



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798




INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 11, November 2024

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.625

 9940 572 462

 6381 907 438

 ijircce@gmail.com

 www.ijircce.com



Investigations on Artificial Intelligence Algorithmic Ethics

Kotersh Naik D, Suhas K C, Abhishek S A, B P Harshith

Assistant Professor, Department of Computer Science and Engineering, CIT, Tumkur, Karnataka, India

Assistant Professor, Department of Computer Science and Engineering, CIT, Tumkur, Karnataka, India

U.G. Student, Department of Computer Engineering, CIT, Tumkur, Karnataka, India

U.G. Student, Department of Computer Engineering, CIT, Tumkur, Karnataka, India

ABSTRACT: The ethical concerns surrounding artificial intelligence have gained widespread attention due to its rapid development. This essay addresses the issues of prejudice, data security, and the incorporation of morality and values into AI algorithms' ethics. It also thoroughly and methodically explains the primary fixes for these three issues. The three main technical approaches now used to address data security concerns are shown to have limited application, high reliance, and significant computation and communication cost. and the future trend is the development and integration of the three paths; the imposition of morality and values is challenging when considering learning ability, adaptability, and interpretability simultaneously; and the solution to the bias problem needs to be improved in terms of interpretability under the condition that the underlying fairness is difficult to determine. In order to offer insights and resources for AI ethics-related research, this study concludes by analyzing and anticipating the evolution of AI ethics.

KEYWORDS: Problems with bias, data security, artificial intelligence algorithm ethics, and morality and value implantation

I. INTRODUCTION

Numerous industries, including business, manufacturing, agriculture, and the military, have made extensive use of the latest generation of Artificial Intelligence (AI) technology. However, the limitations of AI, like its poor interpretability and heavy reliance on data and mathematical power, have created pressing ethical issues for people, society, and the entire world. As a result, research on AI ethics is desperately needed to make sure that the development and application of AI systems are consistent with human morals and values. Despite its quick development, the subject of AI ethics is relatively young and lacks a theoretical framework and systematic set of norms. In conclusion, this paper offers a thorough and methodical explanation of the main issues and solutions in AI algorithmic ethics to assist AI developers and managers in creating and overseeing AI systems more logically. It also encourages policymakers to establish the legal framework and moral standards of AI in a more exacting and transparent manner in order to foster a positive feedback loop between AI ethics and ethical AI.

II. ARTIFICIAL INTELLIGENCE ALGORITHMS' ETHICAL CONCERNS

Algorithm ethics is the study of how to ensure that an algorithm does not depart from particular social ethics while being consistent with prevailing social values throughout its design, development, and implementation. This chapter will examine the ethical concerns surrounding AI algorithms, including those related to bias, data security, and morality and value implantation.

A. Problems with Data Security

Massive data is necessary for artificial intelligence systems to function, and data security issues are one of the many hazards associated with the collecting, storage, and use of this data [1]. In addition to causing user privacy breaches and the exploitation of personal data, data security issues can also endanger national data security and even have an impact on the political climate of the country. In order to address the issue of data security, academics have put forth three



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

main technological avenues: Federated Learning (FL), Multi-party computation (MPC), and Trusted Execution Environment (TEE).

TEE ensures the confidentiality and non-tamperability of data while offering a more secure environment for code execution and storage. The Open Mobile Terminal Platform separated a safe operating system for managing private data outside of the standard operating system in 2006, which is when TEE first emerged [2]. The global platform organization developed pertinent specifications and standards and formally introduced the idea of TEE up until 2010 [3]. At the moment, ARM's TrustZone technology and Intel's SGX (Software Guard Extensions) technology are the two most popular hardware solutions for TEE [4]. TEE has some advantages in terms of algorithmic logic support and computational efficiency. Its reliance on certain hardware support, however, may result in some restrictions. The TrustZone technology-enabled TEE system is one of them; it has significant implementation, architectural, and hardware restrictions that could lead to significant exploitable vulnerabilities [5]. Similar to this, SGX is the target of several attacks. The high probability of SGX being attacked is illustrated by the SmashEx exploit, which uses SGX's enclave interface to handle asynchronous exceptions [6].

