Research Statement

Moussa Koulako Bala Doumbouya
Stanford University. October 2025.

Psychologically Plausible Deep Learning

Modern neural networks represent concepts in Cartesian feature spaces and assess similarity with geometric measures (e.g. cosine similarity) that fundamentally misalign with human psychological similarity judgments. This misalignment creates critical challenges: Neural networks remain black-boxes despite high performance, limiting trust in high-stakes applications. Current models require massive data and compute, with particularly severe impact on under-resourced domains like marginalized languages. Opaque representations hide biases and failure modes that can cause harm at scale. Additionally, we miss opportunities to leverage neural knowledge for human learning through pedagogically-grounded teaching methods.

To address these challenges, I integrated Tversky's (1977) well-established theory of psychological similarity into deep learning, closing the gap between modern neural networks and human cognition. Tversky's set-theoretic similarity model successfully explains human judgments but its non-differentiable discrete operations have prevented integration with gradient-based machine learning. To overcome this challenge, I developed the first differentiable parameterization of Tversky similarity, enabling gradient-based learning while preserving psychological plausibility. My research thus develops machine learning models aligned with human cognition to address four interconnected challenges: (1) Building neural networks humans can understand and interpret. (2) Leveraging interpretable neural representations to enhance human learning through the formalization of pedagogical devices based on similarity assessment. (3) Enabling inspection and correction of harmful behaviors and biases through explicit feature manipulation. (4) Improving efficiency across all domains through psychologically-motivated inductive biases, reducing parameter counts while improving performance, with particular benefits for under-served domains.

Tversky Neural Networks (TNNs)

My core technical contribution addresses the fundamental misalignment between how neural networks and humans represent concepts and similarity [16]. While standard deep learning relies on Cartesian representations and geometric similarity (e.g., dot-products, cosine similarity), psychological research has established that human similarity judgments follow different principles, notably lacking the symmetry and triangle inequality assumed by geometric models. Tversky's feature-matching theory offers a psychologically grounded alternative: objects are represented as sets of features, and their similarity is computed as a weighted combination of common and distinctive features. Despite its success in psychology, this model's non-differentiable set operations have prevented its integration with gradient-based deep learning. I developed the first differentiable parameterization of Tversky similarity that enables gradient-based optimization while preserving psychological plausibility. This breakthrough led to two novel architectural components: the *Tversky Similarity Layer* and the *Tversky Projection Layer*, which can replace standard geometric similarity and multi-layer perceptron modules throughout deep networks. These components demonstrate both theoretical and empirical advantages. Unlike linear layers, a single Tversky Projection layer can model the XOR function. Standard neural network architectures improve when their linear layers or multi-layer perceptrons are replaced with Tversky Projections. TverskyResNet-50 achieves 24.7% relative accuracy improvement on computer vision tasks, while TverskyGPT-2 reduces perplexity by 7.8% and parameters by

34.8% on language modeling tasks. Beyond performance gains, TNNs offer unprecedented interpretability. We developed methods to visualize learned prototypes that reveal human-understandable patterns: MNIST digit prototypes combine salient features from multiple handwriting styles (e.g., the "7" prototype includes both serif and non-serif variants). I demonstrated that set operations on TverskyGPT-2 token representations can algebraically specify semantic concepts, defining linguistic categories like adjectives, comparatives, and superlatives, or disambiguating word senses (e.g., isolating "plant" as factory versus vegetation). These results establish TNNs as a foundation for interpretable AI, computational pedagogy, and efficient deployment, particularly in resource-constrained domains.

Applications

Computational Pedagogy: I led an interdisciplinary team of computer scientists, cognitive scientists, and educators to develop my vision of using neural network representations to enhance human learning. We proposed NeuralPAC [15, 14], which leverages neural network representations to enhance human learning through the automated selection of pedagogical analogical and contrastive feedback stimuli. In randomized controlled trials with 1,088 learners, NeuralPAC improved learning by up to 26.5% (p=0.001) on synthetic stimuli. However, effectiveness diminished on natural stimuli, revealing that geometric representations optimized for machine performance poorly support human learning. This limitation directly motivated developing TNNs, whose psychologically-aligned representations promise substantial improvements in computational pedagogy applications.

Interpretable AI: TNNs' ability to algebraically specify semantic concepts opens new possibilities for interpretable AI. With TverskyGPT2, I demonstrated that linguistic categories like adjective degrees (base, comparative, superlative) and verb forms emerge naturally from set operations on token representations. My industry experience deploying interpretable models for efficient inference [10, 6] and similarity-based forensic information retrieval in video surveillance systems [13, 11, 4, 1, 12, 9] provides examples of application domains that can benefit from the interpretability and efficiency of TNNs in large-scale, real-life, mission-critical applications.

Safe AI: I led the development of h4rm3l [17], a framework for characterizing LLM safety vulnerabilities through synthesized jailbreak attacks. Using h4rm3l, we generated 2,656 successful black-box jailbreak attacks achieving 90% success rates against GPT-4o and Claude-3, revealing safety vulnerabilities in widely deployed large language models. Preliminary experiments with TverskyGPT2 suggest that TNNs could enable better understanding and mitigation of failure modes and hidden implicit biases, addressing critical safety needs and enabling safer deployments [3].

