

Detecting Hate Speech Utilizing Deep Convolutional Network and Transformer Models

Responsible AI for the Public Sector: A Blueprint for Ethical Governance

Edge AI in AR/VR: exploring opportunities and confronting challenges.

Building scalable header platforms for large scale ecommerce websites

Upcoming Live Events  
(Detail inside)

*Ask me Anything (AMA) on Big Data Technologies and Applications*  
Tuesday, November 7th, 2023, 6:00 PM

*The Hotstate Machine – A runtime loadable microcoded algorithmic state machine*  
Tuesday, December 5th, 2023, 6:00 PM

## Editor

Meenakshi Jindal

## Chair

Vishnu S. Pendyala

## Vice Chair

John Delany

## Secretary

Sujata Tibrewala

## Treasurer

SR Venkataraman

## Webmaster

Paul Wesling

## Website & Media

<https://r6.ieee.org/scv-cs/>  
<https://www.linkedin.com/company/78437763/>  
<https://www.linkedin.com/groups/2606895/>  
<https://www.facebook.com/IEEEComputerSocSCVchapter>  
<https://twitter.com/IEEEComputerSoc>

## Mailing List

<http://listserv.ieee.org/cgi-bin/wa?SUBED1=cs-chap-scv&A=1>

## Please note:

Feedforward is published quarterly by the Santa Clara Valley (SCV) of the IEEE Computer Society (CS), a non-profit organization. Views and opinions expressed in Feedforward are those of individual authors, contributors and advertisers and they may differ from policies and official statements of IEEE CS SCV Chapter. These should not be construed as legal or professional advice. The IEEE CS SCV Chapter, the publisher, the editor and the contributors are not responsible for any decisions taken by readers on the basis of these views and opinions. Although every care is being taken to ensure genuineness of the writings in this publication, Feedforward does not attest to the originality of the respective authors' content.

All articles in this magazine are published under a Creative Commons Attribution 4.0 License. For more information, see <https://creativecommons.org/licenses/by/4.0/>



Dear Readers,

## From the Editor's Desk

Welcome to the fourth edition of Volume 2 of Feedforward, the esteemed flagship publication of the IEEE Computer Society, Santa Clara Valley chapter. Within these pages, we aim to not only inform but also inspire our readers, offering fresh perspectives and innovative ideas.

As we step into the upcoming quarter with great anticipation, we're thrilled to present an array of exciting events that will kindle your enthusiasm for technology and innovation. For detailed information about each event, we encourage you to explore the magazine's pages. Join us on this captivating journey of discovery and progress!

Don't miss our "Ask me Anything (AMA) on Big Data Technologies and Applications," scheduled for Tuesday, November 7th, 2023, at 6:00 PM. This event is your opportunity to delve into the world of big data, learn from experts, and gain valuable insights into the latest applications and technologies. Looking further ahead, mark your calendars for the "Hotstate Machine – A runtime loadable microcoded algorithmic state machine" event on Tuesday, December 5th, 2023, at 6:00 PM. This promises to be a fascinating exploration of advanced algorithms and state-of-the-art technologies that you won't want to miss.

These upcoming events are a testament to our commitment to providing valuable resources and fostering networking opportunities for our members. Join us in this exciting journey of discovery and advancement!

In reflecting on the past quarter, we fondly remember our Chapter Open house, Awards Ceremony, and the engaging talk on AI and Conversational Commerce held on Tuesday, September 5, 2023, at 6:30 PM (PT). During this event, Raghu Suram, a distinguished Senior Manager, shared insights on Artificial Intelligence, chatbots, virtual assistants, and their impact on e-commerce.

I'm delighted to share the wonderful news that I've been honored with the Industry Rising Star award. This recognition fills me with immense gratitude for the support and opportunities that have brought me to this point. Thank you to everyone who has been a part of this journey.

As we embark on this journey through the latest edition of Feedforward, we're excited to share not only our featured articles but also the upcoming events that promise to fuel your passion for technology and innovation. In this edition, we have gathered a compelling collection of articles that delve into some of the most exciting and impactful topics in the tech industry. Our doors are always open. We extend a heartfelt invitation to join our mailing list, follow our social media channels, and actively participate in our events. Together, we will continue to nurture connections and provide valuable resources to our members, propelling our chapter into a vibrant future for the technology community.

We trust that you will thoroughly relish this edition of Feedforward and eagerly anticipate engaging with you throughout this exhilarating quarter.

## Submit Articles

<https://r6.ieee.org/scv-cs/?p=2036>

## Stay updated, of upcoming events.

<https://r6.ieee.org/scv-cs/category/upcoming-events/>

## View past events on IEEE.tv and on YouTube

<https://ieeetv.ieee.org/search?q=scv-cs>  
<https://www.youtube.com/playlist?list=PLLsxQYv4DdJlYcGPwqUJsnHmfqMtB3eSj>

With every best wish,

[Meenakshi Jindal](#)

San Jose, California, USA

Monday, October 16, 2023

# Detecting Hate Speech Utilizing Deep convolutional Network and Transformer Models

Utkarsh Mittal, *Manager of Machine Learning and Automation, Gap Inc.*

**Abstract**—Online social networks exhibit a significant prevalence of hate speeches, which poses a potential threat to the society and fosters targeted animosity towards specific communities and authorities. Although online platforms are available to automate few mechanisms of hate speeches but classification w.r.t. to different domains and their accuracy are the big issues, and are challenging the researchers, media, and the academic world. The present study addresses the identifying of Hate Speeches through the comparative analysis of the classification efficacy and model intricacy of four distinct Deep Neural Network models; namely CNN (baseline), bidirectional LSTM with attention, pretrained BERT, and fine-tuned RoBERTa transformer models, and utilizing a ternary classification system (hate, offensive, non-hate). The performance of the subject under consideration was assessed through the application of Accuracy, F1-score, and Matthew's correlation coefficient (MCC) metrics on the test set.

**Keywords:** Hate Speech, Transformer, CNN, LSTM, BERT, MCC.

In today's society, where the people's mind is often cluttered with negative influences, and the environment is fiercely competitive, online social networks (OSNs) and microblogging websites are increasing their popularity and attracting more internet users by surpassing all other websites. These platforms allow individuals to exchange their ideas and involve in open discussions, which sparks conflict for widespread hate speech. It refers to the use of aggressive, violent, or offensive language to target a specific group of people, may be based on gender, ethnicity, race, beliefs, or religion. This is a global issue exacerbated by the rise of OSNs, where the people give controversial statements, become more aggressive and try to prove their actions/speeches to become overnight popular.

Despite the prohibitions of hate/conflict words in the speeches on any platform, the content moderation and filtrations remain a big challenge owing to the volume and diversity of posts, comments, and messages exchanged. Also, the inherent ambiguity of speeches makes them difficult to distinguish between hate, sarcasm, and humor in each sentence/ or word [1]. For e.g., the three sentences available on the twitter, creates lot of confusion because of the ambiguity in the word meaning in the sentence.

within a sentence by identifying positive or negative words or expressions.

- Hey dummy, it has been a while since we last read one of your comments.
- I hate that this team will lose all the time.
- I hated these foreigners.

The first tweet happens to be a sarcastic joke between two friends, the second tweet may be an expression of frustration, while the third tweet may be considered an example of xenophobic hate speech. It has been observed from the literature survey that numerous models generally classify the regular content as hate speech, leading to a challenging dilemma: a noteworthy obstacle arises from the reality that of- fenders deliberately obscure certain terms by intention- ally misspelling and introducing some dual meaning words which are not recognized by vocabulary-based models [2][3].

During the survey it was also observed that the identification methods generally rely on dictionaries of offensive terms or phrases and n-grams. However, this technique led a misclassification of the second tweet where the hate has been included in the sentence as a matter of frustration. Sentiment analysis faces a similar constraint, as it aims to determine sentiment polarity



## RELATED WORK

The principles of computational linguistics are based on the comprehension of the inherent characteristics of language. These characteristics are classified into four primary domains: language modeling, morphology, parsing, and semantics. Language modeling encompasses the implicit syntactic and semantic relationships between words or elements. Morphology focuses on identifying word segments such as roots, stems, prefixes, and suffixes [4][5]. Parsing involves the examination of the relationships between various words and phrases [19]. Semantics, on the other hand, focuses on the understanding of the meanings conveyed by words, phrases, and sentences [6].

The detection of writing patterns has been found to be useful in identifying sarcasm, conducting multi-class sentiment analysis, and measuring sentiment [7][8][9][10]. The identification of hate speech on Twitter was the focus of a study conducted by researchers, who employed unigram-, sentiment-based, and semantic features. Sentiment-centric features encompassed various factors, including the evaluation of positive and negative words, the presence of slang terms, the usage of emojis, and the inclusion of hashtags. The semantic attributes encompassed the quantification of exclamation, question, and period symbols, uppercase words, quotations, exclamatory phrases, expressions denoting laughter, and terms commonly found in tweets. The part-of-speech (PoS) tags assigned to nouns, verbs, adjectives, and adverbs were subject to careful examination. Patterns were employed to associate sentiment and non-sentiment words with Part-of-Speech (PoS) tags [6].

Recently, numerous investigations have focused on the application of transformer-based architectures such as BERT and RoBERTa for the identification of hate speech. The paper titled "Hate Speech Detection with BERT using PyTorch Lightning" by Budhraj and Agarwal, where a BERT-centric model has been suggested for recognizing the hate speech on a social media platform [20]. Another paper titled "RoBERTa: A Robustly Optimized BERT Pre-training Approach" by Liu et al., where the authors have proposed RoBERTa, a pre-training methodology to secure better performance across various NLP tasks helping in detecting the hate speech [11].

## DATA

The Hate Speech and Offensive Language dataset provided by Twitter for the Kaggle Competition is used in the present study [12]. This dataset contained over

25,000 tweets with a three-tier classification system: hate, offensive, and neither. The raw data necessitated

several stages of curation and processing, including eliminating usernames beginning/ending with @, eradicating http links, hashtags, RT, special characters except for the apostrophe ('), emojis, and words with less than three characters [20][3]. The Stemming technique has been applied to remove word suffix and to reduce them to their roots, and Lemmatization is implemented, to account the context and convert the word to its base root form [13][5].

## APPROACH

Four different models are explored to tackle the problem of "Hate" vs. "Offensive" vs. "Neither" speech detection.

### Baseline CNN model

The model architecture includes an embedding layer, a convolutional layer, a max pooling layer, and a dense layer featuring dropout. The convolutional layer operates with 128 filters or kernels, each with a size of 3. Max pooling layers are integrated throughout the model for abstraction and prevention of overfitting. This one-dimensional convolutional structure employs pre-trained GloVe word embeddings [14][15]. Post word-embedding, the input sequence adheres to a maximum sequence length of 100. The data undergo a series of three convolutions and max-pooling processes, with the last pooling layer uses global max-pooling. The model's structure is compact, incorporating a dropout rate of 0.2 to prevent overfitting. A sigmoid activation function is used to determine the probability of three binary classifications. Data is divided into training, and test sets in a ratio of 0.8:0.2. The experimental framework involves a batch size of 128, an RMSProp optimizer with a learning rate of 0.01, a 20% dropout rate, and epochs ranging from 10 to 100. Maximum sequence length of 100 and vocabulary size of 20,000 and embedding dimensions of 100 and 200 are tested.

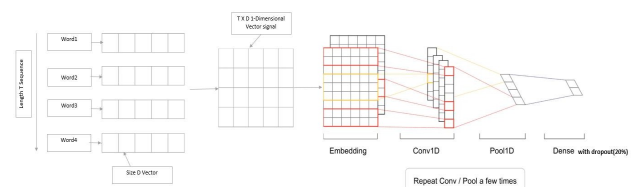


FIGURE 1. Convolutional Neural Network with Pooling la



## Bidirectional LSTM model with attention layer

The Bi-LSTM model utilizes a deep neural network architecture for text classification, with an input composed of token sequences of varying lengths. This model features an embedding layer, followed by a pair of bidirectional LSTM layers, each containing 128 units, a dropout rate of 0.2 [8][16][17]. The BiLSTM layers analyzed the input sequence in both directions, thereby enabling the capture of contextual data from adjacent words. The incorporation of two bidirectional LSTM layers enhances the capacity of the model to identify intimate and abstract connections between words within an input sequence, thus elevating its performance in text classification tasks. Subsequently, an attention layer is included to calculate the attention scores for each word in the input sequence based on its relevance to the classification endeavor. Following the attention layer, a dropout layer with a 0.2 rate is introduced to prevent overfitting. A flattened layer is then added to transform the 3D output from the preceding layers into a 2D output. Finally, a dense layer with an activation function is incorporated to generate a probability distribution across the three potential classes. The model architecture employs glove word embedding with a maximum sequence length of 50 and a maximum vocab size of 50,000. A train/test division of 0.8 and 0.2 has been adopted. During the model's training, the loss function employed is categorical cross-entropy, and the optimization algorithm used is Adam.

The BiDirectional LSTM model may have demonstrated overfitting due to its disproportionately high capacity compared to the available training data. Possible remedies could include minimizing the number of layers or reducing hidden units to 64 or 32, which could help in curtailing this overfitting issue. In addition, the inclusion of regularization terms such as L1 or L2 to the loss function or simplifying the model's complexity could also serve as a potential solution. The training data could possess high variance or noise, which may contribute to the overfitting. By gathering more high-quality training data, this problem can be mitigated.

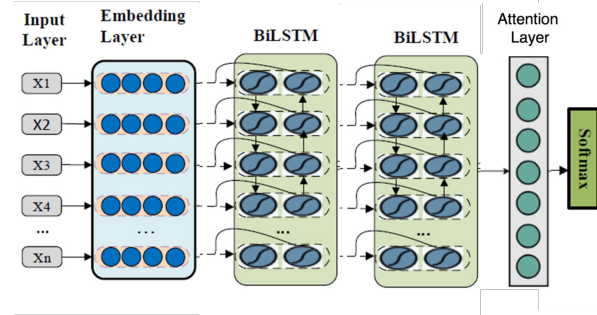
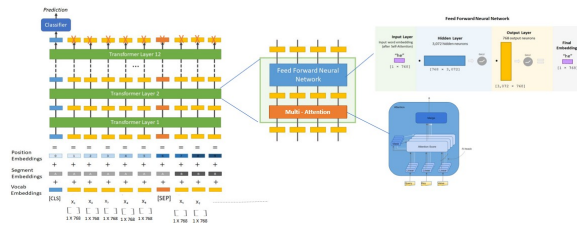


FIGURE 2. Bidirectional LSTM with attention layer.

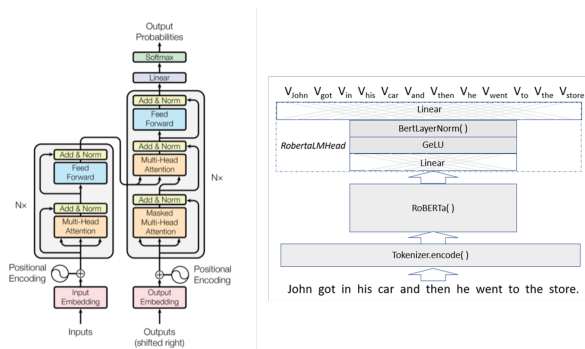
Overfitting might also be a result of excessive training epochs. To prevent this, early termination based on a validation set is advised. Lastly, employing data augmentation techniques such as noise injection could enhance the model's robustness and further prevent overfitting.

## Pretrained Bert with Adamw optimizer Transformer model

The model architecture consists of 12 transformer block layers, featuring a hidden size of 768 (embedding size), 12 self-attention heads, and approximately 110 million trainable parameters [18][19][5][20]. A hugging-face Bert sequence classification library is used for this purpose. The 12 layers' weights are trained, whereas the classification-layer weights are randomly initialized. In the initial layer, the embedding size is set to 768 and the maximum token length as 512. The Adam optimizer with a weight decay fix is used to update the weights during the training loop. The model runs for ten epochs with a batch size of 32 and early stopping is implemented to prevent overfitting. In addition, a scheduler has been incorporated to leverage the learning rate decay (allowing for larger initial steps and smaller subsequent steps). A train/test split of 0.8 and 0.2 is employed. A training process involving categorical cross-entropy and K-fold cross-validation is applied for hyperparameter tuning. The mean performance variation across different results is considered more reliable than relying on a potentially biased validation set, which could overestimate or underestimate the performance. In this model, various experiments are conducted using Glove and FastText embeddings, and FastText significantly outperformed GloVe.