Multiple participants can do calculations and provide conclusions using MPC, a key agreement procedure, without disclosing their individual input data [7]. MPC no longer exists at the theoretical level thanks to Malkhi's 2004 proposal of the Fairplay MPC platform [8]. The Danish beet auction system put forth by Bogetoft [9] is a prime example of MPC in action. It employs MPC to calculate the market clearing price without disclosing the bids of the sellers. Technique put forth by Bogetoft [9], which calculates the market clearing price using MPC without disclosing the seller's bid. Additionally, a joint project between businesses and Bestavro's Boston Women's Labor Council enables businesses to investigate the effect of employees' gender on real salaries and do the relevant statistical analyses without disclosing individual data [10]. Participants can work together on calculations using MPC without having to trust one another. Nevertheless, MPC's computational procedure is intricate and expensive in terms of both computing and communication overhead. Chain signatures have the advantage of supporting the underpinning elliptic curves of several blockchains, including the EVM and Bitcoin chains. The multi-signature wallet flow is displayed in Figure.1.

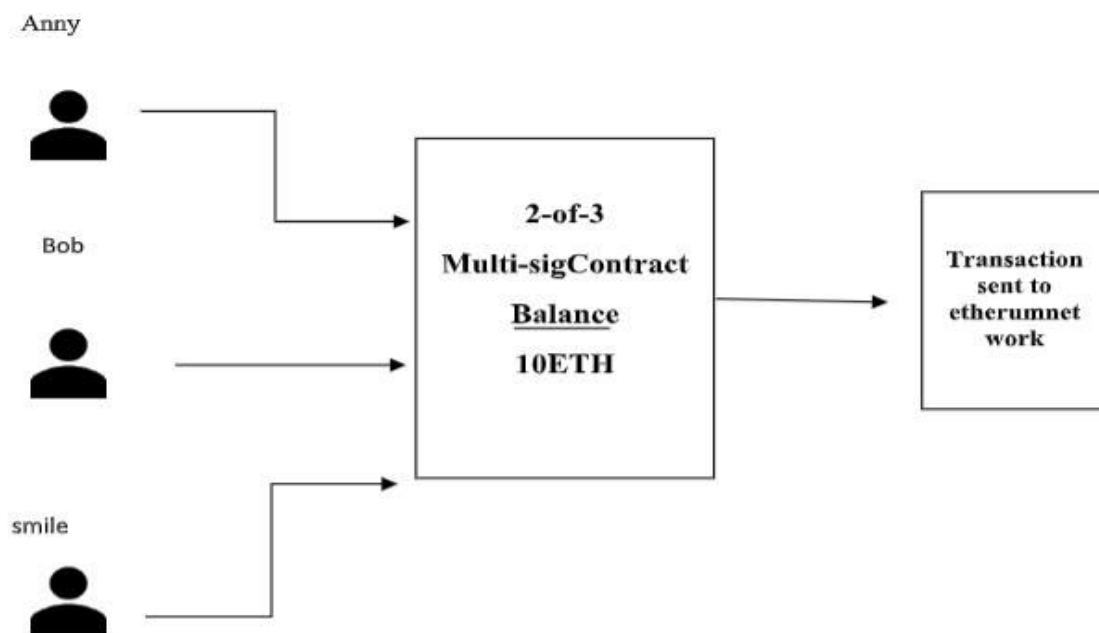


Figure.1. Multi-Signature Wallets



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

FL is a distributed machine learning technique that eliminates the requirement to centralize the original dataset in one place and enables several users to work together on model training. Local updates to the model parameters are calculated by each participant, which are then combined to create a global model. Google first proposed the idea of FL in 2016, outlining the fundamentals and possible uses of federated learning [11]. Google developed a scalable production system for federated learning in the mobile device space later in 2019 using TensorFlow [12]. In addition to its clear benefits in data independence, model stability, statistical heterogeneity, and system heterogeneity, FL enables model training on dispersed devices [13]. FL must, however, deal with the intricacy of synchronization and communication, which poses performance and security issues.

