ABSTRACT
In this presentation, I will explore the manipulation of text-to-image (T2I) generative models through backdoors. I will present our Backdoor Attack on Generative Models (BAGM), which infuses the generated images with subtle manipulative details by injecting backdoors at three stages of the generative process: tokenizer, language model, or the image generator. For evaluation, we use a marketing scenario as the target domain and show that BAGM increases the bias towards target outputs more than fivefold without affecting untriggered model behavior. Beyond intentional manipulation, T2I models naturally contain bias which can propagate unfair social representations or push controversial agendas. We propose an evaluation methodology to quantify general biases in T2I generative models, without any preconceived notions. We assess four state-of-the-art T2I models and compare their baseline bias characteristics to their respective variants, where certain biases have been intentionally induced using BAGM. We propose three evaluation metrics: (i) Distribution bias, (ii) Jaccard hallucination and (iii) Generative miss-rate to measure biases under general and task-oriented conditions. Our method is available as a web application for measuring bias in any T2I model. In a follow up work, we expose the possibility of a dynamic and efficient exploitation of T2I model bias by targeting the language embeddings. By leveraging vector algebra, our technique enables convenient control over the severity of output manipulation and as a by-product, achieves a form of precise prompt engineering to generate images which are implausible with text prompts. Finally, we show a constructive application of this method for model debiasing and report compelling qualitative and quantitative results.

BIOGRAPHY
Ajmal Mian is a Professor of CS at The University of Western Australia. He has received two prestigious fellowships and several research grants from the Australian Research Council and the National Health & Medical Research Council of Australia. He was the West Australian Early Career Scientist of the Year 2012 and has received several awards including the Excellence in Research Supervision Award, EH Thompson Award, ASPIRE Professional Development Award, Vice-chancellors Mid-career Research Award, Outstanding Young Investigator Award, and the Australasian Distinguished Doctoral Dissertation Award. He is an Associate Editor of IEEE Trans on Neural Networks & Learning Systems, IEEE Trans on Image Processing and the Pattern Recognition journal. He was a General Chair of the Int Conf on Digital Image Computing Techniques & Applications (DICTA) 2019, General Chair of the Asian Conference on Computer Vision 2018, Program Chair of DICTA 2012 and Area Chair of WACV 2019, WACV 2018, ICPR 2016, ACCV 2014. Ajmal Mian has supervised 15 PhD students to completion and has published over 180 scientific papers. His research interests are in computer vision, machine learning, 3D point cloud analysis, facial recognition, human action recognition and video analysis.