**FIGURE 3.** Pre-trained Bert with Adam optimizer with weight adjusted.



**FIGURE 4.** Ternary Finetuned Roberta Transformer

## Ternary Finetuned RoBERTa Transformer Model

The RoBERTa-Large configuration employed a transformer framework consisting of 24 layers, a hidden size of 1024, and 16 Attention Heads with a Head size of 64 [19][11]. It undergoes training using a vast corpus that includes BOOK-CORPUS and English WIKIPEDIA, featuring a batch size of 8k, dropout rate of 0.1, and Adam optimization without gradient clipping. This model was fine-tuned using 25K ternary-labeled data. A train/test division of 0.9 and 0.1 has been employed to refine the RoBERTa model across 22,272 instances. The batch size of 32, accompanied by a learning rate of  $2e-5$  for the 355M trainable parameters is used. The model trained over three epochs and uses early stopping to prevent overfitting. FastText embeddings (which represent each word as an n-gram of characters instead of directly learning word vectors) are employed, adhering to a maximum sequence length of 100 following the word embedding.

## RESULTS

In all the scenarios, the information has been divided into training and test sets. A range of metrics, such as Accuracy, Precision, Recall, and F1 Score, are used to evaluate the efficacy of the algorithms in each situation.

Given the dataset's imbalanced characteristics, with one label making up 77% of the text, the Matthews Correlation Coefficient (MCC) is implemented to guarantee a more accurate assessment of the algorithms' performance.

The table presents the performance comparison of various models in terms of precision, recall, F1-score, accuracy, and MCC. The models included a baseline CNN model (Stage1,2) with different optimizers, a Bidirectional LSTM model with an attention layer, a pre-trained Bert model with an Adam optimizer, and a ternary fine-tuned RoBERTa transformer. The models are trained and evaluated using both training and validation datasets.

The baseline CNN model employing the Adam optimizer attained a precision of 0.851 and recall of 0.828 in the training set, along with an F1-score of 0.839 and MCC of 0.730. The validation set achieved a precision of 0.760, recall of 0.739, F1-score of 0.749, and MCC of 0.650. The baseline CNN model using the RMSProp optimizer demonstrated marginally lower scores on both sets, suggesting a somewhat inferior performance compared with the variant using the Adam optimizer.

The Bidirectional LSTM model with an attention layer, trained via the Adam optimizer, displayed remarkable performance in the training set, obtaining high precision, recall, F1 score, and MCC. However, the performance declined in the validation set. It is important to highlight that the bidirectional LSTM model with the attention layer exhibited overfitting, as its performance deteriorated in the validation set relative to the training set.

The pretrained Bert model with the Adam optimizer consistently yielded outstanding outcomes in both the training and validation sets, exhibiting high precision, recall, F1-score, and MCC. Similarly, the ternary fine-tuned RoBERTa transformer leveraging the Adam optimizer exhibited robust performance across all metrics in both the training and test sets.

## CONCLUSION

The study focusses on four distinct NLP techniques to categorize large datasets of tweets into Offensive, hate speeches and neither to check the efficacy and effectiveness of the sentences/words. The models are evaluated using multiple metrics namely Precision, Recall, F1-score, and MCC and their performance has been compared thoroughly. The two transformer-based models, BiDirectional LSTM

with and Pre-trained BERT, exhibits almost similar outcomes with no indication of overfitting, whereas the Fine-tuned RoBERTa Transformer displayed the best performance. Outcome of the study indicates that transformer-based models are effective for addressing and identifying the hate speeches and meeting the complex challenges of any online social platforms. Thus, it is concluded that performance of detection needs improvement despite no of challenges/limitations of the datasets and /or information transformation systems. For e.g.,

1. the data frame is based on the limited datasets and labels.,
2. Limitation of text classification and representation,
3. BERT is pretrained on large, biased datasets, w.r.t. to specific demographic and perspectives. It is noticed that language and expressions when framed with biasness, struggle for detections of speeches as they are meaning differently in a complex domain.

**TABLE 1.** Comparison Performance metrics of Different Models

Model	Parameters	Train/ Test	Pre- cision	Re- call	F1- Score	MCC
Baseline CNN	EmbSize=100, Dropout, Adam	Train	0.80	0.82	0.80	0.73
		Test	0.76	0.73	0.74	0.58
Baseline CNN Model	EmbSize=100, Dropout, RMSProp	Train	0.78	0.77	0.78	0.69
		Test	0.76	0.76	0.76	0.62
BiDirectional LSTM With Attention	EmbSize=100, Dropout, Adam	Train	0.94	0.94	0.94	0.84
		Test	0.89	0.90	0.89	0.72
Pre-trained Bert with Adam	EmbSize=100, Dropout, Adam	Train	0.90	0.91	0.90	0.81
		Test	0.91	0.92	0.91	0.79
Ternary Finetuned RoBERTa	EmbSize=100, Dropout, Adam	Train	0.92	0.91	0.92	0.84
		Test	0.91	0.90	0.90	0.82

## ACKNOWLEDGEMENTS

Thanks to all the reviewers for their valuable suggestions which helped in improving all the sections of the manuscript.

## REFERENCES

1. W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proceedings of the second workshop on language in social media*, 2012, pp. 19– 26.
2. N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 4, pp. 215– 230, 2015.
3. A. V. Yenikar and C. N. Babu, "Sentimlbench: Benchmark evaluation of machine learning algorithms for sentiment analysis," *Indonesian Journal of Electrical Engineering and Informatics (IJEI)*, vol. 11, no. 1, pp. 318–336, 2023.
4. P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," *arXiv preprint arXiv:1806.03822*, 2018.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186. [Online]. Avail- able: <https://www.aclweb.org/anthology/N19-1423/>
6. H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE access*, vol. 6, pp. 13 825– 13 835, 2018.
7. M. Bouazizi and T. O. Ohtsuki, "A pattern-based approach for sarcasm detection on twitter," *IEEE Access*, vol. 4, pp. 5477–5488, 2016.
8. Hamish, "Bidirectional lstm in keras with glove embeddings," 2019. [Online]. Avail- able: <https://medium.com/@hamishvb/bidirectional-lstm- in- keras- with- glove- embeddings- 4be01b4c4fca>
9. D. Davidov, O. Tsur, and A. Rappoport, "Semi-supervised recognition of sarcasm in twitter and amazon," in *Proceedings of the fourteenth conference on computational natural language learning*, 2010, pp. 107–116.
10. O. Tsur, D. Davidov, and A. Rappoport, "lcwsm—a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, no. 1, 2010, pp. 162– 169.
11. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining ap- proach," *arXiv preprint arXiv:1907.11692*, 2019.
12. A. Samoshyn, "Hate speech and offensive language dataset," 2020. [Online]. Available: <https://www.kaggle.com/datasets/vkrahul/twitter-hate- speech> J. Hartmann, M. Heitmann, C. Schamp, and O. Net- zer, "The power of brand selfies," *Journal of Market- ing Research*, 2015.



13. L. M. Kitchell, "Convolutional neural networks," arXiv preprint arXiv:1809.02202, 2018.
14. J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
15. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
16. M. Saeed, "Adding a custom attention layer to a recurrent neural network in keras," 2022. [Online]. Available: <https://towardsdatascience.com/adding-a-custom-attention-layer-to-a-recurrent-neural-network-in-keras-60d3a8c6d7a0>
17. T. Zhang, Y. Jiang, X. Wang, Y. Huang, and H. Zhao, "A robustly optimized bert pre-training approach with post-training," in Neural Information Processing, L. Chen, Q. Wang, X. Hong, J. Liu, Y. Shi, and F. Wang, Eds. Springer International Publishing, 2020, pp. 353–362.
18. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
19. R. Budhraj and B. Agarwal, "Hate speech detection with bert using pytorch lightning," arXiv preprint arXiv:2010.12688, 2020.

**Utkarsh Mittal** is a Machine Learning manager at Gap Inc., a global retail company, has more than ten years of practice experience in machine learning automation and is a leader of big AI-based database projects. He received his Master's in Industrial Engineering with a Supply Chain and Operations Research major from Oklahoma State University, USA. He is closely associated with research groups and editorial boards of high-profile International Journals, and research organizations, and is passionate about solving complex business challenges and encouraging innovation through upcoming technologies.; He is a member of IEEE Computer Society. Contact him at [mittalutkarsh@gmail.com](mailto:mittalutkarsh@gmail.com).


### Upcoming Event.

**Ask me Anything (AMA) on Big Data Technologies and Applications**  
Virtual Event via Zoom and YouTube live.

Date: Tuesday, November 7th, 2023, 6:00 PM Free Registration:

Registration: <https://r6.ieee.org/scv-cs/ask-me-anything-ama-on-big-data-technologies-and-applications/>

Visit [www.youtube.com/channel/UC4kaO4mrTxrCltxb0HalV3Q/channels](https://www.youtube.com/channel/UC4kaO4mrTxrCltxb0HalV3Q/channels)




**Ask me Anything (AMA) on  
Big Data Technologies and Applications**  
... Hadoop, Spark, Mining Streaming data, ...

**Raghavendra K. Chunduri**, *Post Doctoral Fellow in Neuromorphic computing. Department of Electrical and Computer Engineering, University of Colorado, Colorado Springs*

Tuesday, November 7<sup>th</sup>, 2023, 6:00 pm PT (virtual)  
*Via Zoom and YouTube Live*

Register (Free): <https://r6.ieee.org/scv-cs/ask-me-anything-ama-on-big-data-technologies-and-applications/>

**Vishnu S. Pendyala**, Chair  
**John Delaney**, Vice Chair  
**Sujata Tibrewala**, Secretary  
**S.R. Venkatramanan**, Treasurer

 **IEEE  
COMPUTER  
SOCIETY**  
*Santa Clara Valley Chapter*

# Responsible AI for the Public Sector: A Blueprint for Ethical

Lakshmanan Sethu Sankaranarayanan, TAM, Google, Burbank, CA, 91504

**Abstract**—In an era defined by data-driven decision-making and digital transformation, artificial intelligence (AI) stands as a powerful tool with the potential to revolutionize the public sector. The promise of enhanced efficiency, informed policymaking, and improved citizen services through AI adoption is tantalizing. This paper aims to provide a blueprint for responsible use of AI within the public sector, as it is paramount to building trust, safeguarding fairness, and reaping the full benefits of this transformative technology in the public sphere.

The integration of artificial intelligence into the public sector represents a paradigm shift. The allure of predictive analytics, automated administrative processes, and AI-driven decision support systems is undeniable. Yet, as governments embark on this technological journey, they must navigate a complex landscape of ethical, technical, and social challenges. The responsible use of AI is not merely an option but an imperative for the public sectors of today and tomorrow.

## THE POTENTIAL OF AI IN PUBLIC SECTOR

The scope of AI's potential in the public sector is extensive, spanning a diverse array of applications poised to usher in governance that is not only more efficient and effective but also centered around the needs of people. Below, we explore several pivotal domains where AI stands to exert a profound influence within the public sector.

### Predictive Analytics

One of the most promising applications of AI in the public sector is predictive analytics. AI algorithms can analyze vast datasets to predict future events and trends, enabling more effective resource allocation. Here's how it can benefit different areas:

- **Healthcare:** Predictive analytics can help hospitals and healthcare agencies anticipate patient admission rates, disease outbreaks, and medication needs. This enables better allocation of resources, timely response to health crises, and improved patient care.
- **Emergency Services:** Fire departments and emergency responders can use AI to predict the

likelihood of wildfires, floods, or other disasters. This proactive approach allows for resource allocation, evacuation planning, and risk mitigation.

- **Traffic Management:** Cities can use AI to predict traffic congestion, accidents, and optimal traffic signal timings. This reduces commute times, fuel consumption, and environmental impact.
- **Education:** Schools and education departments can use AI to predict student performance and identify students at risk of falling behind. This enables targeted interventions to improve educational outcomes.

### Data Driven Decision-Making

AI has the capacity to revolutionize decision-making processes within government agencies. By analyzing vast datasets and identifying patterns, AI can provide insights that inform policy decisions:

- **Urban Planning:** AI can analyze demographic data, traffic patterns, and environmental factors to inform urban development and infrastructure projects. This leads to more efficient city planning and resource allocation.
- **Criminal Justice:** Predictive policing algorithms can help law enforcement agencies allocate resources more effectively by identifying areas with higher crime rates. This proactive approach aims to reduce crime and improve community safety.
- **Environmental Policy:** AI can process environmental data to model the impact of policy

- decisions on climate change, air quality, and natural resources. This aids in crafting evidence-based environmental policies.
- **Budgeting and Finance:** AI-driven forecasting can assist in budget allocation, revenue prediction, and fraud detection, ensuring government finances are managed efficiently.

## Personalized Services

AI-powered virtual assistants, chatbots, and automated systems can enhance citizen services by providing timely and efficient support:

**Customer Service:** Government agencies can deploy AI-driven chatbots to handle citizen inquiries, provide information, and assist with routine tasks. This reduces the burden on human customer service agents and offers 24/7 support.

**Healthcare Consultation:** Virtual healthcare assistants can offer medical advice, schedule appointments, and provide information on healthcare services, improving access to healthcare for citizens.

**Public Information:** AI chatbots can disseminate information on government programs, policies, and events, ensuring citizens stay informed.

## Public Safety and Security

AI technologies have a crucial role to play in ensuring public safety and security:

- **Predictive Policing:** AI-driven predictive policing can help law enforcement agencies anticipate criminal activity and allocate resources, accordingly, improving public safety.
- **Disaster Response:** AI can analyze real-time data from various sources, including weather satellites and social media, to provide early warnings and guide disaster response efforts during natural disasters.
- **Cybersecurity:** Government agencies can use AI to detect and mitigate cyber threats, protecting sensitive data and critical infrastructure.

## Automated Administrative Processes

AI-driven automation can streamline administrative processes in the public sector:

- **Document Processing:** AI can automate the processing of forms, applications, and documents, reducing bureaucratic overhead and speeding up citizen interactions with government agencies.
- **Data Entry and Management:** Automation can handle data entry, validation, and management, reducing errors and improving data accuracy.
- **Routine Tasks:** Repetitive administrative tasks, such as data entry or appointment scheduling, can be automated, freeing up government employees to focus on more complex and strategic roles.

In summary, AI's potential in the public sector is vast and multifaceted. It has the capacity to optimize resource allocation, enhance decision-making, improve citizen services, bolster public safety, and streamline administrative processes. However, realizing this potential requires careful planning, ethical considerations, and a commitment to responsible AI practices to ensure that the benefits are equitably distributed and that AI technologies serve the greater good of society.

## THE MORAL OBLIGATION

### ISSUES AND DILEMMAS SURROUNDING ETHICAL IMPLEMENTATION OF AI IN THE PUBLIC SECTOR

While the potential benefits of adopting artificial intelligence (AI) in the public sector are profound, it comes with a set of ethical, technical, and practical challenges that governments must address to ensure responsible AI adoption.

### Bias and Fairness

**Challenge:** AI systems may inherit biases from their training data, leading to outcomes that exhibit discrimination. Within the public sector, these biased decisions have the potential to sustain social disparities, fortify historical prejudices, and adversely affect marginalized communities.

**Imperative:** Ensuring fairness in AI algorithms is paramount. Governments and agencies must invest in rigorous testing and validation processes to identify and mitigate bias. This includes regularly auditing AI systems, refining training data, and implementing fairness-aware algorithms to promote equitable outcomes.