Though it is dependent on hardware vendors, TEE offers the best performance and adaptability of the three technological routes and is appropriate for business scenarios requiring high performance and huge volumes of data, Although MPC is the most developed and credible, it has limited application scenarios, a high development cost, and significant computational overhead and performance loss;

Although FL has the best synergistic qualities and works well for multi-party joint modeling, its controllability is poorer and its development time is shorter. The fusion development of the three technology courses has emerged as a current research hotspot due to the drawbacks of utilizing any one technique [14, 15].

B. The Issue of Bias

When an AI system is created, trained, or used, it may be biased against particular people or groups. This is known as the bias problem in AI. Biased AI systems have the potential to worsen societal inequity, cause a crisis of confidence, and cause people to make irrational decisions. As a result, in three distinct fields of artificial intelligence—machine learning, representation learning, and natural language processing—researchers have put up a number of solutions to the issue of bias [16]. Fair classification, fair regression, structured prediction, and fair causal inference are some of the methods used to address biased issues in machine learning. Training data is typically the foundation for fair classification, which teaches how to divide samples or objects into several groups. Dohyung et al. recently enhanced conventional fairness measures and classification performance tradeoffs in their fairness-aware bulk sampling strategy, which helped to alleviate the model's fairness issue [17]; Fair regression aims to achieve a fair outcome by altering the model's prediction results in order to decrease the prediction disparity of sensitive features. Among these, structured prediction predicted outputs with a specific structure by modeling the intricate dependencies between inputs and outputs, and Agarwal et al. regarded the classification problem as a real-valued target prediction problem through a fair regression approach, which successfully addressed the loss and regression problems while reducing bias [18]. Of these, Zhao and colleagues decreased the In order to mitigate the effects of bias amplification in multi-label categorization and visual-semantic role labelling, a collective reference algorithm based on Lagrangian relaxation was designed to reduce the frequency of co-occurrence of gender metrics and prediction task elements by introducing corpus-level constraints [19]. Fair causal inference is an inference technique designed to address bias and uncertainty in the estimation of intervention effects. Schlkopf et al. created algorithms to meet these requirements and put forth a natural causal non-discrimination criterion [20].

In the realm of representation learning, variational auto-encoders and adversarial learning are the primary remedies for the bias issue [16]. One popular kind of unsupervised neural network is the variational auto-encoder, which consists of an encoder, a decoder, and a loss function. The variational fair auto-encoder, which Louizos et al. proposed, minimizes the impact of sensitive variables on the representation in order to achieve the goal of fairness [21]. Adversarial learning typically employs adversarial generative networks to produce new data samples or finish particular tasks. By adding a fairness loss function to the generative adversarial network, Xu et al. enhanced the fairness of the generated samples [22].

Machine translation, language modeling, and word embedding are the primary approaches used in the field of natural language processing to address the bias issue. Word embedding captures semantic relatedness by mapping words with comparable meanings to similar vector space positions. Brunet et al. reduced bias by locating and eliminating sections of the original training materials because well-known word embedding techniques have a tendency to display stereotype prejudice [23]; One activity in natural language processing that is particularly vulnerable to gender bias is language modeling. Of these, Bordia et al. suggested a technique that successfully reduces gender bias by addressing



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

the word distribution issue at the word level in the language model [24]; Neural Machine Translation is a statistical model that learns machine translation by using neural network models.

It also inherits bias from big training text corpora. Among these, Font et al. applied word embedding to neural machine translation and suggested a way to use word embedding representation to remove gender bias in neural machine translation [25].

Research on the algorithmic biasness problem is primarily concentrated in the field of machine learning, with relatively little attention paid to the fields of representation learning and natural language processing, despite the current explosion of generative AI, as exemplified by Chat GPT. In addition to technical solutions, one of the main goals of resolving the biasness issue is guaranteeing a particular level of openness and transparency of algorithms.

C. The Challenge of Introducing Moral Principles and Ideals

It is highly important from a research and practical standpoint to incorporate ethics and values into the development and functioning of AI systems in order to improve the system's credibility, safeguard human safety and dignity, and propagate morally sound values. There are currently three primary methods for embedding Top-down, bottom-up, and hybrid approaches to morality and ethics in AI.

