. Yan, and H. Zhu. Dynamic traffic feature camouflaging via generative adversarial networks. In *2019 IEEE Conf. on Comm. and Network Security (CNS)*, pages 268–276. IEEE, 2019.

[Lin *et al.*, 2011] W.-Y. Lin, Y.-H. Hu, and C.-F. Tsai. Machine learning in financial crisis prediction: a survey. *IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):421–436, 2011.

[Maier *et al.*, 2016] H. R. Maier, J. HA Guillaume, H. van Delden, et al. An uncertain future, deep uncertainty, scenarios, robustness and adaptation: How do they fit together? *Environ. Modelling & Soft.*, 81:154–164, 2016.

[Mantelero, 2018] A. Mantelero. Ai and big data: A blueprint for a human rights, social and ethical impact assessment. *Comp. Law & Security Review*, 34(4):754–772, 2018.

[McSherry and Mironov, 2009] F. McSherry and I. Mironov. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *Proc. of the 15th ACM SIGKDD*, pages 627–636, 2009.

[Mobasher *et al.*, 2006] B. Mobasher, R. Burke, and J. J. Sandvig. Model-based collaborative filtering as a defense against profile injection attacks. In *AAAI*, volume 6, page 1388, 2006.

[Mobasher *et al.*, 2007] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM TOIT*, 7(4):23–es, 2007.

[Moosavi-Dezfooli *et al.*, 2017] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *Proc. of the IEEE Conf. on CVPR*, pages 1765–1773, 2017.

[Nissenbaum, 2004] H. Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

[Norgeot *et al.*, 2019] B. Norgeot, B. S. Glicksberg, and A. J. Butte. A call for deep-learning healthcare. *Nature medicine*, 25(1):14–15, 2019.

[O'Mahony *et al.*, 2004] M. O'Mahony, N. Hurley, N. Kushmerick, and G. Silvestre. Collaborative recommendation: A robustness analysis. *ACM TOIT*, 4(4):344–377, 2004.

[Paluck *et al.*, 2019] E. L. Paluck, S. A. Green, and D. P. Green. The contact hypothesis re-evaluated. *Behavioural Public Policy*, 3(2):129–158, 2019.

[Pinheiro *et al.*, 2020] A. J. Pinheiro, P. F. de Araujo-Filho, J. de M. Bezerra, and D. R. Campelo. Adaptive packet padding approach for smart home networks: A tradeoff between privacy and performance. *IEEE Internet of Things J.*, 8(5):3930–3938, 2020.

[Rastegarpanah *et al.*, 2019] B. Rastegarpanah, K. P. Gummadi, and M. Crovella. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *Proc. of the 12th ACM Inter. Conf. on WSDM*, pages 231–239, 2019.

[Shan *et al.*, 2021] S. Shan, A. N. Bhagoji, H. Zheng, and B. Y. Zhao. Patch-based defenses against web fingerprinting attacks. In *Proc. of the 14th ACM WKSP on AI and Security*, pages 97–109, 2021.

[Shatilov *et al.*, 2021] K. A. Shatilov, D. Chatzopoulos, L.-H. Lee, and P. Hui. Emerging exg-based nui inputs in extended realities: A bottom-up survey. *ACM Trans. Interact. Intell. Syst.*, 11(2), jul 2021.

[Shiue *et al.*, 2018] Y.-R. Shiue, K.-C. Lee, and C.-T. Su. Real-time scheduling for a smart factory using a reinforcement learning approach. *Computers & Industrial Eng.*, 125:604–614, 2018.

[Singh and Joachims, 2018] A. Singh and T. Joachims. Fairness of exposure in rankings. In *Proc. of the 24th ACM SIGKDD*, pages 2219–2228, 2018.

[Smuha, 2019] N. A. Smuha. The eu approach to ethics guidelines for trustworthy artificial intelligence. *CRi-Comp. Law Review Inter.*, 2019.

[Spiegel, 2018] J. S. Spiegel. The ethics of virtual reality technology: Social hazards and public policy recommendations. *Sci. and Eng. ethics*, 24(5):1537–1550, 2018.

[Tang *et al.*, 2019] J. Tang, X. Du, X. He, F. Yuan, et al. Adversarial training towards robust multimedia recommender system. *IEEE Trans. on KDE*, 32(5):855–867, 2019.

[Tavani, 2012] H. Tavani. Search engines and ethics. 2012.

[Torresen, 2018] J. Torresen. A review of future and ethical perspectives of robotics and ai. *Frontiers in Robotics and AI*, 4:75, 2018.

[Tsintzou *et al.*, 2018] V. Tsintzou, E. Pitoura, and P. Tsaparas. Bias disparity in recommendation systems. *arXiv:1811.01461*, 2018.

[Wang and Moulden, 2021] J. Wang and A. Moulden. *AI Trust Score: A User-Centered Approach to Building, Designing, and Measuring the Success of Intelligent Workplace Features*. ACM, USA, 2021.

[Wang *et al.*, 2013] Y. Wang, L. Wang, Y. Li, et al. A theoretical analysis of ndcg ranking measures. In *Proc. of the 26th Conf. COLT*, volume 8, page 6. Citeseer, 2013.

[Williams *et al.*, 2007] C. A. Williams, B. Mobasher, and R. Burke. Defending recommender systems: detection of profile injection attacks. *Service Oriented Comp. & Apps.*, 1(3):157–170, 2007.

[Xiao *et al.*, 2019] K. Xiao, S. Mao, and J. K. Tugnait. Tcp-drinc: Smart congestion control based on deep reinforcement learning. *IEEE Access*, 7:11892–11904, 2019.

[Xin and Jaakkola, 2014] Y. Xin and T. Jaakkola. Controlling privacy in recommender systems. NeurIPS, 2014.

[Xu *et al.*, 2007] Y. Xu, K. Wang, B. Zhang, and Z. Chen. Privacy-enhancing personalized web search. In *Proc. of the 16th Inter. Conf. on WWW*, pages 591–600, 2007.

[Zehlike and Castillo, 2020] M. Zehlike and C. Castillo. Reducing disparate exposure in ranking: A learning to rank approach. In *Proc. of The Web Conf. 2020*, pages 2849–2855, 2020.