## Transparency and Accountability

**Challenge:** AI systems frequently function as enigmatic 'black boxes,' rendering it difficult for citizens and stakeholders to grasp the decision-making mechanisms driving AI-powered choices. This opacity in operation can undermine trust and provoke apprehensions regarding accountability.

**Imperative:** Establishing transparency and accountability mechanisms is essential. Governments should demand transparency in AI systems and provide clear explanations of AI decisions to citizens. Open-source AI models and algorithms where possible to allow for public scrutiny. Additionally, appoint AI ethics boards or officers to oversee responsible AI practices.

## Privacy and Data Protection

**Challenge:** The collection and processing of vast amounts of citizen data for AI applications raise legitimate privacy concerns. Mishandling sensitive information can lead to privacy breaches and erode public trust.

**Imperative:** Ensuring robust data privacy and security measures is imperative. Government entities must prioritize actions such as data encryption, the enforcement of stringent access controls, and strict adherence to data protection regulations like GDPR. Additionally, the establishment of data anonymization techniques and consent mechanisms is essential to safeguarding citizen privacy while facilitating the responsible use of AI applications.

## Job Displacement

**Challenge:** The automation potential of AI may lead to job displacement in the public sector, particularly in roles that involve routine tasks. Addressing the workforce implications of AI adoption is crucial.

**Imperative:** Governments must proactively address job displacement by investing in workforce reskilling and job transition support programs. These initiatives can help affected employees acquire new skills and transition to roles that are complementary to AI systems, fostering a smooth and inclusive transition.

## Accountability for AI Decisions

**Challenge:** AI-driven decisions may lack a clear entity responsible for the outcomes. When AI systems make mistakes or cause harm, determining accountability can be challenging.

**Imperative:** It is crucial for governments to delineate unambiguous lines of accountability concerning AI decisions. This encompasses the definition of precise roles and responsibilities for both individuals and agencies responsible for the oversight of AI systems. Furthermore, comprehensive legal frameworks should be established to ensure that responsible parties can be held accountable for the outcomes stemming from AI deployments.

## Ethical Dilemmas and Trade-offs

**Challenge:** The responsible use of AI often involves ethical dilemmas and trade-offs. Balancing priorities such as privacy, security, and efficiency can be complex and requires careful consideration.

**Imperative:** Government agencies should engage in ethical deliberation and conduct impact assessments before implementing AI systems. Ethical guidelines should be developed to help agencies navigate these complexities and make informed decisions that align with societal values.

## Public Trust and Perception

**Challenge:** Public perception and trust in AI adoption within the public sector can shape its success. Mistrust or fear of AI can hinder its acceptance and effectiveness.

**Imperative:** Fostering and upholding public trust stands as an absolute necessity. Governments should engage in transparent communication regarding their AI initiatives, educate the public about the advantages and protective measures in place, and actively include citizens in AI decision-making processes to cultivate a sense of ownership and inclusivity.

The responsible adoption of AI in the public sector transcends mere choice; it is an ethical imperative. Governments bear the responsibility of ensuring that AI technologies are harnessed for societal benefit while minimizing negative impacts and upholding principles of fairness, transparency, and accountability. Confronting these ethical challenges

directly is indispensable for establishing trust and harnessing the full potential of AI for the betterment of both citizens and society at large.

In the following sections, we will explore the principles and strategies that governments can adopt to implement responsible AI in the public sector, ensuring that AI technologies serve the public interest while upholding ethical standards and values.

## RESPONSIBLE AI PRINCIPLES

As governments and public institutions embark on their AI journeys, it's crucial to establish clear principles and guidelines that promote the responsible and ethical use of artificial intelligence. These principles serve as a compass, ensuring that AI technologies are harnessed to benefit society while upholding fundamental values and avoiding potential pitfalls.

### Ethical Guidelines

**Principle:** Develop and implement ethical guidelines that prioritize fairness, equity, and non-discrimination in AI deployment within the public sector.

**Implementation:**

- **Implementing Fairness-Focused Algorithms:** Employ algorithms explicitly engineered to recognize and alleviate biases in decision-making procedures.
- **Balanced Results:** Guarantee that AI systems do not unduly affect specific demographic segments and that outcomes are fair and balanced.
- **Ethical Training Data:** Meticulously select and validate training data to minimize biases, promoting diversity and representativeness.

### Transparency

**Principle:** Promote transparency in AI systems to enable stakeholders, including people, to understand how decisions are made.

**Implementation:**

- **Explainable AI:** Use models and algorithms that offer transparent explanations of their decisions, allowing users to comprehend the rationale behind AI-driven outcomes.
- **Documentation:** Maintain comprehensive documentation of AI systems, including data sources, algorithms used, and decision criteria.
- **Open Source:** Encourage the open-source release of AI models and algorithms to allow for independent scrutiny and auditing.

### Data Privacy and Security

**Principle:** Implement robust data privacy and security measures to safeguard sensitive information and protect citizen rights.

**Implementation:**

- **Information Encryption:** Encrypt sensitive data to prevent unauthorized access and protect confidentiality.
- **Access Controls:** Establish strict access controls to ensure that only authorized personnel can access and process sensitive data.
- **Compliance:** Adhere to data protection regulations such as GDPR, HIPAA, or relevant local laws.
- **People Privacy:** Anonymize data where necessary to protect individual privacy.

### Accountability and Oversight

**Principle:** Establish mechanisms for accountability and oversight of AI systems to ensure responsible and ethical practices.

**Implementation:**

- **Ethics Boards:** Set up AI ethics boards or officers responsible for monitoring and evaluating AI systems' ethical use.
- **Audit Trails:** Maintain comprehensive audit trails of AI-driven decisions and actions for accountability and transparency.
- **Regulatory Compliance:** Ensure that AI systems comply with all relevant laws and regulations.
- **Reporting Process:** Implement mechanisms for reporting AI-related issues, including ethical concerns or violations.

### Education and Training

**Principle:** Invest in educating public sector employees about AI and its ethical use to ensure responsible AI practices.

**Implementation:**

- **Training & Learning:** Develop training programs to upskill government employees who work with AI systems.
- **Ethics Knowledge:** Include ethics training as a core component of AI education programs.
- **Awareness Campaigns:** Raise awareness among employees about the ethical implications of AI and the importance of responsible AI practices.

These principles serve as the cornerstone for the conscientious integration of AI within the public sector. They serve as compass points for governments and public institutions, assisting them in navigating the intricate terrain of AI adoption while upholding ethical standards, transparency, and alignment with societal values. Embracing these principles is not merely a means to engender public confidence but also to cultivate a climate of responsible AI governance, one that has the potential to yield substantial and favorable results for both citizens and society at large.

In the subsequent sections, we will explore the importance of international collaboration in addressing AI challenges and examine real-world case studies of responsible AI implementations in various public sector domains.

## INTERNATIONAL COLLABORATION

The responsible use of artificial intelligence (AI) in the public sector is a global imperative that transcends national borders. To address the ethical, technical, and practical challenges associated with AI adoption and to foster a harmonized, responsible AI ecosystem, international collaboration is essential. Governments, organizations, and stakeholders worldwide must work together to establish global standards, share best practices, and ensure that AI technologies serve the public interest.

### International Standards

Importance: International standards provide a common framework for the ethical and responsible use of AI technologies. Collaborative efforts to develop and adopt these standards ensure that AI systems align with global values and principles.

Initiatives:

- **ISO Standards:** The International Organization for Standardization (ISO) has developed and continues to develop standards related to AI ethics, governance, and safety. Governments should actively participate in these initiatives to shape global AI standards.
- **UN and International Organization:** The United Nations (UN) and other international bodies are exploring AI governance frameworks. Collaboration with these organizations can lead to the establishment of global norms for responsible AI.

### Information Sharing across borders.

Importance: Sharing information and best practices across borders is essential for building collective knowledge and expertise in responsible AI adoption. It enables governments to learn from one another's experiences and successes.

Initiatives:

- **International Workshops and Conferences:** Governments and organizations can participate in international AI workshops and conferences to share insights and case studies on responsible AI implementation.
- **Global AI Communities:** Joining global AI communities and networks fosters collaboration and enables knowledge exchange on responsible AI practices.

### Collaborative Initiatives

Importance: Collaborative initiatives bring together governments, academia, industry, and civil society to address AI challenges collectively. These initiatives facilitate the development of ethical guidelines and best practices.

Initiatives:

- **Global AI Ethics Consortia:** Governments can participate in or establish international AI ethics consortia that focus on developing ethical guidelines and principles for AI use in the public sector.
- **Public-Private Partnerships:** Collaborations between governments and private sector AI providers can promote responsible AI development and deployment. Governments can work with industry partners to establish ethical AI standards.

### Cross-Border Research and Development

Importance: Cross-border research and development collaborations accelerate AI advancements while ensuring that responsible AI principles are integrated into emerging technologies.

Initiatives:

- **International Research Associations Encourage** international partnerships between research institutions and universities to advance responsible AI research.



- Global Innovation Center: Establish global innovation hubs that promote responsible AI development and support cross-border innovation.

Global Cooperation in Responsible AI is Imperative to guarantee that AI advances are dedicated to the welfare of humanity while upholding ethical benchmarks and values. Governments ought to participate actively in international entities, exchange insights and best practices, and collaborate towards shaping a world where public-sector AI innovations benefit people across the globe. Through fostering synergy and joint efforts, the international community can jointly tackle AI hurdles and advocate for responsible AI governance on a worldwide scale.

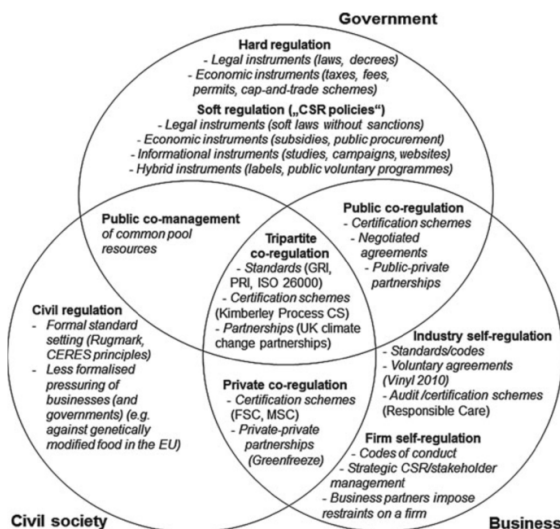


Figure 1. Public Private Partnership

In the following sections, we will delve into real-world case studies that exemplify responsible AI implementations in various public sector domains, providing concrete examples of the principles and collaborative efforts discussed in this article.

## CASE STUDIES

To understand the practical application of responsible AI principles in the public sector, let's explore real-world case studies that showcase how governments are harnessing AI for the benefit of their citizens while upholding ethical standards.

## Healthcare: Predictive Analytics Saves Lives

### Case Study: Predictive Analytics for Disease Outbreaks

In Taiwan, the government has implemented a sophisticated AI-driven system that analyzes vast amounts of health data, including information on patient symptoms, travel history, and environmental factors. This system enables healthcare authorities to predict disease outbreaks, such as flu epidemics of dengue fever, with remarkable accuracy. By identifying potential hotspots and allocating medical resources accordingly, Taiwan has been able to respond swiftly to health crises, saving lives and reducing the economic impact of disease outbreaks.

Responsible AI Practices:

**Ethical Data Handling:** The Taiwanese government ensures the responsible use of citizen data by anonymizing and protecting sensitive information.

**Transparency:** The AI system is transparent, providing clear explanations for its predictions.

**Public Trust:** Transparent communication and collaboration with healthcare experts and the public have built trust in the system.

## Government Decision Support: AI for Informed Policymaking

### Case Study: AI-Powered Decision Support in Singapore

The government of Singapore has integrated AI into its decision-making processes to enhance public policy development. Machine learning algorithms analyze extensive data, including citizen feedback, economic indicators, and environmental data, to provide policymakers with insights and recommendations. This data-driven approach has enabled Singapore to make informed decisions on urban planning, traffic management, and

environmental policy, leading to improved quality of life for its citizens.

Responsible AI Practices:

**Ethical Considerations:** Singapore's government actively considers ethical implications and consults experts when implementing AI in policymaking. Department of Transportation is using AI to develop new bike lanes and pedestrian crosswalks.

**Transparency:** The AI-driven decision support system provides policymakers with transparent insights into the data used to make recommendations. The city has developed a set of guidelines for the ethical use of AI. The guidelines require that AI systems be

**Public Engagement:** Public engagement initiatives ensure that citizens are informed and involved in decisions that affect them. used in a way that is fair, transparent, and accountable. The guidelines also require that AI systems be used in a way that respects the privacy and security of residents.

### Case Study: City of Los Angeles, California

The City of Los Angeles is using AI to improve the efficiency and effectiveness of its public services. For example, the city's Department of Transportation is using AI to optimize traffic flow and reduce congestion. The city's Department of Water and Power is using AI to detect leaks and reduce water waste. And the city's Police Department is using AI to predict crime and improve public safety.

The city has developed a set of principles to guide its use of AI, including transparency, fairness, and accountability.

### Case Study: State of Virginia

The State of Virginia is using AI to improve the delivery of healthcare services. For example, the state's Department of Health is using AI to identify patients who are at risk of developing chronic diseases. The state's Department of Medical Assistance Services is using AI to improve the quality of care for Medicaid recipients. And the state's Department of Veterans Affairs is using AI to help veterans manage their healthcare needs.

The state has developed a set of standards for the responsible use of AI in healthcare. The standards require that AI systems be transparent, fair, and accountable. The state also requires that AI systems be used in a way that respects the privacy and security of patients.

### Case Study: City of New York, New York

The City of New York is using AI to improve the lives of its residents. For example, the city's Department of Education is using AI to identify students who are struggling and provide them with extra help. The city's Department of Housing and Preservation is using AI to identify and repair public housing units. And the city's

### Virtual Assistants: Chatbots for Citizen Services

#### Case Study: Estonia's AI-Powered Virtual Assistants

Estonia, known for its innovative approach to e-governance, has deployed AI-powered virtual assistants to enhance people services. These chatbots provide citizens with 24/7 support, answering queries related to government services, deadlines, and requirements. By automating routine tasks, Estonia's government has streamlined citizen interactions, reduced bureaucracy and improving overall user experiences.

#### Responsible AI Practices:

**Data Privacy:** Estonia prioritizes data privacy, with strict controls in place to protect citizen information.

**Transparency:** Virtual assistants provide clear explanations for their responses, ensuring transparency in AI-driven interactions.

**Language Accessibility:** AI chatbots are designed to accommodate multiple languages, making government services more accessible to a diverse population.

### Public Safety: AI Enhancing Security

#### Case Study: Predictive Policing in Los Angeles

The Los Angeles Police Department (LAPD) harnesses the power of AI for predictive policing. Machine learning algorithms analyze historical crime data to predict areas with a higher likelihood of criminal activity. This information allows the LAPD to allocate resources proactively, reduce crime rates, and enhance public safety. Importantly,

LAPD has

implemented measures to address ethical concerns and maintain community trust.

Responsible AI Practices:

**Fairness and Bias Mitigation:** LAPD continuously evaluates its predictive policing algorithms to ensure fairness and reduce potential bias.

**Community Engagement:** The LAPD actively engages with communities to build trust and transparency in its use of AI for public safety.

### **Administrative Efficiency: Streamlining Government Operations**

#### **Case Study: AI Automation in the UK's National Health Service (NHS)**

The UK's NHS has adopted AI-driven automation to streamline administrative processes. This includes the automation of appointment scheduling, data entry, and routine administrative tasks. By reducing bureaucratic red tape, the NHS has improved efficiency and freed up healthcare staff to focus on patient care.

Responsible AI Practices:

**Data Security:** The NHS always puts data security and patient confidentiality first when implementing AI automation.

**Efficiency and Service Quality:** AI is used to enhance service quality and ensure that administrative processes do not compromise patient care.