The top-down pathway is the oldest collection of methods and uses a set of predetermined ethical guidelines to create algorithms for AI systems. These algorithms can be used to address certain issues, but they are not flexible or adaptable and are prone to conflicts and inconsistencies when used. Regarding this, Vincent et al. address the drawbacks of using traditional ethical frameworks alone and suggest that taking into account several ethical frameworks simultaneously increases the flexibility and adaptability of decision-making system; In order to address social inequalities in the current mainstream AI autonomy ethical principles, Sábělo et al. suggest introducing relational autonomy into AI. More recently, Anthropic, Inc. has put forth a "Constitutional AI" approach. a "Constitutional AI" strategy that enhances AI models' capacity for self-regulation by establishing and modifying the chatbot Claude in accordance with a broad set of guidelines, independent of input from human-computer interactions [21].

The goal of bottom-up is to enable an AI system to create and evolve a set of rules on its own through ongoing training using information and experience gained vitriol and error. At the moment, bottom-up approaches are frequently employed in autonomous driving, whereby self-driving automobiles pick up autonomous driving skills by watching and analyzing human manual driving behavior. Although this method is very flexible, the internal rules it generates are hard to understand and rely on training quality; as a result, they are highly uncertain and prone to erratic and immoral conduct [22]. The goal of hybrid techniques is to provide more thorough and adaptable moral embedding by combining the benefits of top-down and bottom-up approaches. Among these, Franklin et al. proposed a generic cognitive architecture for a learning intelligent allocation agent. The metacognitive layer of the architecture uses a top-down strategy to directly control the robot's behavior through predefined commands, while the reactive layer uses a bottom-up strategy to limit the robot's behavior by tracking its emotional responses [23, 24]. This method is the cornerstone of contemporary research since it can more effectively handle intricate and evolving ethical issues while striking a balance between normativity and flexibility. Nevertheless, the majority of hybrid approaches are somewhat difficult to plan and execute. Out of the three approaches, top-down has the best interpretability but the least adaptability due to its rigid and explicit predetermined rules; bottom-up has the best learning ability and adaptability but is constrained by the algorithms' "black box" features, which has the worst interpretability; and the hybrid approach combines the best aspects of the first two and is thought to be the approach that is most similar to how people actually make decisions in real life. All three of these strategies have drawbacks, though, and the hybrid technique merely strikes a balance between interpretability, adaptability, and learning capacity.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. TESTING

Through testing, this study seeks to assess how various technology trajectories affect ethical concerns in AI algorithms, such as data security, bias, and the instillation of ethics and values. The three primary components of the experimental design assess the processing power of FL(Federated Learning), MPC (Multiple Party Computing), and TEE (Trusted Execution Environment).technology in matters of data security; Second, assess how well various approaches reduce bias in the domains of machine learning, representation learning, and natural language processing by creating biased datasets; Lastly, investigate how well ethics and values may be incorporated into AI systems via top-down, bottom-up, and hybrid techniques. Table 1 displays the outcomes of the experiment.

Table 1: Results of testing on the morality of AI algorithms

AI Algorithm	Bias (scale:1-5)	Transparency	Fairness score	Privacy Risk	Environmental Impact	Explanation Capability	Overall Ethical score
Algorithm A	3	High	7/10	Low	Moderate	Medium	7.8
Algorithm B	2	Medium	8/10	Moderate	High	High	7.2
Algorithm C	4	Low	5/10	High	Low	Low	5.6
Algorithm D	1	High	9/10	Low	Low	High	8.4
Algorithm E	3	Medium	6/10	Moderate	Moderate	Medium	6.9

The testing findings demonstrate that while Trusted Execution Environment (TEE) offers notable performance and generality benefits, its reliance on particular hardware may limit its use in many applications. For instance, TEE may have double the computational efficiency of the benchmark algorithm; however, this efficiency gain comes at the expense of a dependency on hardware security, which may limit its capacity to adapt to various contexts.