Efficient and Inclusive AI: My early PhD work focused on bringing NLP to marginalized languages through speech recognition [8, 2] and machine translation [7, 5], where data scarcity prevents using large-scale approaches. We also developed handwritten code recognition for pen-and-paper CS education [18], combining deep learning with domain-specific algorithms to overcome limited training data. TNNs' improved data efficiency through psychologically-motivated inductive biases positions them to democratize AI access in resource-constrained domains, from minority language processing to accessible education tools.

References

[1] Tulio Alcantara, Moussa Doumbouya, Eric Sjue, Hannah Valbonesi, and William Christopher Weston. Method and System for Interfacing with a User to Facilitate an Image Search for a Person-of-Interest, October 2020. US Patent 10,810,255.

[2] Martijn Bartelds, Ananjan Nandi, Moussa Koulako Bala Doumbouya, Dan Jurafsky, Tatsunori Hashimoto, and Karen Livescu. CTC-DRO: Robust Optimization for Reducing Language Disparities in Speech Recognition. *arXiv preprint arXiv:2502.01777*, 2025. Under review.

[3] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[4] Richard Butt, Alexander Chau, Moussa Doumbouya, Levi Glozman, Lu He, Aleksey Lipchin, Shaun P. Marlatt, Sreemanananth Sadanand, Mitul Saha, Mahesh Saptharishi, et al. System and Method for Appearance Search, July 2020. US Patent 10,726,312.

[5] Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, Djibrila Diane, and Solo Farabado Cissé. Smol: Professionally translated parallel data for 115 under-represented languages. In *Proceedings of the Conference on Machine Translation (WMT)*, 2025.

[6] Moussa Doumbouya, Xavier Suau Cuadros, Luca Zappella, and Nicholas E. Apostoloff. Semantic Coherence Analysis of Deep Neural Networks, November 2023. US Patent 11,816,565.

[7] Moussa Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory Conde, Kalo Mory Diané, et al. Machine translation for nko: Tools, corpora, and baseline results. In *Proceedings of the Eighth Conference on Machine Translation*, pages 312–343, 2023.

[8] Moussa Doumbouya, Lisa Einstein, and Chris Piech. Using radio archives for low-resource speech recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14757–14765, 2021.

[9] Moussa Doumbouya, Lu He, Yanyan Hu, Mahesh Saptharishi, Hao Zhang, Nicholas John Alcock, Roger David Donaldson, Seyedmostafa Azizabadifarahani, and Ken Jessen. Method and System for Facilitating Identification of an Object-of-Interest, January 2021. US Patent 10,891,509.

[10] Moussa Doumbouya, Lu He, and Mahesh Saptharishi. System and Method for CNN Layer Sharing, April 2020. US Patent 10,628,683.

[11] Moussa Doumbouya, Yanyan Hu, Kevin Piette, Pietro Russo, Mahesh Saptharishi, and Bo Yang Yu. Sensor Fusion for Monitoring an Object-of-Interest in a Region, September 2020. US Patent 10,776,672.

[12] Moussa Doumbouya, Yanyan Hu, Kevin Piette, Pietro Russo, Peter L. Venetianer, and Bo Yang Yu. Alias Capture to Support Searching for an Object-of-Interest, June 2021. US Patent 11,048,930.

[13] Moussa Doumbouya, Mahesh Saptharishi, Eric Sjue, and Hannah Valbonesi. Method, System and Computer Program Product for Interactively Identifying Same Individuals or Objects Present in Video Recordings, November 2018. US Patent 10,121,515.

[14] Moussa KB Doumbouya, Gabriel Poesia Reis e Silva, Lisa Einstein, Noah D Goodman, Daniel Schwartz, and Chris Piech. Finding pedagogically-effective contrasting cases in feature space: A pre-registered study. *Open Science Foundation*, 2022.

[15] Moussa Koulako Bala Doumbouya et al. Neuralpac: Teaching with analogies and contrasts in neural feature spaces. In *Under Review*, 2024. Under Review.

[16] Moussa Koulako Bala Doumbouya, Dan Jurafsky, and Christopher D Manning. Tversky neural networks: Psychologically plausible deep learning with differentiable tversky similarity. *arXiv preprint arXiv:2506.11035*, 2025. Under Review.

[17] Moussa Koulako Bala Doumbouya, Ananjan Nandi, Gabriel Poesia, Davide Ghilardi, Anna Goldie, Federico Bianchi, Dan Jurafsky, and Christopher D. Manning. h4rm3l: A language for composable jailbreak attack synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. Published as a conference paper at ICLR 2025.

[18] Md Sazzad Islam, Moussa Koulako Bala Doumbouya, Christopher D Manning, and Chris Piech. Handwritten code recognition for pen-and-paper cs education. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 200–210, 2024.