These case studies exemplify the responsible use of AI in the public sector, with governments adopting ethical principles, transparency, and accountability measures to ensure that AI technologies benefit citizens while upholding societal values. By learning from these real-world examples, governments can better navigate the complexities of responsible AI adoption and inspire trust among citizens and stakeholders.

In the subsequent sections, we will explore the ethical dilemmas and trade-offs that governments face in their pursuit of responsible AI adoption, as well as the critical role of government leadership in driving responsible AI initiatives.

## **ETHICAL DILEMMA & TRADE-OFF**

The conscientious integration of artificial intelligence (AI) within the public sector frequently necessitates the resolution of intricate ethical quandaries and the consideration of tough compromises. Governments find themselves in the position of carefully harmonizing diverse priorities to guarantee that AI technologies serve the greater good while preserving ethical norms and principles. Below, we explore some of the ethical quandaries and compromises frequently encountered in this endeavor.

### **Privacy vs. Security**

Dilemma:

Balancing the need to protect citizens' privacy with the imperative of ensuring national security and public safety is a significant challenge. Government agencies may need access to sensitive data for security purposes, but this can raise concerns about mass surveillance and infringement on individual privacy.

Trade-off:

Governments must establish clear guidelines and oversight mechanisms to determine when and how AI technologies can access and use sensitive data. Ensuring that data collection practices are proportionate, transparent, and subject to legal safeguards is crucial.

### **Fairness vs. Efficiency**

Dilemma:

Efficiency gains from AI may conflict with fairness objectives. For example, optimizing resource allocation based solely on historical data may perpetuate existing biases and inequalities in public services.

Trade-off:

Public sector should invest in fairness-aware AI algorithms that explicitly consider and address bias. While this may require additional time and resources, it ensures that AI-driven decision-making is fair and equitable.

### **Accountability vs. Autonomy**



**Dilemma:**

AI systems can operate autonomously, making decisions without direct human intervention. Determining accountability when AI systems make mistakes or cause harm can be challenging.

**Trade-off:**

Governments should establish legal frameworks that define clear lines of accountability for AI-driven decisions. These frameworks should specify the roles and responsibilities of individuals, agencies, and AI developers. Striking a balance between AI autonomy and human oversight is essential.

## Data Accessibility vs. Data Protection

**Dilemma:**

Access to a wide range of data is critical for training AI models effectively. However, data access can raise concerns about privacy, data protection, and the potential for misuse.

**Trade-off:**

Governments should prioritize data protection and compliance with privacy regulations. They can establish data-sharing agreements that strike a balance between enabling AI innovation and safeguarding citizen data rights. Implementing robust anonymization and encryption measures also helps protect sensitive information.

## Innovation vs. Risk Mitigation

**Dilemma:**

Innovation is a driving force behind AI adoption, but it can lead to the rapid deployment of technologies without thorough risk assessment. Governments must manage the tension between encouraging innovation and mitigating potential harms.

**Trade-off:**

Implementing regulatory sandboxes and controlled testing environments can enable innovation while minimizing risks. Governments should encourage responsible AI development through guidelines and incentives while maintaining the ability to intervene when necessary to protect the public interest.

## Cost vs. Impact

**Dilemma:**

Responsible AI practices may involve significant upfront costs, including investment in fairness assessments, audits, and ethical oversight. Allocating resources for these purposes can be challenging in budget-constrained environments.

**Trade-off:**

Governments should recognize that responsible AI adoption is an investment in long-term societal benefits. While there may be initial costs, the reduction of biases, the promotion of fairness, and improved decision-making can lead to significant cost savings and positive impacts on public services over time.

Navigating these ethical dilemmas and trade-offs requires careful consideration, ethical leadership, and ongoing collaboration between government agencies, AI developers, civil society, and citizens. By prioritizing responsible AI principles and adopting a holistic approach to AI governance, governments can minimize risks, maximize benefits, and ensure that AI technologies serve the public interest while upholding ethical standards and values.

## GOVERNMENT LEADERSHIP

Public sector leadership plays a pivotal role in shaping the responsible adoption of artificial intelligence (AI) in the public sector. Effective leadership sets the tone for ethical AI practices, fosters innovation, and ensures that AI technologies align with public values and interests. Here are key aspects of government leadership in driving responsible AI adoption:

### Policy and Regulation

**Setting the Ethical Framework:**

Government leaders must establish clear policies and regulations that define the ethical principles and standards for AI adoption. This includes guidelines for fairness, transparency, accountability, and data privacy. These policies provide a foundation for responsible AI development and deployment.

**Regulatory Oversight:**

Government agencies should exercise regulatory oversight to ensure that AI systems comply with established ethical standards. This oversight includes assessing AI algorithms, data usage, and the impact on citizens and society.

### Investment and Resources

#### Funding for Responsible AI:

Government leadership involves allocating resources and funding to support responsible AI initiatives. This includes investing in AI research, workforce training, and the development of AI ethics boards or oversight bodies.

#### Public-Private Partnerships:

Collaborating with the private sector and academia is crucial. Government leaders can facilitate partnerships that encourage responsible AI development, leveraging the expertise of industry leaders and researchers.

### Education and Awareness

#### Workforce Training:

Leadership entails recognizing the importance of workforce training and education on AI ethics and responsible practices. Governments should promote AI literacy among employees, ensuring that those working with AI systems understand the ethical implications.

#### Public Awareness:

Government leaders should engage in transparent communication with the public. This includes educating citizens about AI initiatives, their benefits, and the safeguards in place to protect their rights and privacy.

### Ethical AI Procurement

#### Ethical Guidelines for Procurement:

Government procurement processes should include ethical guidelines that require AI vendors to adhere to responsible AI principles. This ensures that AI technologies acquired by government agencies are aligned with ethical standards.

#### Vendor Accountability:

Leadership involves holding AI vendors accountable for the ethical use of their technologies. Government contracts should include provisions for auditing and monitoring AI systems' compliance with ethical guidelines.

### International Collaboration

#### Global Leadership:

Government leaders can take an active role in international forums and collaborations related to AI governance. This includes contributing to the development of global standards and norms for responsible AI.

#### Sharing Best Practices:

Leaders should promote the sharing of best practices and lessons learned with other governments. By participating in international networks and initiatives, governments can collectively address AI challenges.

### Ethical AI Implementation

#### Leading by Example:

Government agencies should lead by example in adopting responsible AI practices. This involves conducting ethical impact assessments, promoting fairness, and ensuring transparency in their AI initiatives.

#### Accountability and Reporting:

Leadership requires establishing mechanisms for reporting and addressing ethical concerns related to AI systems. Governments should create channels for citizens and employees to voice their ethical concerns.

### Adapting to Ethical Challenges

#### Ethical Adaptability:

Government leaders must be adaptive in the face of evolving ethical challenges posed by AI technologies. Flexibility in policy and regulation allows governments to respond to emerging ethical dilemmas.

#### Continuous Evaluation:

Leadership involves the continuous evaluation of AI systems and their impact on society. Regular audits, reviews, and updates to policies ensure that ethical standards are upheld.

Government leadership is instrumental in driving responsible AI adoption in the public sector. By setting ethical frameworks, allocating resources, fostering collaboration, and promoting transparency, governments can harness the potential of AI while safeguarding ethical principles and values.

## RESPONSIBLE AI ADOPTION CHALLENGES

The path to responsible AI adoption is not without its challenges. Examine the practical obstacles governments face:

### Budget Availability: Balancing Resources

Discuss the financial considerations involved in implementing AI solutions in the public sector.

### Data Availability and Quality: The Foundation of AI

Explore the challenges related to data access, quality, and interoperability, which are essential for successful AI deployment.

**Change Management: Preparing the Workforce**  
Address the importance of change management strategies to ensure a smooth transition to AI-driven processes and workflows.

## FUTURE OF RESPONSIBLE AI

The journey towards responsible AI adoption in the public sector is an ongoing evolution that holds immense potential for transforming government operations, enhancing public services, and fostering ethical governance. Looking ahead, several key trends and developments are poised to shape the future of responsible AI in the public sector:

### Advanced AI Governance Frameworks

The development of comprehensive AI governance frameworks will become increasingly critical. Governments will refine and expand existing ethical guidelines and regulations to address emerging AI challenges. These frameworks will encompass not only the principles of fairness, transparency, and accountability but also evolving considerations like AI-generated content, deep fakes, and AI in autonomous systems.

There are several responsible AI tools and frameworks available that can help public sector organizations to develop and deploy responsible AI systems. These tools and frameworks can help public sector organizations to assess the fairness of their data and algorithms, to monitor their AI systems for bias, and to explain the decisions that their AI systems make.

Here are a few examples of responsible AI tools and frameworks:

**AIF360:** AIF360 is an open-source toolkit for bias mitigation and model interpretability in machine learning models. It provides a suite of tools to help developers to identify and mitigate bias in their data and algorithms, and to explain the decisions that their models make.

- **Fairlearn:** Fairlearn is a Python library for assessing and mitigating bias in machine learning models. It provides several tools to help developers to assess the fairness of their models, and to identify and mitigate bias in their data and algorithms.
- **IBM Watson OpenScale:** IBM Watson OpenScale is a cloud-based platform that helps organizations to manage and monitor their AI systems. It provides several tools to help organizations to assess the fairness and performance of their AI systems, and to monitor their systems for bias.
- **Microsoft Responsible AI Toolkit:** The Microsoft Responsible AI Toolkit is a set of tools and resources to help developers to build and deploy responsible AI systems. It includes tools to help developers to assess the fairness of their data and algorithms, to monitor their AI systems for bias, and to explain the decisions that their systems make.

## ETHICAL AI AUDITING AND CERTIFICATION

To ensure that AI systems adhere to responsible practices, there will be a growing demand for ethical AI auditing and certification processes. Independent organizations may emerge to assess and certify AI solutions for their adherence to ethical standards. Governments will collaborate with these entities to establish recognized certification programs.

### AI for Public Engagement and Decision-Making

AI will play a more prominent role in public engagement and decision-making processes. Chatbots and virtual assistants will become ubiquitous for citizen interactions. AI-driven tools will help analyze public sentiment and input to inform policy making, enhancing government responsiveness and citizen involvement.

### AI in Public Safety and Disaster Management

AI technologies will bolster public safety efforts and disaster management. Predictive AI models will provide early warnings for natural disasters, optimizing resource allocation and response times. AI-driven analytics will assist law enforcement agencies in proactive crime prevention and emergency response.

## Responsible AI Education and Workforce Development

Governments will prioritize AI education and workforce development programs to equip public employees with the skills needed for responsible AI implementation. Training modules will emphasize AI ethics, transparency, and responsible AI practices.

## Public-Private Collaboration

Collaboration between the public and private sectors will deepen. Governments will partner with industry leaders to co-create ethical AI solutions and leverage private sector expertise. Public-private alliances will drive innovation while ensuring that AI serves public interests.

## International Cooperation on AI Ethics

International collaboration on AI ethics will gain momentum. Governments will work together to harmonize ethical standards and regulations, enabling responsible AI adoption on a global scale. International agreements and conventions may be established to address cross-border AI challenges.

## AI-Powered Healthcare and Education

AI will revolutionize healthcare and education in the public sector. Predictive analytics will aid in disease prevention and healthcare resource allocation. AI-driven personalized learning platforms will enhance education accessibility and quality.

## Ethical AI in Autonomous Systems

As autonomous systems become more prevalent in the public sector, governments will emphasize ethical considerations. Responsible AI will underpin autonomous vehicles, drones, and smart infrastructure, ensuring safety, fairness, and transparency.

## Continuous Ethical Oversight

Government agencies will establish ongoing ethical oversight mechanisms for AI systems. This includes regular audits, impact assessments, and the adaptation of ethical guidelines to evolving technologies and societal needs.

The future of responsible AI in the public sector holds the promise of a more transparent, equitable, and citizen-centric government. By embracing ethical AI

practices and staying at the forefront of AI governance, governments can harness the transformative potential of AI while upholding the principles of fairness, transparency, and accountability. As AI technologies continue to advance, responsible AI adoption will remain central to creating a brighter and more inclusive future for citizens and societies worldwide.

## CONCLUSION

In closing, summarize the key takeaways and emphasize the critical role of responsible AI in shaping the future of the public sector. Reiterate that responsible AI is not merely an ethical choice but a strategic imperative for governments worldwide. By adhering to ethical principles, fostering transparency, and prioritizing fairness, governments can harness AI's transformative potential while maintaining public trust and ensuring a brighter future for all.

## REFERENCES

1. Brynjolfsson, E., & McAfee, A. "Artificial Intelligence and the End of Work". MIT Sloan Management Review 2014. <https://sloanreview.mit.edu/article/artificial-intelligence-and-the-end-of-work/>
2. Brundage, M., Avin, S., Clark, J., et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation". <https://arxiv.org/abs/1802.07228>
3. AI Now Institute, New York University. "AI Now Report 2019". [https://ainowinstitute.org/AI\\_Now\\_2019\\_Report.pdf](https://ainowinstitute.org/AI_Now_2019_Report.pdf)
4. Nick Bostrom & Eliezer Yudkowsky. "The Ethics of Artificial Intelligence". <https://www.nickbostrom.com/ethics/ai.html>
5. Josh Tenenbaum, et al. "Building Machines That Learn and Think Like People". Behavioral and Brain Sciences, 2017. [https://www.mitpressjournals.org/doi/full/10.1162/NECO\\_a\\_00990](https://www.mitpressjournals.org/doi/full/10.1162/NECO_a_00990)
6. Selbst, A. D., & Barocas, S., "Fairness and Abstraction in Sociotechnical Systems". 2018. <https://dl.acm.org/doi/10.1145/3176349.3176356>
7. Daniel Crispin. "The AI Spring: How Artificial Intelligence Might End Climate Change". 2020. <https://www.greenpeace.org/international/publication/30353/the-ai-spring/>
8. Brundage, M., et al. "Responsible Artificial Intelligence". <https://arxiv.org/abs/2006.03586>



9. Jobin, A., Ienca, M., & Vayena, E. "AI Governance: A Review of the Concept and Its Implementation". <https://link.springer.com/article/10.1007/s00146-019-00850-5>
10. Diakopoulos, N. "Algorithmic Accountability: A Primer". [https://datasociety.net/pubs/ia/DataAndSociety\\_Algorithmic\\_Accountability\\_Primer\\_2016.pdf](https://datasociety.net/pubs/ia/DataAndSociety_Algorithmic_Accountability_Primer_2016.pdf)
11. Ayush Goyal, Anupam Datta, and Pedro Domingos. "A Review of AI Fairness Tools and Frameworks". Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML), 2022, pp. 1-20.
12. Hoda Heidari, Christoph Molnar, and Kristian Lum. "A Framework for Evaluating AI Fairness Tools and Frameworks". Proceedings of the 2022

Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML), 2022, pp. 1-12.

**Lakshmanan Sethu Sankaranarayanan** is a Technical Account Manager for Google Cloud in North America where he works with enterprise customers to help them succeed on Google Cloud. He is currently focusing on Generative AI/ML & Data Solutions. With more than a decade of experience in the technology industry, Lakshmanan has a passion to transform enterprise organizations with Google Cloud. Prior to Google, He had helped fortune 500 companies to migrate to cloud across North America & Asia & UK regions

### Upcoming Event

#### **The Hotstate Machine – A runtime loadable microcoded algorithmic state machine**

Virtual Event via Zoom and YouTube live.

Date: Tuesday, December 5th, 2023, 6:00 PM Free Registration:

Registration: [2312-hotstate.eventbrite.com](https://2312-hotstate.eventbrite.com)



#### **The Hotstate machine – A runtime loadable microcoded algorithmic state machine**

... *Hotstate machine, SystemVerilog, Finite State Machine* ...