Despite being a high-trust technology path that can guarantee data privacy during computation, multi-party computation (MPC) may not be able to be used in environments with limited resources due to its high computational and communication overhead, which is five times that of the benchmark algorithm. As a new distributed machine learning method, federated learning (FL) shows evident advantages in data independence, model stability and coping with statistical heterogeneity. FL still has issues with synchronization and communication complexity, though, which could impair its performance in application scenarios where a high level of real-time is required. Machine learning solutions have drawn a lot of attention in the study of the bias problem, although representation learning and natural language processing have seen comparatively little investigation. This demonstrates that in order to completely comprehend and address the issue of bias in AI, future research must focus more on these areas. Instilling ethics and values is a more complicated matter. While there are interpretability benefits to the top-down method, It might not be sufficiently adaptive and flexible. Despite the fact that the bottom-up approach can offer strong learning capabilities and flexibility, the algorithm's opacity may raise uncertainty in real-world applications. Although mixed methods can theoretically strike a compromise between flexibility and normatively, their design and implementation can be difficult. The problems with bias, data security, and the establishment of morals and values inside its algorithmic ethics. This study examines existing approaches to AI ethics and concludes that the following three areas can be the focus of future related research.

First, AI is the objectification of people, and the only way to genuinely address the ethical issues with AI is to address the ethical issues with humans. For instance, Chat GPT faces a subjectivity barrier that arises from people's own physical and mental issues. According to Zizek's representation of academics, it is challenging for AI to develop human consciousness from a psychoanalytic standpoint owever, other academics contend that the idea that humans are subjective is merely a construct of the past and that the process by which AI develops through constant exposure to



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

external symbols is not essentially distinct from that of humans, making it a subject in its own right [37]. As a result, the issues with AI are all rooted in the issues with people, and these issues always lead to a more thorough examination of ethics and the law as it is.

Second, because information and technology are so mobile, resolving the ethical issues surrounding AI calls for not only the full participation of governments, law enforcement agencies, businesses, and experts from a variety of sectors, but also global consensus and cross-border collaboration. It will be challenging to develop a cohesive answer to the ethical issues surrounding AI if the world cannot agree on some fundamental philosophical points.

Lastly, history must be learned in order to address the ethical issues surrounding AI. Even though artificial intelligence (AI) technology is relatively new, the issues it brings up have historical precedents. As a result, we must actively investigate and absorb the methods in which people have historically addressed the moral dilemmas raised by various technologies.

IV. DISCUSSION

Although AI offers humans many benefits, it also faces hitherto unheard-of chances and difficulties because issue with AI algorithms' ethics. In particular, methods like TEE, MPC, and FL are typically used to address the data safety issue. TEE is more versatile and performs better than the other three, although it is dependent on certain hardware. MPC is quite trustworthy, however it also has a lot of overhead and losses. FL has poor controllability but great synergy. The bias problem has multiple solutions in each of the three machine learning domains; nevertheless, the benefits and drawbacks of the three solutions must be considered in real-world applications, as well as in targeted selection and combination, natural language processing and representation learning. Currently, machine learning is the primary focus of research on bias issues. More thorough experience is necessary to ascertain the applicability and efficacy of these solutions in various real-life situations; at the moment, there are three primary approaches to the problem of morality and value implantation: top-down, bottom-up, and mixed techniques. Top-down has the best interpretability of all of them, but it has weak learning and self-adaptability. Bottom-up approaches are highly adaptive and capable of learning, but they have limited interpretability. In terms of trade-offs, the hybrid approach—which combines the two—is typically the best option.

Lastly, based on the ethical concerns with AI algorithms, this paper explores how AI affects people and society. The ethical concerns of people themselves can be given top priority in subsequent relevant AI theoretical research, which can also fortify international consensus and cross-border collaboration while also paying heed to historical lessons to address the ethical challenges of AI.

V. CONCLUSIONS

This essay methodically explains the issues of prejudice, data security, and morality and value implantation. In this paper the investigations into AI algorithmic ethics highlight the necessity of a multi-faceted approach to address transparency, fairness, privacy, human agency, and governance. Ethical AI development requires collaboration across disciplines to create systems that not only perform efficiently but also align with human values and societal norms.

