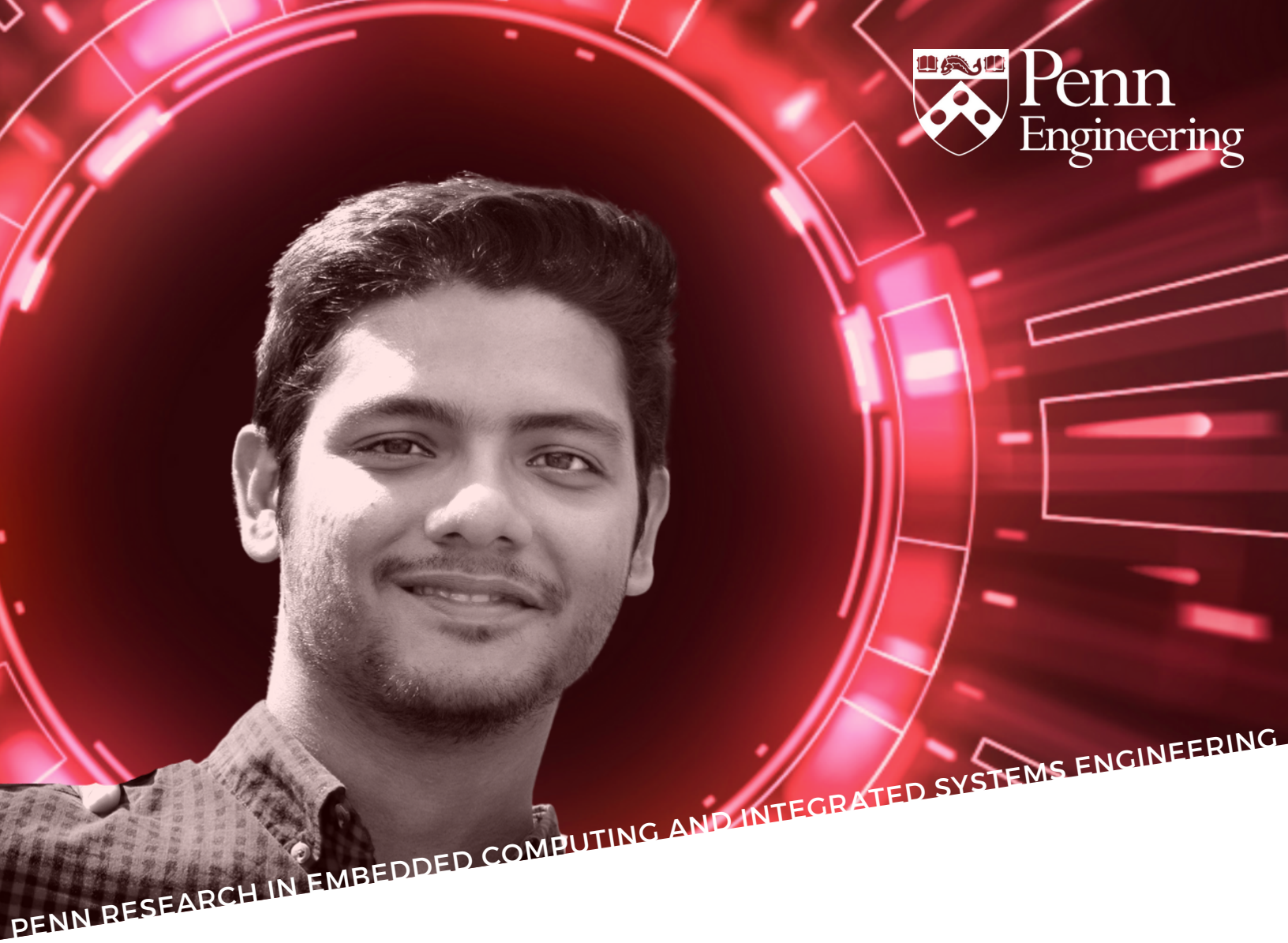

Penn Engineering

PENN RESEARCH IN EMBEDDED COMPUTING AND INTEGRATED SYSTEMS ENGINEERING

IN THE SPOTLIGHT

10 OCT 2023

SOURADEEP **DUTTA**

TRUSTWORTHY AI: TAMING THE BEAST FOR A BETTER FUTURE

Effective AI systems need to be resilient, secure and, above all, trustworthy.

Research by Souradeep Dutta, a postdoctoral researcher at Penn Engineering's PRECISE Center, focuses on building tools and algorithms for teaching computers to process data in a way similar to the human brain.

"Over the past few years, I have designed algorithms for verifying neural networks, and proposed memory-based deep neural networks for guaranteed learning outcomes," Dutta says. "My intellectual neighborhoods would be cyber-physical systems, artificial intelligence, formal methods, and reinforcement learning."

A native of Kolkata India, Dutta moved to the U.S. in 2016, earning his Ph.D. in electrical computer and energy engineering from the University of Colorado at Boulder. He came to Penn Engineering and the PRECISE Center in 2020 as a postdoctoral researcher advised by PRECISE Center Director Insup Lee.

"I have made some great connections at PRECISE, both in terms of meeting great mentors to guide me, as well as lifelong friendships, which have proved to be extremely helpful in this academic journey," Dutta says.

Souradeep with his close friends, Michele Caprio and Kuk Jin Jang, from the PRECISE Center

Dutta's research projects include developing a software tool called SHERLOCK, which helps to analyze neural networks and see whether they are processing data accurately.

"Using SHERLOCK, we were able to prove correctness of Airborne Collision Avoidance systems, and models for robotic systems and medical devices like the artificial pancreas," Dutta said.

For another project, Dutta created a model to help robots safely navigate anomalies by using training data to define a "comfort zone" for the system.

A third project proposes a new method for boosting the robustness of classifiers in machine learning. Classifiers are algorithms that organize and categorize data – for example, scanning email and organizing messages into "Spam" and "Not Spam."

"We introduced a two-stage classification framework called memory classifiers," Dutta says. "We applied this to improve robustness of electrocardiogram classification to predict different heart conditions."

In his free time, Souradeep enjoys hiking. Here is Granite Peak in Washington where he last visited

Dutta and his co-authors also applied the model to improve the accuracy of skin-lesion classification for acne grading in collaboration with experts at the Perelman School of Medicine.

In addition, Dutta is also working on a project to help AI systems conform to "common sense" practices and facts that would be obvious to humans, but are often absent in large machine learning models, and a project to better verify the safety of "Black Box" AI systems.

Finally, his research includes developing an online technique that aims to make autonomous systems more accurate by taking advantage of "distribution shifts" – for example, when a self-driving car receives a corrupt image of the road due to snow. In those types of situations, the issue tends to persist rather than disappear immediately.

"This gives the autonomous system a scope to adapt and recover from this shift by computing some semantic preserving transformation to the data: Essentially the system seeks to alter the appearance of the images without changing their meaning," Dutta says. "We discovered that this has a positive impact on the accuracy of subsequent tasks."

Dutta will be on the academic market beginning October 2023. If anyone is interested in interviewing him, please email precise@seas.upenn.edu to get the most recent version of his CV and to set up a meeting.