**Steve Casselman**, CEO HotWright Inc.

Tuesday, December 5<sup>th</sup>, 2023, 6:00 pm PT (virtual)

*Via Zoom and YouTube Live*

Register (Free):

<https://r6.ieee.org/scv-cs/the-hotstate-machine-a-runtime-loadable-microcoded-algorithmic-state-machine/>

Vishnu S. Pendyala, Chair  
John Delaney, Vice Chair  
Sujata Tibrewala, Secretary  
S.R. Venkatramanan, Treasurer



IEEE  
COMPUTER  
SOCIETY

*Santa Clara Valley Chapter*

# Edge AI in AR/VR: exploring opportunities and confronting Challenges

Dwith Chenna, MagicLeap Inc. USA

## Abstract

***Dynamic fusion of Edge AI with Augmented Reality/Virtual Reality (AR/VR), unraveling a landscape rife with exciting opportunities and intricate challenges. The synergy between Edge AI and AR/VR technologies unveils a horizon where real-time performance, reduced latency, and heightened user immersion become achievable realities. The paper examines the manifold opportunities that arise when Edge AI is seamlessly integrated into AR/VR systems. It highlights the potential for local data processing to unleash unprecedented levels of interactivity and responsiveness. Furthermore, the empowerment of AR/VR devices to make split-second decisions on the edge promises to reshape user experiences across various domains. In this paper we explore different Edge AI algorithms to run on resource constrained hardware available on AR/VR systems. By navigating these challenges prudently, stakeholders can unlock the transformative potential of this synergy, ensuring that the fusion of Edge AI and AR/VR augments human experiences responsibly and innovatively.***

**Keywords:** Edge AI, Augmented Reality (AR), Virtual Reality (VR)

Augmented Reality (AR) and Virtual Reality (VR) are transformative technologies, each offering unique immersive experiences and applications. AR enriches the real world by overlaying digital information, such as 3D models or contextual data, onto our physical surroundings. This technology is accessible through devices like smartphones, tablets, and AR glasses, such as Microsoft's HoloLens or Apple's ARKit. AR finds applications in diverse fields, including gaming e.g. Pokemon Go, navigation providing real-time directions and location-based information [1], retail enabling virtual try-ons and interactive shopping experiences[2], education [3](enhancing learning with interactive 3D models), and maintenance and repair supporting technicians with digital instructions[4].

On the other hand, Virtual Reality (VR) transports users to entirely computer-generated environments, often achieved through specialized headsets like the Oculus Rift or HTC Vive, accompanied by motion controllers and high-performance computing systems. VR is an advanced technology in gaming, offering immersive, 360-degree experiences where users feel present within the digital world. Beyond gaming, VR has profound applications in training simulations [5] (e.g., flight or medical training), psychological therapy[6](utilizing exposure therapy to treat various conditions), design and architecture[7] (enabling architects to explore 3D models in a virtual space),

and entertainment[8] (with applications in virtual cinemas and live events). Moreover, Mixed Reality (MR) bridges the gap between AR and VR, allowing digital and real-world elements to interact seamlessly. Notable examples include Microsoft's HoloLens and MagicLeap, which blend holographic content with the user's environment. MR opens up possibilities for remote collaboration, complex data visualization, and hands-free interactions in industries like healthcare, engineering, and design. As these technologies continue to evolve and become more accessible, they hold immense potential for shaping the future of various sectors, promising innovative solutions and enriched human experiences.

AR/VR devices incorporate a diverse array of sensors to create immersive digital experiences as shown in Fig 1. RGB cameras capture the visual world, enabling users to see the real environment. Infrared (IR) cameras, on the other hand, detect infrared light, allowing for night vision and heat mapping applications. Depth sensors, often using technologies like LiDAR or Time-of-Flight, measure the distance to objects, facilitating accurate spatial mapping and gesture recognition. Inertial Measurement Units (IMUs) combine accelerometers and gyroscopes to track head and body movements[9], ensuring precise positional data. Audio plays a pivotal role, with microphones capturing real-world sound and speakers delivering spatialized audio[10], enhancing immersion by providing a 360-degree soundscape[11].

Collectively, these sensors converge to bridge the gap between the virtual and physical worlds, making AR/VR experiences truly transformative.

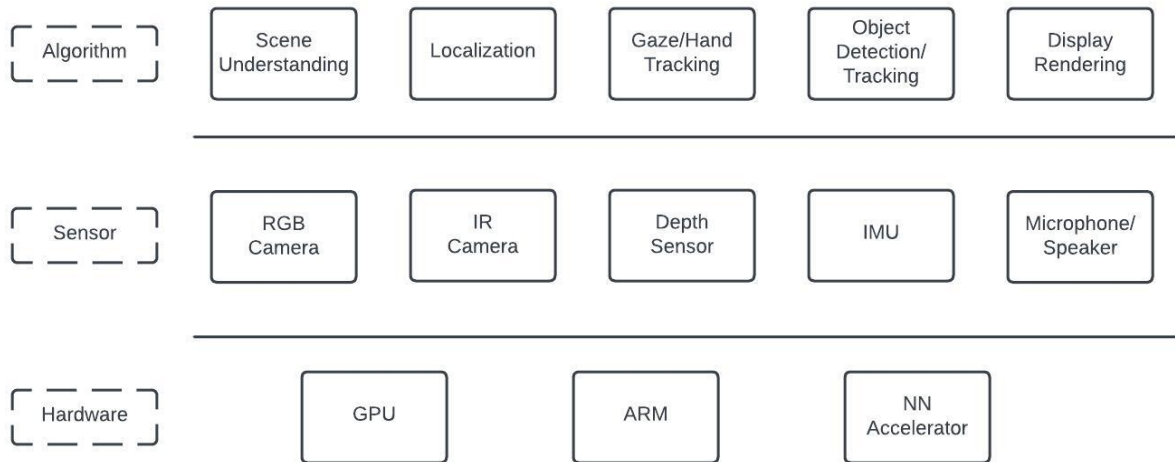


Fig 1. Overview of the sensors, algorithms, and Hardware in AR/VR systems

**RGB cameras:** AR/VR headsets use RGB cameras to serve multifaceted roles, from mapping real-world surroundings for obstacle detection to enabling augmented reality experiences by blending virtual content with the physical environment. They also support hand and gesture tracking, pass-through video for quick real-world glimpses, mixed reality experiences, and stereo depth sensing for precise object interactions. These cameras greatly enhance the overall immersion, safety, and interactivity of AR and VR applications, making them indispensable components in modern headsets.

**IR cameras:** In AR/VR systems, infrared (IR) cameras have diverse functions, including precise tracking of head and hand movements, enabling gesture recognition, tracking eye movements for performance optimization, and capturing facial expressions for realistic avatars. They are also essential in low-light conditions and for night vision, while their combination with depth-sensing technology allows for 3D environment mapping. IR cameras play a pivotal role in enhancing immersion and interactivity, making them indispensable components in modern AR and VR setups.

**Depth sensor:** These sensors provide critical spatial information about the environment critical for AR/VR. These sensors typically use technologies like Time-of-Flight (ToF) to measure distances accurately. In AR, depth sensors enable precise object placement and occlusion, ensuring virtual objects interact convincingly with the real world. They also facilitate hand and gesture tracking, enhancing user

interaction. In VR, depth sensors contribute to more immersive experiences by allowing users to interact with virtual objects and navigate environments in a natural and responsive manner. Additionally, depth sensors support features like foveated rendering, where graphics quality is optimized based on where the user is looking, improving performance and realism.

**Inertial Measurement Unit (IMU):** Incorporates three essential sensors into its design. Firstly, there's the accelerometer, which is responsible for determining linear acceleration along the x, y, and z axes while also factoring in the force of gravity. Secondly, the gyroscope is employed to precisely measure rotations and angular movements. Lastly, the magnetometer comes into play, aiding the IMU in estimating absolute orientation. Collectively, these sensors enable the IMU to capture data regarding motion, orientation, and gravitational effects, making it a crucial component in an array of applications, from robotics and aerospace to virtual reality systems[12].

**Microphone/Speaker:** Enabling voice commands, spatial audio for realism, social interaction, and environmental immersion. They also support narration and storytelling, provide feedback and guidance, allow for voice recognition, and capture user responses, contributing to user engagement and adaptability within virtual environments. Microphones enhance both the interactivity and immersion of AR and VR experiences by adding a crucial audio layer to the virtual world.

In AR/VR systems, the data from these sensors need to be processed through various deep learning algorithms to enhance the overall experience and functionality. Some key applications include:

**Scene Understanding:** Computer vision algorithms are central to scene understanding. Techniques such as object detection[], semantic segmentation[], and instance segmentation[] are used to identify and classify objects within the camera's view. This helps in recognizing the scene's content. AR/VR glasses continuously update their understanding of the scene as the user moves, making use of real-time mapping and localization techniques to maintain accurate spatial awareness. It is essential to be able to process this data at the edge, allowing EdgeAI algorithms to perform tasks like semantic segmentation, which divides the camera's view into meaningful segments, and depth estimation, which provides information about the distance of objects from the viewer. These capabilities improve the realism and context-awareness of AR/VR experiences.

**Localization:** Simultaneous Localization and Mapping (SLAM) algorithms enable AR/VR glasses to understand their precise position and orientation in the physical world while simultaneously mapping the surrounding environment. This technology is crucial for creating a seamless mixed-reality experience where virtual objects are anchored to real-world locations. Deep learning can be used for creating 3D maps of the environment using data from RGB cameras and depth sensors (e.g., ToF cameras). These maps are crucial for precise object placement and occlusion in mixed reality scenarios. In addition to traditional mapping techniques like Simultaneous Localization and Mapping (SLAM)[13], deep learning algorithms such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), can enhance the accuracy and robustness of spatial mapping[14] as shown in Fig 2. CNNs can be employed for feature extraction and recognition, aiding in the identification of features, depth, objects and surfaces within the environment. RNNs, on the other hand, can assist in tracking and predicting the movement of objects and users, further improving the quality of spatial mapping data. These deep learning-powered enhancements contribute to more seamless and immersive mixed reality experiences.

**Gaze/Hand Tracking:** These algorithms recognize hand and body gestures, enabling users to interact intuitively with AR/VR environments. Hand tracking and gesture recognition are fundamental for controlling applications, navigating menus, and enhancing the sense of immersion. Eye tracking technology significantly enhances the capabilities and user experience of AR/VR glasses by enabling foveated rendering, reducing motion sickness, supporting natural interaction, providing gaze analytics, enhancing social interactions, and improving accessibility. Deep learning algorithms play a vital role in eye and hand

tracking. Convolutional Neural Networks (CNNs) are used to accurately detect and monitor users' eyes for gaze estimation[15-16], which is used for features like foveated rendering, optimizing graphics where users are looking, and hand tracking enables intuitive and realistic interactions within these virtual worlds. It is also used for tracking hand movements and recognizing gestures[17]. This technology enhances immersion by allowing users to interact naturally with virtual environments, making it possible to control and navigate through AR/VR experiences with precision.

**Object Recognition and Tracking:** EdgeAI algorithms are employed to detect and track objects or markers within the camera's field of view. This capability is essential for recognizing physical objects and for seamlessly integrating virtual elements into the real-world environment. It allows for dynamic interactions between the user and the augmented or virtual content. Deep learning models like Convolutional Neural Networks (CNNs) are used to recognize and track objects in the real-world environment captured by AR/VR cameras. This is essential for placing virtual objects accurately within the scene. Convolutional Neural Networks (CNNs) are instrumental in recognizing and tracking real-world objects. Various types of CNN-based algorithms serve specific purposes: object detection algorithms like Faster R-CNN, YOLO[18], and SSD[19] identify and locate objects in real-time, ensuring precise placement of virtual objects. Semantic segmentation with CNNs such as Mask R-CNN [20] and U-Net[21] distinguishes object boundaries, facilitating more convincing occlusion effects. Deep learning-based tracking algorithms have gained prominence in computer vision, offering solutions for robust and efficient object tracking. Long Short-Term Memory (LSTM) networks predict object trajectories[22], and DeepSORT[23] enhances multiple-object tracking. Integration with Kalman filters and CNNs further extends their versatility, making these algorithms valuable for a wide range of applications, from surveillance to augmented reality. This comprehensive use of CNNs enhances the immersion and realism of AR/VR experiences by seamlessly integrating virtual and real-world elements.

**Display Render:** Display rendering is vital in AR and VR applications, as it creates immersive environments, enhances realism, and ensures user comfort and engagement. AI algorithms work to replicate life-like environments by dynamically fine-tuning aspects such as lighting, physics, and textures. This results in immersive experiences that closely mirror the physical world[32]. Realistic graphics and well-rendered user interfaces improve interaction, navigation, and educational value, making these technologies valuable in fields like education, healthcare, marketing, and scientific research. In this context, deep learning algorithms further enhance



AR/VR by enabling advanced image recognition, object tracking, and real-time content optimization, ensuring that virtual elements seamlessly integrate with the real world and delivering a richer and more responsive user experience. Furthermore, in order to provide an immersive visual experience, modern displays require head mounting, high image resolution, low latency, as well as high refresh rate. This poses a challenging computational problem, foveated rendering [30] is a technique that uses gaze estimation to estimate the real time gaze and perform gaze dependent rendering and compression. Deep learning algorithms, specifically Generative Adversarial Networks (GANs), are instrumental in generating highly realistic and immersive environments[24,25], characters, or objects within the AR/VR world. These technologies significantly reduce the manual effort required for content creation, enabling more efficient and dynamic AR/VR experiences. GANs leverage adversarial training to create authentic-looking content to enhance the visual richness of AR/VR environments. DeepFovea[31], uses deep learning to foveated rendering.

**Audio Processing:** Audio processing is paramount in AR and VR, enhancing immersion and realism. Deep learning, a vital component, brings about significant improvements in audio processing. It enables spatial audio, lifelike sound cues, responsive feedback, and personalized experiences. Realistic soundscapes not only engage users more deeply but also add a critical layer of immersion to these technologies. Additionally, deep learning assists in speech recognition, adapting audio to environmental conditions, and creating interactive audio elements, culminating in more intuitive and compelling AR/VR interactions. Recurrent Neural Networks and Convolutional Neural Networks can be used for audio processing in AR/VR to create spatial audio, simulate real-world sound environments, and enhance the sense of immersion. Deep learning algorithms are pivotal in audio processing for AR/VR applications[26]. They enable accurate speech recognition, noise reduction for clearer audio, realistic voice synthesis, spatial audio simulation, emotion detection, sound source localization, and auditory scene analysis. These advancements enhance the overall auditory experience, making AR/VR environments more immersive and interactive, with applications ranging from natural language interaction to 3D sound modeling and emotion-aware virtual experiences.

In essence these diverse source sensors through visual, spatial and auditory, allow the usage of multi-model interactions. Large language models, which can analyze multi-modal these inputs to provide personalized experiences. These models excel in analyzing multimodal inputs, enabling the creation of personalized experiences. Multi-modal systems possess an enhanced capability to comprehend user context, thereby enabling more context aware-

responses and actions within AR/VR environments. Consequently, this contributes to the creation of immersive and captivating experiences. For instance, by discerning a user's preferences from their interactions, the system can tailor its content and responses to better align with individual needs and inclinations. These models also enhance accessibility in AR/VR for individuals facing disabilities and language barriers. This is achieved through the incorporation of voice commands, text input, and gesture controls, ensuring a broader range of users can engage with these technologies, thus fostering greater inclusivity.

These diverse deep learning algorithms enhance the functionality and immersion of AR/VR experiences by enabling object recognition, gesture control, voice interaction, and personalized content delivery, among other capabilities. They are crucial for creating interactive and engaging virtual environments. While Edge AI in AR/VR holds immense promise, it also brings forth a set of challenges that must be addressed for its successful integration. To execute these EdgeAI algorithms in real-time while conserving battery life, AR/VR glasses often incorporate specialized hardware accelerators such as GPUs and Neural Network Accelerators. Additionally, privacy-focused algorithms are essential to ensure that sensitive data, like faces, is processed securely and that user privacy is maintained. EdgeAI algorithms contribute significantly to creating immersive, responsive, and context-aware experiences in AR/VR glasses, reducing the need for cloud processing and minimizing latency. In the next section, we will explore the opportunities and challenges of Edge AI in the realm of AR/VR.