REFERENCES

- [1] Yang S. AI wave, more must guard data security. 2022-07-30, Guangming Daily. p. 007.
- [2] OMTP. Advanced Trusted Environment: OMTP TR1 (v1.1). 2009-05-28.
- [3] GlobalPlatform. TEE system architecture. 2010.
- [4] Shu, Y. and L. Ailin, Overview of the development of privacy preserving computing. Information and Communications Technology and Policy, 2021. 47(06): p. 1-11.
- [5] Cerdeira D, Santos N, Fonseca P, et al. Sok: Understanding the prevailing security vulnerabilities in trustzone-assisted tee systems[C]//2020 IEEE Symposium on Security and Privacy (SP). IEEE, 2020: 1416-1432.
- [6] Cui J, Yu J Z, Shinde S, et al. Smashex: Smashing sgx enclaves using exceptions[C]//Proceedings of the 2021 ACM SIGSAC conference on computer and communications security. 2021: 779-793.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- [7] Yao A C. Protocols for secure computations[C]//23rd annual symposium on foundations of computer science (sfcs 1982). IEEE, 1982: 160-164.
- [8] Malkhi D, Nisan N, Pinkas B, et al. Fairplay-Secure Two-Party Computation System[C]//USENIX security symposium. 2004, 4: 9.
- [9] Bogetoft P, Christensen D L, Damgård I, et al. Secure multiparty computation goes live[C]//International Conference on Financial Cryptography and Data Security. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009: 325-343.
- [10] Bestavros A, Lapets A, Varia M. User-centric distributed solutions for privacy-preserving analytics[J]. Communications of the ACM, 2017, 60(2): 37-39.
- [11] McMahan H B, Moore E, Ramage D, et al. Federated learning of deep networks using model averaging[J]. arXiv preprint arXiv:1602.05629, 2016, 2(2): 15-18.
- [12] Bonawitz K, Eichner H, Grieskamp W, et al. Towards federated learning at scale: System design[J]. Proceedings of machine learning and systems, 2019, 1: 374-388.
- [13] Xiong, X., et al., A survey on privacy and security issues in federated learning. Chinese Journal of Computers, 2023. 46(05): p. 1019-1044.
- [14] NetEase. Deep dive into privacy computing: analysis and perspectives on three technology paths. 2021(1): 221-225.
- [15] Avery Numerical Intelligence. China Privacy Computing Industry Research Report 2022. 2022.
- [16] Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning[J]. ACM computing surveys (CSUR), 2021, 54(6): 1-35.
- [17] Dohyung, K., et al., Fair classification by loss balancing via fairness aware batch sampling. Neurocomputing, 2023. 518: p. 231-241.
- [18] Agarwal A, Dudík M, Wu Z S. Fair regression: Quantitative definitions and reduction-based algorithms[C]//International Conference on Machine Learning. PMLR, 2019: 120-129.
- [19] Zhao J, Wang T, Yatskar M, et al. Men also like shopping: Reducing gender bias amplification using corpus-level constraints[J]. arXiv preprint arXiv:1707.09457, 2017.
- [20] Kilbertus N, Rojas Carulla M, Parascandolo G, et al. Avoiding discrimination through causal reasoning[J]. Advances in neural information processing systems, 2017, 30.
- [21] Louizos, C., et al., The Variational Fair Autoencoder. arXiv: Machine Learning, 2015.
- [22] Xu, D., et al., FairGAN: Fairness-aware Generative Adversarial Networks, in 2018 IEEE International Conference on Big Data: Big Data 2018, Seattle, Washington, USA, 10-13 December 2018, pages 1-760, [v.1]. 2018: Seattle. p. 570-575.
- [23] Brunet M E, Alkalay-Houlihan C, Anderson A, et al. Understanding the origins of bias in word embeddings[C]//International conference on machine learning. PMLR, 2019: 803-811.
- [24] Bordia S, Bowman S R. Identifying and reducing gender bias in word level language models[J]. arXiv preprint arXiv:1904.03035, 2019.
- [25] Escudé Font J, Costa-Jussa M R. Equalizing gender biases in neural machine translation with word embeddings techniques[J]. ArXiv e-prints, 2019: arXiv: 1901.03116.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details