## OPPORTUNITIES

EdgeAI, the convergence of artificial intelligence and edge computing, presents a realm of compelling opportunities within the domain of AR/VR. These immersive technologies demand real-time processing for seamless experiences, placing significant emphasis on the computational power of edge devices. Moreover, as privacy and security concerns loom large, EdgeAI provides a promising avenue for on-device data processing, ensuring that sensitive information remains secure within the user's immediate environment. By offloading AI inference from remote servers to the edge, bandwidth optimization becomes a reality, mitigating latency issues and reducing the reliance on high-speed internet connections. As user experience reigns supreme in AR/VR, EdgeAI can further enhance

Cloud AI	Mobile AI	AR/VR AI	Tiny ML/AI
<ul style="list-style-type: none"> <li>• DNN</li> <li>• Large Models</li> <li>• X Million parameters</li> <li>• TFLOPs</li> <li>• Focus on Accuracy</li> <li>• H/W: GPU/TPU/FPGA</li> <li>• AlexNet, Inception, ResNet, VGGnet</li> <li>• Data: storage, sharing (1%)</li> </ul>	<ul style="list-style-type: none"> <li>• Optimized Algorithms, lite CNN</li> <li>• Constrained resources: memory few GB RAM, application size limitation</li> <li>• GFLOPs</li> <li>• Focus on accuracy-efficiency trade-off</li> <li>• H/W: SoC, NPU</li> <li>• MobileNet, ShuffleNet, SqueezeNet</li> <li>• Data: pics, audio, clicks</li> </ul>	<ul style="list-style-type: none"> <li>• Optimized Algorithms, lite CNN</li> <li>• Constrained resources: memory few GB RAM, application size limitation</li> <li>• H/W: SoC, DSP with hardware accelerator</li> <li>• Sensors: CMOS sensor, IR camera, audio, IMU, accelerometer</li> <li>• Data: sensing physical world, video render(95%)</li> </ul>	<ul style="list-style-type: none"> <li>• CNN Micro</li> <li>• Severely constrained hardware</li> <li>• ~100KB RAM</li> <li>• MCU with hardware accelerator</li> <li>• Sensors: CMOS sensor, IR camera, audio, IMU, accelerometer, temp, chemical</li> <li>• Data: sensing physical world (95%)</li> </ul>

Table 1. Comparison of different compute elements [33]

these experiences by enabling smoother interactions and more responsive content rendering.

Additionally, the ability to function offline, leveraging local AI models, adds a layer of versatility and accessibility to AR/VR applications. EdgeAI is poised to redefine the way we perceive, interact with, and secure our AR/VR experiences, shaping a future where these technologies seamlessly blend with our physical world. In this section, as discussed above, we will delve into specific prerequisites within the AR/VR domain that underscore the critical need for EdgeAI.

**Real-Time Processing:** One of the primary advantages of Edge AI in AR/VR is its ability to process data locally on edge devices rather than relying on distant cloud servers. This reduces the latency significantly, enabling real-time data analysis and response. As a result, users can experience smoother interactions and seamless transitions in AR/VR environments. Commonly recognized as

"motion-to-photon latency" [27] within the realm of AR/VR

technologies, this term denotes the time required for a user's physical movement to be fully displayed on a screen. For instance, if there's a 100ms delay between your movement and its representation on your AR/VR headset's screen after you shift your gaze, that 100ms constitutes the motion-to-photon latency. It's imperative to maintain a low motion-to-photon latency (< 20ms) in order to effectively immerse the user's mind into the virtual environment, a sensation referred to as "presence." Presence captures the feeling of being within the simulated world. To attain this state of presence, one crucial prerequisite is minimal latency. Specifically, a motion-to-photon latency of under 20ms is necessary. Table 2, shows the latency requirements for different systems with edge and hybrid compute. Edge compute uses on device compute where hybrid compute uses combine the edge and cloud compute.

Component	Edge Compute (ms)	Hybrid Compute (ms)
Sensors	1	1
Data Transport	2 (USB and HDMI)	40 (network to cloud)
Computing	3-5	100+
Display	10 (refresh)	15 (decode and refresh)
Total	18	150+

Table 2. Latency of AR/VR components with different approaches [28]

Conversely, elevated motion-to-photon latency leads to an unsatisfactory virtual reality encounter, often

resulting in motion sickness and a sense of nausea. Should there be a delay between the user's movement and the screen's response, it can lead to feelings of

disorientation and motion sickness. Ultimately, this detrimentally disrupts the entire experience for the user.

By harnessing the computational power at the edge devices, AI algorithms can rapidly analyze and respond to sensory inputs from the user's environment. This instantaneous processing ensures that virtual elements seamlessly integrate with the real world, eliminating perceptible delays in moving data to the cloud. Whether it's tracking movements, rendering immersive graphics, or providing context-aware information, EdgeAI's ability to process data in real-time enhances the overall fluidity and responsiveness of AR/VR experiences, setting a new standard for immersion and interactivity.

**Privacy and Security:** AR/VR devices are composed of several different information-gathering technologies, each presenting unique privacy risks and mitigation approaches. AR/VR devices and applications gather significant amounts of personal data, including information provided by users, information generated by users, and information inferred about users. Individual AR/VR devices come with built-in sensors collecting biometric data, brain patterns, users' speech and facial expressions and poses, and also the surrounding

environment to render a high-quality immersive experience. These technologies are essentially a collection of sensors and displays that work in concert to create an immersive experience for the user. Edge AI allows AR/VR applications to process sensitive user data on the device itself, enhancing privacy and security [29]. Since data doesn't need to be transmitted to remote servers, the risk of data breaches and unauthorized access is minimized, building user trust and confidence in using AR/VR technologies.

**Bandwidth Optimization:** The data generated in AR/VR experiences can be voluminous, putting a strain on network bandwidth when transmitted to the cloud for processing. Edge AI mitigates this issue by processing data locally, reducing the amount of data sent to the cloud, thereby optimizing bandwidth usage. Table 3 below shows the bandwidth requirements for different applications and shows the need to improve bandwidth requirements for immersive AR/VR applications. This clearly shows a need to process data locally, which helps to preserve the bandwidth requirement.

Bandwidth	Application
1Mbps	Image and Workflow downloads
2Mbps	Video Conference
2 to 20 Mbps	3D Model and Data visualization
5 to 25 Mbps	Two-way telepresence
10 to 50 Mbps	Current-gen 360 video (4K)
50 to 200 Mbps	Next-gen 360 videos (4K + 90 FPS)
200 to 500 Mbps	6 DoF Video or free viewpoints

Table 3. Network bandwidth requirements for AR/VR applications []

The advent of 5G technology and edge computing has brought significant advancements and is poised to revolutionize AR/VR experiences by addressing key challenges bandwidth requirements. 5G networks offer substantially higher bandwidth compared to previous generations of mobile networks. This increased bandwidth enables the seamless streaming of high-definition and even 4K/8K content, essential for delivering immersive VR experiences.

**User Experience:** Edge AI enables personalized and context-aware experiences in AR/VR applications. Instead of relying on centralized cloud servers for data analysis, edge AI brings this processing capability

directly to the user's device. This means that the AR/VR application can assess user behavior, preferences, and contextual information in real-time, without the need for continuous internet connectivity. The ability to analyze user behavior and preferences on the device empowers the application to tailor content and interactions to everyone, leading to a more immersive and engaging experience. This localized analysis empowers the application to tailor content and interactions to the individual user. For example, it can adapt the lighting, audio, or object placement based on the user's preferences and environment. If a user frequently interacts with certain virtual elements or exhibits particular behavioral patterns, the AR/VR system can anticipate these

actions and respond accordingly, creating a more seamless and intuitive experience. Edge AI in AR/VR applications represents a significant step toward creating immersive and engaging experiences that adapt and respond to each individual user, fostering a stronger sense of presence and interaction within virtual environments.

**Offline Functionality:** With Edge AI, AR/VR applications can continue functioning even in offline or low-connectivity scenarios. This is particularly valuable in remote locations or crowded events where internet access may be limited. It is crucial not only in military and first responder applications but also in various generic use cases of AR/VR. In healthcare, it ensures that medical data and training modules can be accessed for surgical planning, medical training, and remote consultations even in areas with unreliable internet connectivity, thus safeguarding patient care and healthcare education. Likewise, in education, offline capability allows students in underserved or remote regions to access immersive educational content, bridging the digital divide and expanding access to quality learning experiences. Furthermore, in the realm of entertainment, offline AR/VR experiences enable users to enjoy immersive gaming and interactive storytelling while on the move, without the need for a stable internet connection. Overall, offline functionality in AR/VR devices extends their reach and utility across diverse domains, guaranteeing access to crucial resources and immersive experiences irrespective of internet availability.

## CHALLENGES

Running deep learning algorithms on the edge within AR/VR systems presents a promising avenue for real-time, responsive, and privacy-enhanced experiences. However, it also comes with its own set of formidable challenges. These challenges span computational limitations of edge devices, the need for energy-efficient inference, optimization of model sizes, and the inherent trade-offs between performance and power consumption. Furthermore, ensuring seamless integration of AI into AR/VR systems while addressing issues of latency, accuracy, and resource constraints remains a complex endeavor. This discussion will explore these challenges in detail, shedding light on the evolving landscape of EdgeAI in the context of augmented and virtual reality.

**Power and Energy Efficiency:** Edge devices, such as smartphones and AR/VR headsets, have limited processing power and battery life. Implementing AI algorithms locally demands significant computational resources, which may strain the device's hardware and drain the battery quickly. Optimizing AI models for energy efficiency becomes crucial to overcome this challenge. Hardware optimization should incorporate

energy-efficient hardware components. This includes low-power CPU, GPU, DSP and specialized AI accelerators when applicable. Efficient hardware design can significantly reduce power consumption during AI inference. Optimizing AI algorithms for inference on edge devices is essential. Techniques like quantization, pruning, and model compression reduce the computational demands, lowering power requirements without sacrificing accuracy. In scenarios where Edge AI devices use sensors (e.g., cameras, accelerometers), sensor fusion techniques can reduce the need for continuous data acquisition, conserving power while maintaining context awareness. Implementing energy monitoring features within Edge AI devices can help users track and manage power consumption. Alerts and notifications can inform users when energy-hungry processes are active. Specialized event-based sensors that can trigger AI algorithms will allow efficient battery management algorithms and techniques can help extend the operational lifespan of AR/VR devices. By carefully selecting hardware components, optimizing algorithms, and implementing energy-efficient strategies, Edge AI solutions can provide reliable, real-time AI processing while conserving power and maximizing device longevity.

**Model Size:** Advanced AI models designed for AR/VR applications can be large and resource-intensive, making it difficult to deploy them directly on edge devices with limited storage capacity. Shrinking the model size without sacrificing performance becomes essential to facilitate seamless integration. Edge AI in AR/VR requires striking a balance between model complexity and accuracy. While complex models can deliver more precise results, they may be impractical for edge devices due to their limited resources. Choosing the right AI model for a specific task can impact energy efficiency. Smaller, more lightweight models may consume less compute and power during inference while still delivering satisfactory results for certain applications. Finding the right trade-off is essential to ensure optimal performance.



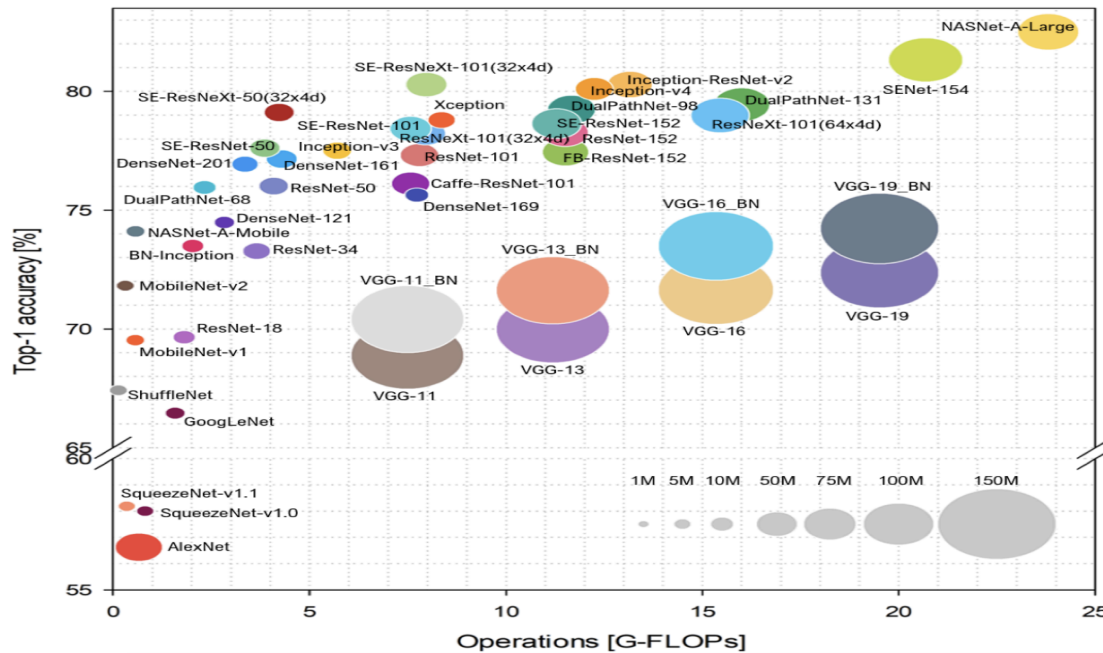


Figure 2. Comparison of Model Size vs Accuracy of Deep Learning models [34]

**Deployment:** The AR/VR ecosystem encompasses a broad spectrum of devices, featuring diverse hardware configurations and operating systems. Ensuring seamless compatibility and achieving peak performance across this multifaceted landscape presents a formidable hurdle for Edge AI developers. Similarly, the embedded systems domain exhibits remarkable diversity, ranging from microcontrollers with 8 to 32-bit processors to System-on-Chips (SoCs), low-power GPUs, FPGAs, and specialized Neural Network Accelerators. Each device category encompasses different manufacturers, each offering distinct development tools, programming environments, and interfaces. This amalgamation of hardware diversity and toolset variations can be particularly daunting. For instance, consider the deployment of AI models on embedded devices. Developers face a myriad of choices, often requiring trade-offs between model performance, accuracy, and ease of implementation. While there's an abundance of AI models available, not all of them prioritize the essentials for edge computing—compact model sizes, optimization for underlying hardware, efficiency-enhancing techniques, computational frugality, and the ability to train with limited datasets. In essence, just as Edge AI developers grapple with ensuring AR/VR compatibility and performance, those in the embedded space contend with a heterogeneous landscape of hardware devices and tools, all while navigating the delicate balance between model capabilities and resource constraints.

**Data Privacy:** Edge computing enhances privacy by processing data locally, reducing the need for constant cloud-based data transmission. However, this approach introduces security challenges. Edge AI developers must prioritize robust security measures to protect sensitive data on devices. To mitigate risks, developers should implement end-to-end encryption, secure data storage, and regular audits. Privacy by design and user education are crucial. Balancing privacy and security is key to realizing the benefits of edge computing while ensuring data protection, fostering user trust in edge AI applications.

## CONCLUSION

The fusion of Edge AI and AR/VR unveils a wealth of promising prospects, marked by minimal latency, enhanced privacy, and tailor-made interactions. Nevertheless, conquering the obstacles associated with processing capabilities, model sizes, and the diverse array of devices is indispensable for a harmonious integration. As technology makes strides and research advances, the trajectory of Edge AI within the AR/VR domain holds the potential to revolutionize our interactions with the world, endowing them with heightened engagement, seamlessness, and immersion. While challenges are evident, the synergistic union of Edge AI and AR/VR paints a picture of a forthcoming era where the divisions between the tangible and digital realms become indistinct, ushering in a redefined era of human interactions with augmented and virtual realities.

## REFERENCES

- [1] Nayyar A, Mahapatra B, Le DN, Suseendran G (2018) Virtual reality (VR) & augmented reality (AR) technologies for tourism and hospitality industry. *Int J Eng Technol* 7(2):156–160.
- [2] A. Javornik, Y. Rogers, A.M. Moutinho, R. Freeman, Revealing the shopper experience of using a "magic mirror" augmented reality make-up application, in: *Conference on Designing Interactive Systems*, vol. 2016, 2016, pp. 871–882. Association for Computing Machinery (ACM).
- [3] H. Ardiny and E. Khanmirza, "The role of ar and vr technologies in education developments: opportunities and challenges," in *6th RSI International Conference on Robotics and Mechatronics*. IEEE, 2018, pp. 482–487.
- [4] Eswaran, M.; Bahubalendruni, M.R. Challenges and opportunities on AR/VR technologies for manufacturing systems in the context of industry 4.0: A state of the art review. *J. Manuf. Syst.* 2022, 65, 260–278.
- [5] Satava RM, Jones SB. Medical applications of virtual reality. In: Stanney KM, editor. *Handbook of Virtual Environments: Design, Implementation, and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.; 2002. pp. 368–391.
- [6] Tsao, Y. C., Shu, C. C., & Lan, T. S. (2019). Development of a reminiscence therapy system for the elderly using the integration of virtual reality and augmented reality. *Sustainability*, 11(17), 1–10.
- [7] Davila Delgado, J.M.; Oyedele, L.; Demian, P.; Beach, T. A research agenda for augmented and virtual reality in architecture, engineering and construction. *Adv. Eng. Inform.* 2020, 45, 101122.
- [8] Jenny, S. Enhancing Tourism with Augmented and Virtual Reality; HAMK: Espoo, Finland, 2017.
- [9] I. Arun Faisal, T. Waluyo Purboyo, and A. Siswo Raharjo Ansori, "A Review of Accelerometer Sensor and Gyroscope Sensor in IMU Sensors on Motion Capture," *Journal of Engineering and Applied Sciences*, vol. 15, no. 3, pp. 826–829, Nov. 2019.
- [10] Auria, D.D.; Mauro, D.D.; Calandra, D.M.; Cutugno, F. A 3D audio augmented reality system for cultural heritage management and fruition. *J. Digit. Inf. Manag.* 2015, 13, 203–209.
- [11] Hurtado Soler, A., Botella Nicolás, A. M., & Martínez Gallego, S. (2022). Virtual and Augmented Reality Applied to the Perception of the Sound and Visual Garden. *Educ. Sci.*, 12, 377. <https://doi.org/10.3390/educsci12060377>
- [12] N. Ahmad, R. A. R. Ghazilla, N. M. Khairi, and V. Kasi, "Reviews on various inertial measurement unit (IMU) sensor applications," *International Journal of Signal Processing Systems*, vol. 1, no. 2, pp. 256–262, 2013.
- [13] Aulinas, J., Y. R. Petillot, J. Salvi, and X. Lladó. 2008. "The SLAM Problem: A Survey." *Conference on Artificial Intelligence Research and Development*, 11th International Conference of the Catalan Association for Artificial Intelligence, Spain, October 22–24.
- [14] Ruihao Li, Sen Wang, and Dongbing Gu. Deepslam: A robust monocular slam system with unsupervised deep learning. *IEEE Transactions on Industrial Electronics*, 68(4):3577–3587, 2021. doi: 10.1109/TIE. 2020.2982096.
- [15] P. L. Mazzeo, D. D'Amico, P. Spagnolo, and C. Distantè, "Deep learning based eye gaze estimation and prediction," in *Proc. 6th Int. Conf. Smart Sustain. Technol. (SpliTech)*, Sep. 2021, pp. 1–6, doi: 10.23919/SpliTech52315.2021.9566413.
- [16] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*, 2021.
- [17] Sanchez-Riera, J.; Srinivasan, K.; Hua, K.-L.; Cheng, W.-H.; Hossain, M.A.; Alhamid, M.F. Robust rgb-d hand tracking using deep learning priors. *IEEE Trans. Circuits Syst. Video Technol.* 2017, 28, 2289–2301.
- [18] Terven, J.; Cordova-Esparza, D. A Comprehensive Review of YOLO: From YOLOv1 and Beyond. *arXiv* 2023, arXiv:2304.00501
- [19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *ECCV*, 2016.
- [20] P. Bharati, A. Pramanik, Deep learning techniques—r-cnn to mask rcnn: A survey, in: *Computational Intelligence in Pattern Recognition*, Springer, 2020, pp. 657–668.
- [21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 234–241.
- [22] Gao, C.; Yan, J.; Zhou, S.; Chen, B.; Liu, H. Long short-term memory-based recurrent neural networks for nonlinear target tracking. *Signal Process.* 2019, 164, 67–73.
- [23] B. Veeramani, J. W. Raymond, and P. Chanda, "DeepSort: Deep Convolutional Networks for sorting haploid maize seeds," *BMC Bioinformatics*, vol. 19, no. 9, p. 289, 2018.
- [24] Lee, B., Jo, Y., Yoo, D. & Lee, J. Recent progresses of near-eye display for AR and VR. In Stella, E. (ed.) *Multimodal Sensing and Artificial Intelligence: Technologies and Applications II*, vol. 11785, 1178503, DOI: 10.1117/12.2596128. International Society for Optics and Photonics (SPIE, 2021).

[25] Li, Z., Chen, L., Liu, C., Zhang, F., Li, Z., Gao, Y., Ha, Y., Xu, C., Quan, S., & Xu, Y. (2021). Animated 3D human avatars from a single image with GAN-based texture inference. *Computers and Graphics*, 95, 81–91. <https://doi.org/10.1016/j.cag.2021.01.002>

[26] Maximo Cobos, Jens Ahrens, Konrad Kowalczyk, and Archontis Politis, “An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–21, 2022.

[27] <http://www.chioka.in/what-is-motion-to-photon-latency/>

[28] S. Mangiante et al., “VR is on the Edge: How to Deliver 360° Videos in Mobile Networks,” in *Proc. ACM SIGCOMM. Workshop on Virtual Reality and Augmented Reality Network (VR/AR Network)*, 2017.

[29] R. Di Pietro and S. Cresci, “Metaverse: Security and privacy issues,” in *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, Atlanta, GA, Dec. 2021, pp. 281–288.

[30] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D Graphics. *ACM Trans. Graph. (Proc. SIGGRAPH)* 31, 6, Article 164 (2012), 164:1–164:10 pages.

[31] Anton S. Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. 2019. DeepFovea: Neural Reconstruction for Foveated Rendering and Video Compression using Learned Statistics of Natural Videos. *ACM Trans. Graph.* 38, 4, Article 212 (July 2019), 13 pages. <https://doi.org/10.1145/3355089.3356557>

[32] Singh, M.; Fuenmayor, E.; Hinchy, E.P.; Qiao, Y.; Murray, N.; Devine, D. Digital Twin: Origin to Future. *Appl. Syst. Innov.* 2021, 4, 36.

[33] Stanislava Soro. 2021. TinyML for ubiquitous edge AI. *arXiv preprint arXiv:2102.01255* (2021).

[34] Benchmark Analysis of Representative Deep Neural Network Architectures, S. Bianco, R. Cadene, L. Celona, P. Napoletano



Dwith Chenna is a research and development professional with a strong focus on algorithm development and optimization in the fields of computer vision, deep learning and human computer interaction. He has extensive experience in developing state-of-the-art, performance-critical perception systems, and a deep understanding of the complexities involved in developing and optimizing deep learning models on resource-constrained hardware, such as digital signal processors. Dwith's responsibilities include evaluating embedded algorithms for performance and accuracy, and driving key performance metrics such as latency, memory, bandwidth and power consumption—often through integration and development of tooling and automation. He is also responsible for quantizing, optimizing and tuning the performance of deep learning mod

# Building scalable header platforms for large scale ecommerce websites

Damodaran Sathyakumar, eBay, Salt Lake City, UT, 84020, USA

**Abstract—** *Header widgets serve a very important role for websites in general & specifically to e-commerce websites. On e-commerce websites, they contain some of the most critical customer functionalities like – Sign In, Shopping Cart, Search, Notifications, Site Navigation, Language switch, Currency settings, Shipping Location, Shopping Lists, User Profile, and Settings. This article explores the various intricacies, optimization techniques and approaches in building scalable header platforms that have very good availability, reliability, resiliency, fault tolerance, render and interaction speed, thereby unlocking delightful customer experiences.*

Trivial at first glance and always assumed to be there, header widgets on e-commerce websites contain very crucial and critical functionality. The effects of broken pages and widgets are detrimental to the business and have a direct impact on the revenue as highlighted in the research by Nexcess [1] and Rewind [2].

Although occupying minimal real estate on a website, if the header widgets went down, none of the functions such as sign-in, shopping cart, search, site navigation, notifications would be available for the customer to make use of. Loss of functionality or incorrect functionality can be detrimental to business as evidence from the Citi Bank case [3]

Users would not being able to login or logout, switch accounts, set currencies, set shipment locations, perform search, access notifications, or view the shopping cart and other lists, to name a few. This adds to the list of root causes for increased bounce rate or lowered engagement rate as highlighted by Aillum [4].

On e-commerce websites this breakage translates to loss of revenue as highlighted in the research by Uptrends [5]. While pages on large scale websites get millions of views in site traffic, header widgets, being a horizontal function, end up serving billions of views in site traffic.

Their unique horizontal presence therefore provides an opportunity for websites to be able to serve other non-visual cross cutting horizontal concerns – meaning they become delivery platforms for non-

visual components like – common site-wide utils such as service workers, http clients, cookie utils.

Besides this, they also integrate with many other platforms like tracking, cart, payments, risk, identity, search, speed scripts, experimentation, security platforms and so on which need to deliver assets across pages.

But what does it take to build such a platform? How can they be built to serve traffic reliably & remain always available, while still maintaining a good metric for first content paint (FCP)? In this article, we will embark on a journey of the design considerations, principles, techniques & best practices in achieving this platform.

## DESIGN CONSIDERATIONS IN BUILDING HEADER PLATFORMS

The guiding design principles include:

- **Scalability:** The design should be able to handle billions of requests in traffic (for large scale e-commerce websites). This could be achieved by separating the UX as static and dynamic fragments & tackling them. For example, a link to the shopping cart is static and common to all users. But the modal / flyout, that loads the cart items data (upon hover) is specific to every user.
- **Decoupled View Layers:** In a large-scale e-commerce company, different pages may be powered by different view frameworks. So, it's necessary that the Header widget's view



- framework is neither an influencing factor in deciding the view layer for the entire page nor is it impaired in any way by the frameworks used by the domain product pages. In short, they need to be decoupled.
- *Decoupled apps & code Rollouts for increased release velocity and failure isolation:* Engineering velocity is essential for rapid iteration. It's crucial that header platforms can be built & iterated upon independently off the release cycles of other teams. This calls for decoupled apps. Also, besides helping on independent release and development cycles, decoupling apps helps isolate both the header platform & domain product pages from each other's failures.
- *Site Speed & Render Speed:* Site speed is very crucial for e-commerce websites as evident from research by Akamai [6]. Header widgets being the first element on the page, have to be first flushed to improve perceived experience such as FCP. It's important that our stack supports async rendering & streaming, while also not impacting the site speed metrics of product pages.
- *Interaction speed:* is a measure of how quickly the Header component becomes interactive after being rendered by the browser. There are some features to the header, such as search – that need to be immediately interactive, versus features like cart that can be loaded lazily.
- *Reliability:* It's crucial that the app is designed with reliability in mind. This includes predictable and reliable response times.
- *Resiliency:* While runtime errors may occur, it's essential that the app is resilient to it and not cause a degrading experience.
- *Fault Tolerance (quick recovery):* When runtime errors occur or if the app restarts due to out-of-memory issues, high central processing unit (CPU) usage, how quickly is our system able to revert to being able to serve traffic.
- *Availability:* Besides a good infrastructure cloud, it's important that the app deployed is generally

available 99.9% of the time and does not go down due to poor design.

- *Personalization:* The scale & design should include personalization. For example, the shopping cart for each user differs.

## SCALABILITY VIA SEPARATION OF STATIC & DYNAMIC PORTIONS

Static portions of the header widget include features like search, site navigation, sign-in links, greeting message, quick access links, universal site utils, partner platform integrations and so on.

Dynamic portions of the header widget include links / buttons when clicked / hovered, display the shopping cart, user profile (account / settings page), recently accessed / wish lists, notifications, user personalized dashboard page and so on.

During the design phase, it's necessary to come up with such a UX design that provides an on-demand access experience for such dynamic content.

This would enable us to tackle for scale by differentiating the content that is served. As indicated above, the dynamic portions are accessed only upon a user action while the static portions are always available. Also, the dynamic portions vary for every user while the static portions although varying, will however remain largely the same for large segments of users.

And note that, it's the static portions that will be first visible to the user and would need to be tackled first. With this segmentation, we would deal with separate production pools that serve the static & dynamic portions. This is highlighted in Figure 1 and Figure 2. If our intended UX was to display the dynamic portions in modals / flyouts, even if there are errors in serving the dynamic data, we can still prevent degrading the experience by offering a call-to-action link that takes the user to the intended landing page. But such errors don't grow to impact the static portions & lead us to a scenario where the user isn't even able to search / login / logout.

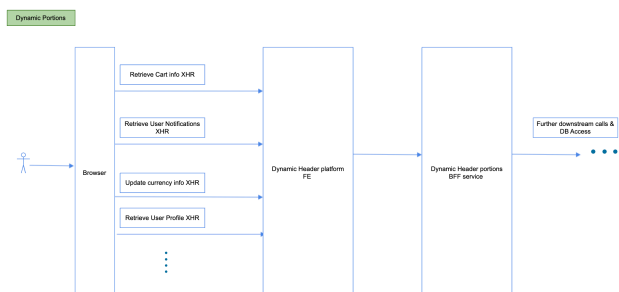


FIGURE 1. Possible Header platform Architecture for Dynamic Fragments

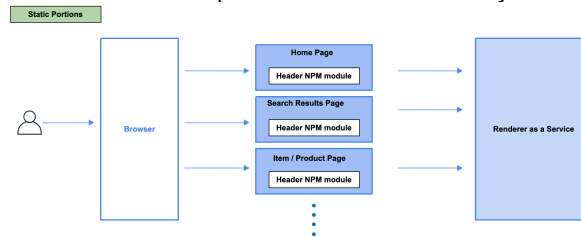


FIGURE 2. Possible Header platform Architecture for Static Fragments

## DECOUPLED APP VIEW LAYERS FOR MAINTAINABILITY

Today, e-commerce websites use a variety of view layer Frameworks like Vue, React, Solid, Angular, Marko, Qwik, Ionic, Flutter, React Native and so on, which come with their own reusable component libraries.

Being a cross platform function, header widgets cannot be re-built, functionality duplicated and maintained in a wide variety of frameworks. The only point of integration

with domain pages therefore is an adaptor (which are tag libraries in the specific framework) that call a downstream renderer-as-a-service or a back end for front-end service (BFF) (a cloud architecture pattern that tackles different devices [7]), where the actual rendering of the Header widget happens (refer to Figure2, Figure 3, and Figure 4).

With such a design, the Header widget is fully abstracted away from the framework upgrades & intricacies of the domain pages – which may be pages like the home page / search results page / item or product pages to name a few and devices – IOS vs Android and so on.

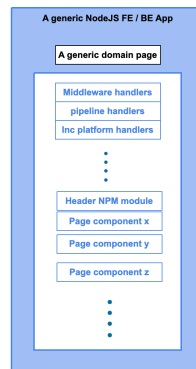


FIGURE 3. Possible Application Architecture for a generic Front-End application

## DECOUPLED APPS AND CODE ROLLOUTS

The need for such isolation and independence is because, teams want to be able to iterate & build fast without being tied to the release cycle of the various product page teams.

If the release of the Header widget features were to be coordinated with over 100 different product page teams that maintain numerous pages, our widget would never see the light of the day.

What we need here is a system of distributed front-end pools that power the various pages of the website. And the Header platform has its own production pools that, while separately serving the static and dynamic content, are also separated from the rest of the front-end production pools (as highlighted in Figure 1 and 2).

The other advantage of having separate production pools is that we are fully encapsulating both the Header

& the domain pages from each other's side effects - errors.

## SITE SPEED AND RENDER SPEED

Site Speed & Render speed are very significant factors that decide upon your tech stack's view layer. While there is a plethora of templating solutions, it's required that it is capable of doing asynchronous rendering & streaming.

By asynchronous renders, portions / fragments of your template can continue to render while the rest are awaiting data (as explained by Patrick Steele Idem in [8]). By streaming, portions of the rendered template can be flushed out of the buffer without waiting for the entire template to be rendered.

Large scale websites make use of a technique called progressive rendering [8] that flushes the output buffer multiple times either in-order or out-of-order. In selecting such a view solution, the Header

widget can be the first one to be flushed on the page, thereby improving on the core web vitals like FCP.

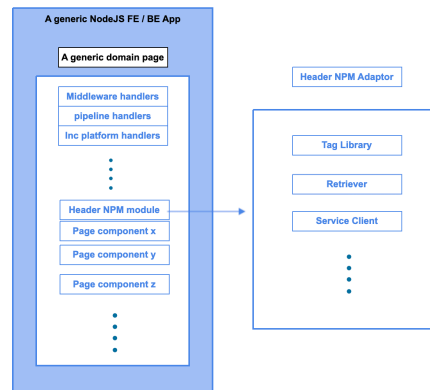


FIGURE 4. Possible exploded view of the adaptor in the application architecture detailing decoupled apps.

## INTERACTION SPEED

There are some portions of the Header that would need to be immediately interactive. For instance, the search box & the autocomplete. Usually, the JavaScript assets for interaction are bundled together to reduce network calls for improving performance as highlighted in research [9].

But to be able to achieve instant interactivity of specific fragments & an overall good Time to Interact (TTI) score on core web vitals, a bundler with declarative code splitting [10], based on slots that can be injected into the page, while still maintaining logical execution consistency is needed.

Besides, the view framework will either support the islands architecture [11] for lazy hydration or be a resumable framework [12], that eliminates hydration completely. Frameworks like Marko & Qwik are very good choices due to their default support for the island's architecture and resumability respectively.

## RELIABILITY

It's crucial that the Header widget has a very reliable response time, as it is the first widget that shows up on the top of any given page.

As mentioned earlier, whilst each of the important pages on a given ecommerce website can receive a few hundred million hits, the header platform, being a cross function across pages, would end up serving over a billion hits.

So, it's practically not feasible to be serving so much traffic by letting domain page adaptors drive all the traffic to the downstream renderer service. This is where the adaptor shines by employing a multi-tier least recently used (LRU) caching mechanism. The importance of cache cannot be downplayed in this scenario as highlighted by the talk "Caching for Cash" [13].

To understand caching for the header widget, we would need to dive deeper into the static portions. Although,

these are mostly static portions, that contain a window for personalization & variation.

The variations are determined by certain aspects. These aspects may be elements such as site ID (that uniquely identifies a site version – like UK / US / DE), page ID (that uniquely identifies a page / shell within a page if it's a Single page app), language (for all those users that

access a US version of the site from say, Russia). A cache key is generated based on these aspects.

The static portions are cached on the adaptor through this cache key. The cache is in-memory and maintained on the production pools of the product pages, within the adaptors.

Note that while there are variants via the aspects, each variant may end up serving millions of users. For example, the Header widget for the CA site, for the home page, with language English may serve a few million users, while the same Header widget for the CA site, for home page, with language French may serve a few more million users.

The key can also be composed of other factors like experimentation that divide the users into multiple treatment & control groups, thereby letting features be A/B tested on the Header widget.

With the multi-tier caching mechanism, the header widget is now scaled for individual pages to serve millions of users in traffic. Every production box of any domain product page, would have the respective Header adaptor installed, containing its own cache & collectively serving traffic to millions of users belonging to a segment as mentioned above.

Besides serving traffic reliably, the multi-tier caching is also able to provide us with predictable response times. As the content is cached, the static portions of the Header can be flushed out in less than 20ms predictably, improving on perceived site speed for the page.

## RESILIENCY AND FAULT TOLERANCE (QUICK RECOVERY)

While a major part of traffic is handled & served by the Adaptors of the respective pools. They still have to make calls to the renderer service when there is a cache-miss or when the cache entry is disposing upon reaching its time to live (TTL).

In such cases, those service calls, in a microservice environment can be prone to failures as highlighted by Dmytro in his research [14]. It's critical, that at that moment, the Header does not fail to show up on the page.

A common error which is a side effect of the domain pages is increased CPU usage & out of memory issues that cause boxes to re-start. While the load balancer infrastructure will handle this by diverting traffic to other

boxes, this can still cause problems with defective boxes. Using a circuit breaker pattern (popularized by Netflix's Hystrix [15]), we can provide some time for the boxes to recover from their failure, by marking down such boxes & periodically checking on their health.

There are other potential runtime errors that can be threatening. Such threats can be addressed by a system.

of disk based cold cache, that is immediately loaded into memory, when the app re-starts or when a service call has errored out. The cold cache basically contains variants of the header responses pre-generated. This can be immediately served while the main in-memory LRU cache is being warmed.

A setup like this (refer to Figure 5) can fully encapsulate the static portions of the header from any issues that could potentially prevent it from showing up on the page.

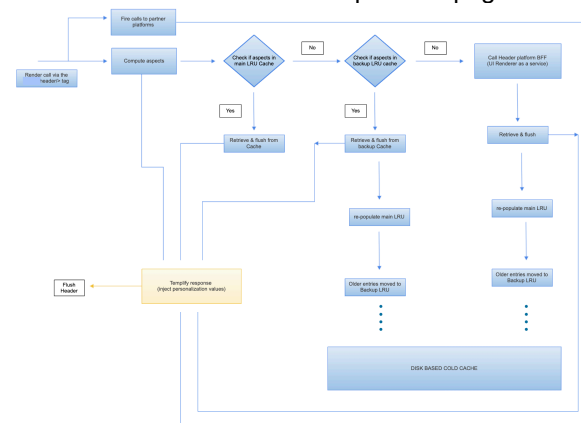


FIGURE 5. Possible architecture for the Header Adaptor with the caches.

## AVAILABILITY

While the architecture has scaled (via the multi-tier caching mechanism in the adaptors) to serve the Header widget on individual domain product pages, there is still quite a lot of traffic received by the renderer service. This is because, while the adaptors reside on the machines of the respective product pages and serve those individual pages, the renderer service must serve all the website's pages.

The adaptors offset a certain amount of the load via the multi-tier caching mechanism. But given that the renderer may receive quite a lot of traffic, it needs to be improved in the manner it handles incoming requests. Different pages may send requests to render the header, but can we optimize for the similar simultaneous requests that may come from multiple production boxes of the same page.

This is where we use a technique called request collapsing [16], which helps to handle multiple similar requests by processing just one of the N requests & re-using the same response for the rest of the N-1 requests. This prevents the thundering herd problem. Besides this, a similar multi-tier caching mechanism can be employed at the renderer service as well. Or a distributed key-value store can also be used if network latency is negligible.

One of the other concerns is when production boxes receive their first request. At this point the main in-memory LRU cache may not be warmed. This can be circumvented by a system of warmups where the production boxes self-trigger requests to pre-warm the LRU cache, as part of app bootstrap.



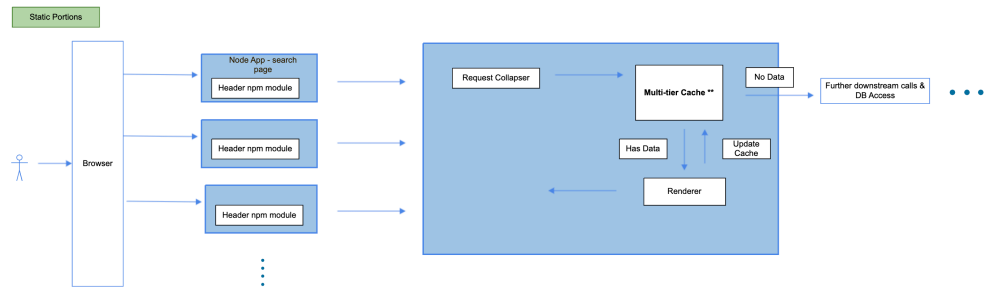


FIGURE 6. Possible architecture with the Renderer as a service included.

## PERSONALIZATION IN STATIC FRAGMENTS

While it was earlier mentioned how dynamic portions of the Header widget such as notifications, shopping cart, user profile contains personalized information and that static portions contains critical functionality, like search, but not any personalized information. This is not entirely true. There may be greeting / welcome messages, notification count, cart count and other very specific info that may need to be displayed immediately as the header shows up. This can be handled via data injection (Refer to “Templify” in Figure 5). While the generalized version of the header for a given segment of users is retrieved from the cache, its immediately templified (template injection) with the required personal information injected into it before final flush of the response buffer.

## DYNAMIC PERSONALIZED CONTENT

The full-blown user personalized dynamic fragments such as viewing notifications, the shopping cart, recently viewed lists, recently watched lists, user profile, account settings, language settings switch and so on are handled as on-demand actions which are triggered after the header shows up on the page.

These on-demand actions trigger multiple fetch requests to the production pool that is wired up to serve only dynamic user content, that is specific to every single user. Some of these actions may not be available for guest users.

But it’s essential to note that the links / buttons which trigger these calls are still served by the static portions of the header. Scaling this portion pretty much involves the best practices of scaling just another page on your website. These include front-end and back-end optimization techniques as highlighted by the research in [9].

Some important optimization techniques again come in Front End due to asset sizes that impact page load time

& JavaScript parse time. It’s generally useful that both the static portions and dynamic portions share the same templating view solution. Thereby, the runtime of the view / templating solution can be shared between the static & dynamic portions via solutions like module federation [17].

With that setup, the assets for the dynamic portions can be loaded lazily or individually for every on-demand hover / touch action. This can be achieved via a bundler plugin that computes & extracts an intersection of the dependencies into one JavaScript asset, while the individual components are loaded separately for every call & rendered on the client.

## CONCLUSION

Employing these techniques and considering the design cornerstones mentioned above would help build vastly scalable cross function Header platforms that bring the most crucial functionality to e-commerce web sites.

Note that, in the above architectures, you could easily substitute the BFF Renderer service for a GraphQL service. It is just a cloud design pattern to serve a variety of devices – web & mobile.

Also, further improvements come in the form of using a distributed caching mechanism for the renderer as a service.

## ACKNOWLEDGMENTS

We thank eBay Inc. This article is based on learnings from building a similar platform there.

## REFERENCES

1. Maddy Osman, “How to calculate the true cost of ecommerce downtime and reduce its impact,” [Online]. Available: <https://www.nexcess.net/blog/ecommerce-downtime/> (URL)
2. Corey Pollock, “The hidden costs of ecommerce downtime – and how to prevent it,

- [Online]. Available: <https://rewind.com/blog/ecommerce-downtime-cost-prevent/> (URL)
3. Lee, Timothy B. 2021. "Citibank just got a \$500 million lesson in the importance of UI design." Ars Technica. <https://arstechnica.com/techpolicy/2021/02/citibank-just-got-a-500-million-lesson-in-the-importance-of-ui-design/>
  4. Ailum, "7 causes of high bounce rates," [Online]. Available: <https://www.aillum.com/blog/7-causes-of-high-bounce-rate-on-websites/>
  5. Blog by Uptrends, "Calculate the true cost of website downtime," [Online]. Available: <https://blog.uptrends.com/technology/calculate-the-true-cost-of-website-downtime/> (URL)
  6. Akamai, "How web and mobile performance optimize conversion and user experience." [Online]. Available: <https://www.akamai.com/site/en/documents/white-paper/how-web-and-mobile-performance-optimize-conversion-and-user-experience-white-paper.pdf>
  7. Microsoft, "Backends for Frontend pattern," [Online]. Available: <https://learn.microsoft.com/en-us/azure/architecture/patterns/backends-for-frontends> (URL)
  8. Patrick Steele Idem, "Re-discovering progressive HTML Rendering with Marko," [Online]. Available: <https://tech.ebayinc.com/engineering/async-fragments-rediscovering-progressive-html-rendering-with-marko/> (URL)
  9. Ruchi Agarwal, "Building High performance modern web applications," *IEEE Feed Forward Santa Clara Valley* Chapter, July-Sep 2023, [https://r6.ieee.org/scv-cs/wp-content/uploads/sites/81/2023/08/FeedForward\\_Vol\\_2\\_Number3-1.pdf](https://r6.ieee.org/scv-cs/wp-content/uploads/sites/81/2023/08/FeedForward_Vol_2_Number3-1.pdf)
  10. Steele Idem Patrick, Piercey Dylan, Michael Rawlings, 'Code Splitting in Lasso JS,' [Online]. Available: <https://github.com/lasso-js/lasso#code-splitting>
  11. Lydia Hallie and Addy Osmani "Islands Architecture," [Online] Available: <https://www.patterns.dev/posts/islands-architecture>
  12. Ryan Carniato, "Resumable JavaScript with Qwik," [Online]. Available: <https://qwik.builder.io/docs/concepts/resumable/> (URL)
  13. Kent C Dodds, "Caching for Cash," [Online]. Available: <https://www.youtube.com/watch?v=amfVh0CZ78>
  14. Dmytro Semenov, "Building resilient platforms" [Online] Available: <https://codeburst.io/building-resilient-platform-part-1-51b852588fb3>
  15. Ben Christensen, "Introducing Hystrix for resilience engineering," [Online] Available: <https://netflixtechblog.com/introducing-hystrix-for-resilience-engineering-13531c1ab362?gi=409e2d4fef38>
  16. Fastly.com, "Request Collapsing," [Online]. Available: <https://developer.fastly.com/learning/concepts/edge-state/cache/request-collapsing/> (URL)
  17. Zack Jackson, "Webpack module federation, a game changer in JavaScript architecture," [Online] Available: <https://medium.com/swlh/webpack-5-module-federation-a-game-changer-to-javascript-architecture-bcdd30e02669> (URL)

**Damodaran Sathyakumar**, is a Staff Engineer at eBay Inc., Salt Lake City, USA. At eBay, he is known for his work on eBay's Header platforms, cross functional & horizontal widget platform, collectibles workflows & eBay's refurbished platform which together serve a billion impressions in traffic for eBay. He holds a Bachelors in Electronics Engineering from Anna University. With over 12 years of experience in building scalable web applications, his interests include all things web, Node JS, JavaScript & view frameworks.

## Acknowledgment of Reviewers

We would like to express our sincere appreciation to the following reviewers who generously dedicated their time and expertise to review the featured papers. Their commitment to scholarly excellence, dedication to the peer-review process, and valuable insights have significantly contributed to the high quality of the articles presented herein:

- Raghavan Muthuregunathan - Senior Engineering Manager, Search AI
- Meenakshi Jindal - Senior Software Engineer, Netflix
- Charankumar Akiri – Senior Application security engineer, Reddit
- Dwith Chenna - Senior Embedded DSP Engineer
- Utkarsh Mittal – Data Science Manager, Gap



Santa Clara Valley Chapter

**Chair**

Vishnu S Pendyala, PhD

# IEEE SCV CS Chapter Open House

**Vice Chair**

John Delaney

**Secretary**

Sujata Tibrewala

**Treasurer**

SR Venkatramanan



## Industry Rising Star Award 2023: Meenakshi Jindal

**Webmaster**

Paul Wesling

**Connect with us**

<https://r6.ieee.org/scv-cs/>

<https://www.linkedin.com/groups/2606895/>

<http://listserv.ieee.org/cgi-bin/wa?SUBED1=cs-chap-scv&A=1>

<http://www.youtube.com/user/ieeeCSStaClaraValley>

<https://www.linkedin.com/company/ieee-computer-society-scv-chapter/>

